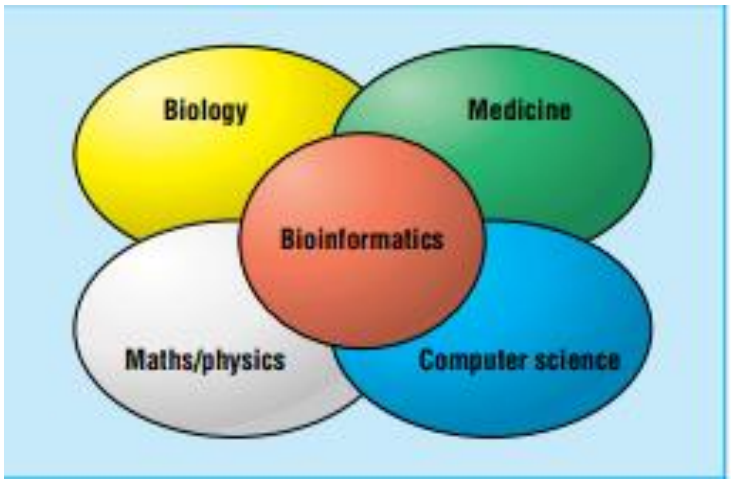
# BAB III
# LANDASAN TEORI

## 3.1 Bioinformatika

Bioinformaika lahir pada era tahu 70-an yang dimotori oleh seorang ilmuwan Amerika yang melakukan suatu inovasi pengembangan teknologi DNA rekombinan. Kemudian dibentuklah suatu perusahaan bioteknologi bernama Genentech yang memproduksi hormon insulin dalam bakteri (Aprijani dan Elfaizi, 2004). Perkembangan bioinformatika terus meningkat hal ini berkaitan dengan tingginya jumlah informasi molekuler teknologi biologi. Beberapa bidang Bioinformatik yang sangat pentik diantaranya adalah Pengolahan urutan DNA dan protein,*gene expression* anaisis, dan analisis struktur biologi. Bioinformatik adalah sutu aplikasi dari suatu teknologi informasi dengan studi biologi pada tingkatan molekul (Santamaria,2009).

Raza (2012) bioinformatika dapat diartikan sebagai salah satu cabang penerapan ilmu komputer terhadap pengelolahan informasi biologis yang bekerja sebagai pengalian informasi, pengelompokan, analisis, interpretasi dan pemanfaatan informasi dari urutan biologis dan molekul dengan tujuan utama dari ilmu ini adalah peningkatan pemahaman dari proses biologis.

Wargasetia (2006) bioinformatika merupakan sebuah teknologi yang digunakn untuk mengumpulkan, menyimpan, danmenganalisisdata biologi molekul yang juga berperan dalam bidang klinis, identifikasi mutasi, hingga identifikasi diagnosis penyakit baru.Sedangkan menurut Bayat (2002) bioinformatika merupakan suatu aplikasi dan alat komputasi yang digunakan untung menginterprestasikan data-data biologi. Bioinformatika juga merupakan suatu ilmu yang interdisipliner yang saing menujang dengan disiplin ilmu lainya diantaranya adalah bidang biologi, kedokteran, fisika dan matematika, serta ilmu komputer.




Sumber : Bayat (2002)

**Gambar 3.1***Kontribusi Ilmu Terapan Lainya Pada Bidang Bioinformatika*

Terdapat sepuluh area ilmu yang berkembang pada bidang bioinformatika, yaitu ( Raza, 2012) :

1. *Sequencing analaysis*

   Analisis *sequencing* atau urutan DNA biasanya akan mencari bagian urutan yang sama atau berbeda pada saat analisis medis dan proses pemetaan genom.
2. *Genome annotation*

   Anotasi atau *annotation* adalah proses menandai gen dan feature biologis lainya dalam suatu urutan DNA.
3. *Analysis of gene expression*

   Suatu ekspresi dari sejumlah gen dapat ditentukan dengan mengukur level mRNA dengan menggunkan berbagai teknik seperti *microarray*, *expression cDNA sequencing tag* (EST), *serial analysis of gene expression* (SAGE), *massively parallel signature squencing* (MPSS).

4. *Analysis of protein expression.*

   Analisis ekspresi protein adalah suatu analisis yang memberikan petunjuk yang baik pada aktivitas gen. hal ini dikarenakan oleh protein merupakan katalis akhir pada aktivitas sel.

5. *Analysis of mutation in cancer.*

   Genom dari sel yang terkena kanker akan di tata ulang dengan cara yang kompleks sehingga terkadang tidak dapat di prediksi. Adanya *massive sequencing* adalah salah satu upaya untuk mengidentifikasi suatu titik mutasi gen pada kanker. Ilmu bioinformatika semakin berkembang menghasilkan suatu sistem otomatis untuk mengukur volume data *squencing* yang dihasilkan dan membuat suatu algoritma baru dan software untuk membandingkan hasil *sequencing*dengan kumpulan genom manusia yang terus bertambah *sequence*dan garmaline polimorfisme. Analisis untuk melihat titik lesi yang ditemukan berulang diantara banyak tumor merupakan analisis lain yang sekarang berkembang.

6. *Protein structure prediction.*

   Urutan asam amino protein yang biasanya disebut dengan struktur primer dapat dapat ditemukan pada gen. Mengetahui struktur protein sangat penting untuk memahami fungsi protein yang berfungsi dalam kajian desain obat dan desain enzim baru.

7. *Comperative genomic*

   *Comperative genomic* adalah suatu ilmu yang mempelajari hubungan dari struktur genom dan fungsi biologis yang berbeda jenis.*Comperative genomic*memanfaatkan persamaan dan perbedaan protein, RNA, dan daerah bagian regulasi dari organisme yang berbeda.

8. *Modeling biological system*

   Pemodelan biologis bertujuan untuk mengembangkan algoritma yang efisien, struktur data, visualisasi, dan alat komunikasi untuk mengintegrasikan data biologis dalam jumlah besar dengan tujuan pemodelan computer. Hal ini melibatkan simulasi sistem biologi, seperti subsistem sel jaringan metabolic dan enzim.

9. *High-troughput image analysis.*

   Analisis ini bertujuan untuk mengukur dari kumpulan gambar yang besar dan komplek. Contoh dari penelitian tersebut adalah *clinical image analysis* dan visualisasinya dan informatik *bioimage*

### 3.2 *Microarray*

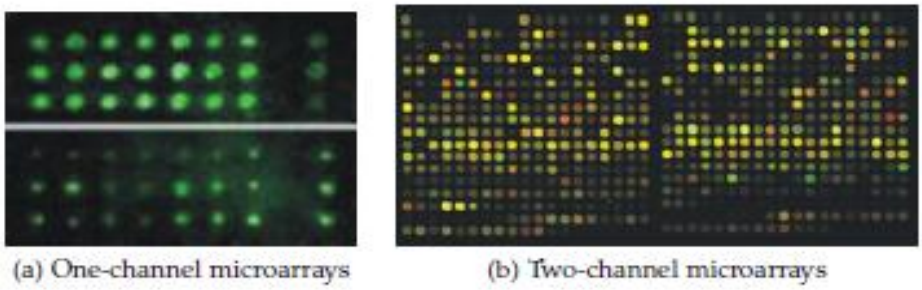
*Microarray*adalah suatu *chip* atau *slide* mikroskop yang berisikan serangkaian sampel berupa DNA, RNA, protein, dan jaringan (Pasanen,2003). *Chip* gen dari *microarray* terbuat dari suatu silikon atau kaca dimana bahan genetik akan ditempatkan dan memiliki struktur seperti grid, setiap spot yang mengandung rangkaian *nukleotide* tunggal yang berbeda disebut sebagai *probe* dan etiap *spot* akan mempunyai jutaan salinan probe (Santamaria, 2009).

Data *microarray* pada awalnya adalah sebuah gambar, dimana untuk mendapatkan tingkat ekspresi data gambar tersebut akan dianalisis terlebih dahulu dengan cara setiap titik pada gambar akan diidentifikasi,diukur itensitaskemudian dibandingkan dengan latar belakang yang disebut dengan kuantitas gambar. Hasil dari data tersebut akan berbentuk matriks *gene expression* (Brazma dkk, 2001).*Gene expression* akan terlihat apabila DNA yang mengandung informasi molekul akan direkam untuk menampilkan RNA-nya(Noviani dan Yoga,2010).

Terdapat dua jenis *microarray* mengarah pada jumah gen yang dimasukan pada setiap *chip*(Santamaria,2009) :

1. *One-channel microarray*adalah urutan gen dari suatu sampel akan dimasukan pada *microarray* untuk proses hybridisasi, adapun contoh dari jenis ini adalah *affymetrics genechip. One channel* membutuhkan dua *chip* untuk menampilkan perbandingan sampel, dan dilakukan secara komputasi
2. *Two-channel microarray*adalah urutan gen dari dua sampel yang berbeda kemudian dimasukan kedalam *microarray* utuk proses hibrisasi kompetitif. Jenis ini diberi warna neon yng berbeda yaitu warna hijau (Cy3) atau merah (Cy5) sehingga biasanya disebut dengan Cy3/Cy5 *microarray*s. Jenis tersebut hanya membutuhkan satu *chip* untuk

menampilkan perbandingan sebuah sample kontrol dengan sampel percobaan penyakit.




Sumber : Santamaria (2009)

**Gambar 3.2***Jenis Microarrays*

Menurut Cherkas (2010) *microarray* teknologi memiliki dua pendekatan yaitu :

1. *cDNA Array*

   cDNA dapat disebut *dual channel* atau *microarray* dengan dua warna mRNA dari dua sampel biologis yang bebeda ditranskrpsi ulang menjadi cDNA , dan diberi lebel dengan warna hijau (Cy3) atau merah (Cy5).

2. *Oligonucleotide array*

   *Oligonucleotide array* adalah suatu platform yang digunakan untuk mengukur*gene expression* yang memiliki kepadatan tinggi. Susunan *chip* silikon mengandung probe dari untaian *oligonucleotide* yang pendek yang terdiri dari 11-20 pasang *oligonucleotide* yang masing-masig memiliki panjang 25 pasang basah.

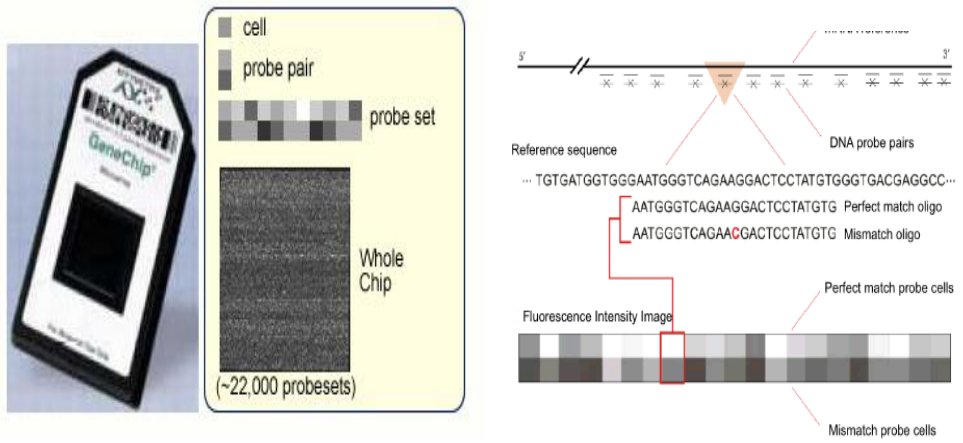
Metodologi dalam proses analisis aplikasi *microarray* terbagi atas *Single nucleotide polymorphism (SNP) microarray, Cromatin inmuno precipitation on gene chip (Chip on chip),* dan *Comparative genomic* Santamaria (2009).

*Microarray* manusia HG-U133 memiliki tiga *probe* sebagai berikut (Serin,2011) :

1. *_at* apabila suatu probe set disambung dengan *_at* diartikan sebagai suatu keadaan dimana semua probe mengenai suatu transkip yang telah diketahui.
2. *s _at* suatu probe diberi tanda *s_at* diartikan bahwa ketika semua probe sama persis dengan transkrip alternatif dari gene yang sama.Hal tersebut juga dapat terjadi pada transkrip dari gen yang homolog.
3. *x _at* penambahan nama *x_at* dapat diartikan ketika beberapa probe saling identik atau memiliki persamaan yang sangat tinggi, dengan urutan (*squencing*) tidak seling berkaitan, hal ini dapat meyebabkan *cross-hybdrization* ke *squence* yang merupakan bukan traskrip targetnya.

### 3.3 *Affymetrix*

*Affymetrix genechip* merupakan *chip* yang hanya memiliki satu target yang dimasukan pada setiap *chip*. Suatu tingkatan intensitas tertentu akan dibaca dari setipa spot atau sel dalam *chip*, dimana setiap gen akan disajikan dengan beberapa bagian DNA (probe) yang pendek dan sesuai dengan gen pada setiap *spot* kumpulan dari probe disebut juga sebagai probe setkemudian hasil dari PM/MM akan dilakukan proses *summarized* untuk dapat membaca hasil rata-rata dari setiap gen. *Affymetrix* memiliki rangkain *chip* yang sangat besar memiliki sekitar 500.000 *spot* sehingga membutuhkan RAM yang besar. (Genstat, 2017)

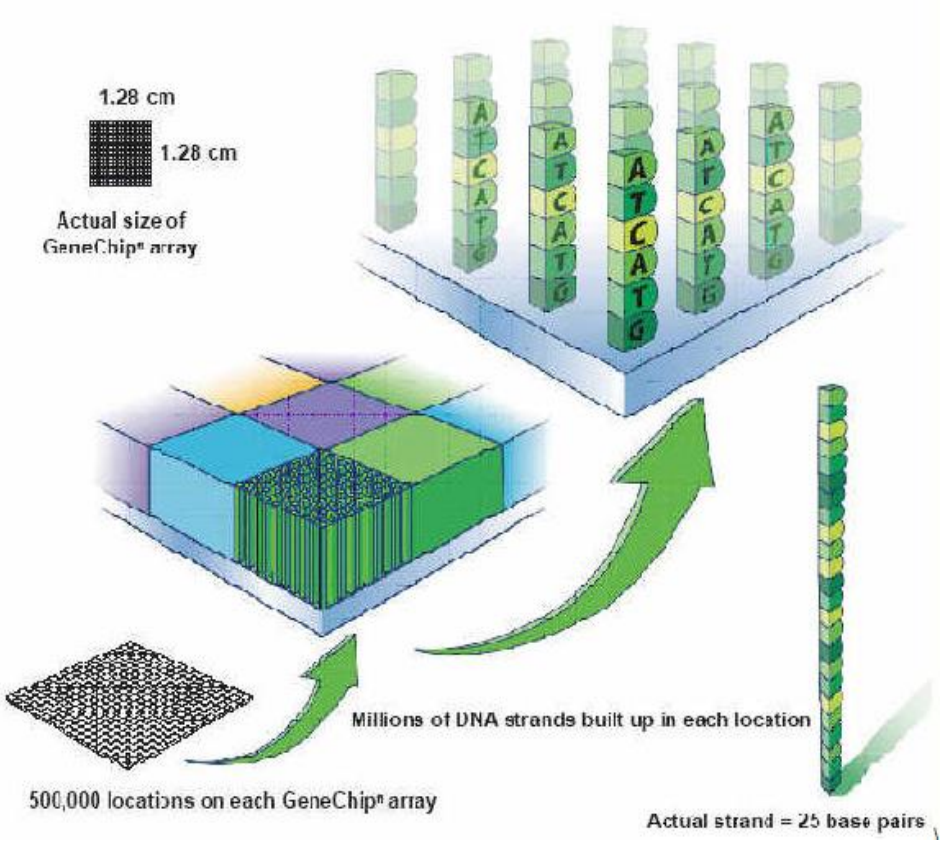



Sumber : (https://www.vsni.co.uk/software/genstat/htmlhelp/marray/AffymetrixChips.htm)

**Gambar 3.3** *Affymetrix Microarray*

DNA yang terdiri dari basa, yaitu adenin (A), guanin (G), timin (T), citosin (C) dimana C akan berpasangan dengan G dan A akan berpasangan dengan T.

Suatu untai DNA hanya akan berpasangandengan RNA yang sesuai. Kemudian *microarray* akan memakai pasang basa yang telah berpasangan yang merupakan hasil dari proses hibdrisasi (Affymetrix,2007)




Sumber : (https://www.vsni.co.uk/software/genstat/htmlhelp/marray/Affymetrix*Chip*s.htm)

**Gambar 3.4** *Proses Aymetrix*

### 3.4 *Gene Expression*

Proses*gene expression* terbagi atas dua tahap yaitu transkripsi dan translasi. Traskripsi adalah prosesinformasi genetik yang adapada DNA akan menghasilkan RNA diantaranya adalah RNA duta (mRNA), RNA transfer (tRNA), dan RNA ribosomal (rRNA) pada akhirnya hanya mRNA yang akan di translasikan ke dalam protein. mRNA dan tRNA masih memberikan peranan penting didalam proses translasi dimana rRNA yang berfungsi pada pembentukan ribosom, dan tRNA yang membawa asam amino yang sesuai dengan informasi yang ada didalam molekul mRNA pada proses translasi (Suharsono, 2005)

*Gene expression* merupakan informasi komplek yang dikodekan dalam gen yang dignakan untuk menghasilkan produk yang fungsional seperti protein yang menetukan fungsi sel. Matrik pada eksprsi gen terdiri dai baris dan kolom dimana setiap baris akan mewakili jumlah gen *N* dan kolom akan mewakili kondisi *M*(Panda, 2012). Setiap satuan matrik atau sel adalah suatu bilangan real dan mewakili tingkat*gene expression* di bawah suatu kondisi percobaan (Chekouo,2012).

**Tabel 3.1***Gene Expression Dataset*

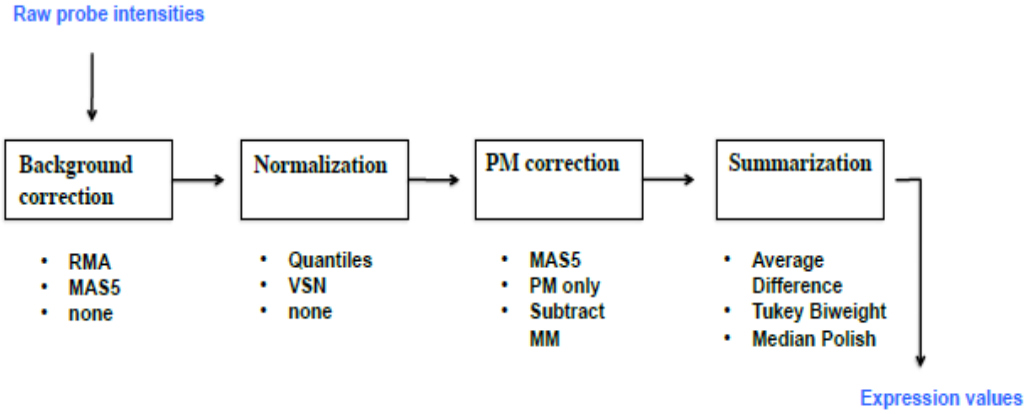
| | **sample 1** | **sampel 2** | **sampel3** | **.....** | **sampel M** |
|---|---|---|---|---|---|
| gene 1 | | | | | |
| gene 2 | | | | | |
| gene 3 | | | | | |
| gene 4 | | | | | |
| ..... | | | | | |
| gene N | | | | | |

Jamous (2015) menyatakan sampel suatu *gene expression*akan mengandung beberapa kondisi berikut :

1. Perbedaan kondisi biologis sematik seperti perbedaan tipe jaringan seperti jaringan kulit, tulang, darah dan saraf.
2. Perbedaan waktu pada rentang linier ataupun nonlinear, seperti serangkaian sampel yang diambil setiap menit, jam, ataupun hari yang berbeda.
3. Berdasarkan tahap kronik dalam proses biologi, seperti sampel yang diambil dari tahap awal, intermediate dan akhir yang didefinisikan dengan baik dari sel yang berkembang dalam proses.

### 3.5 *Preprocessing*

*Preprocesssing* adalah suatu proses yang digunakan untuk membuang efek non biologi pada data dan membantu memberikan hasil yang lebih baik. Setiap proses *preprocessing* memiliki pilihan yang berbeda seiap tahapnya (Serin,2011) .




Sumber : Serin (2011)

**Gambar 3.5** *Proses preprocessing*

Gambar 3.5menjelaskan bahwa proses *preprocessing* terbagi atas tiga tahap yaitu *background correction, normalization,* dan *summerization*yang memiliki fungsi sebagai berikut (Bolstad,2004) :

1. *Background correction* adalah suatu metode yang berfungsi sebagai menghilangkan *background noise*, menyesuaikan *cross hibdrization* yang merupakan pengikat dari DNA non-spesifik yang melekat pada *array*.
2. *Normalization* adalah suatu proses yang digunakan untuk menghilangkan variansi non biologis yang tidak diinginkan dan mungkin ada pada *microarray.*
3. *Summerization* adalah suatu proses pengabunga beberapa intensitas probe menjadi probe set yang nantinya akan menghasilkan nilai ekspresi.

### 3.6 *Filtering*

*Filtering*merupakan suatu proses yang digunakan untuk mengurangi jumah gen dan meningkatkan kekuatan dalam suatu analisis. Ketika jumlah gen sangat banyak dan proporsi tingkat*gene expression* sangat rendah dapat mengakibatkan

akurasi yang buruk dalam suatu pengamatan analisis. *Filtering*berkerja dengancara mendeteksi variansi, mendeteksi rata-rata sinyal gen atau *MAS detection call* (Hackstadt dan Hess, 2009)

### 3.7 *Feature Selection*

*Feature selection* atau sering dikenal dengan *gene selection* merupakan suatu metode yang digunakan untuk memilih gen yang relevan yang akan digunakan dalam proses klasifikasi. *Feature selection* terbagi atas beberapa pendekatan diataranya adalah *filter* dan *wrapper*. Pendekatan filter dapat dilakukan dengan menggunaka t-test, wilcoxon, dan *analysis of variance* (ANOVA). *Feature selection* dengan menggunakan *filter* memiliki kekurangan, yaitu setiap interaksi dan korelasi antara gen lainya akan diabaikan.

### 3.8 Kanker

Kanker adalah suatu penyakit yang disebabkan oleh pertumbuhan dan penyebaran yang tidak terkontrol dari seltidak normal (ACS,2017). Secara umumkanker dibagi atas dua jenis yaitu kanker jinak dan kanker ganas perbedaan antara kedua jenis adalah tipe jinak memiliki proses penyebaran yang lambat dibandingkan tipe ganas. Penyebaran kanker yang berkembang lebih cepat dari sel normal akan menarik pembuluh darah, berjuang melawan sistem pertahanan tubuhdan menyebar keorgan lain yang disebut metastasis (Faried,2015).

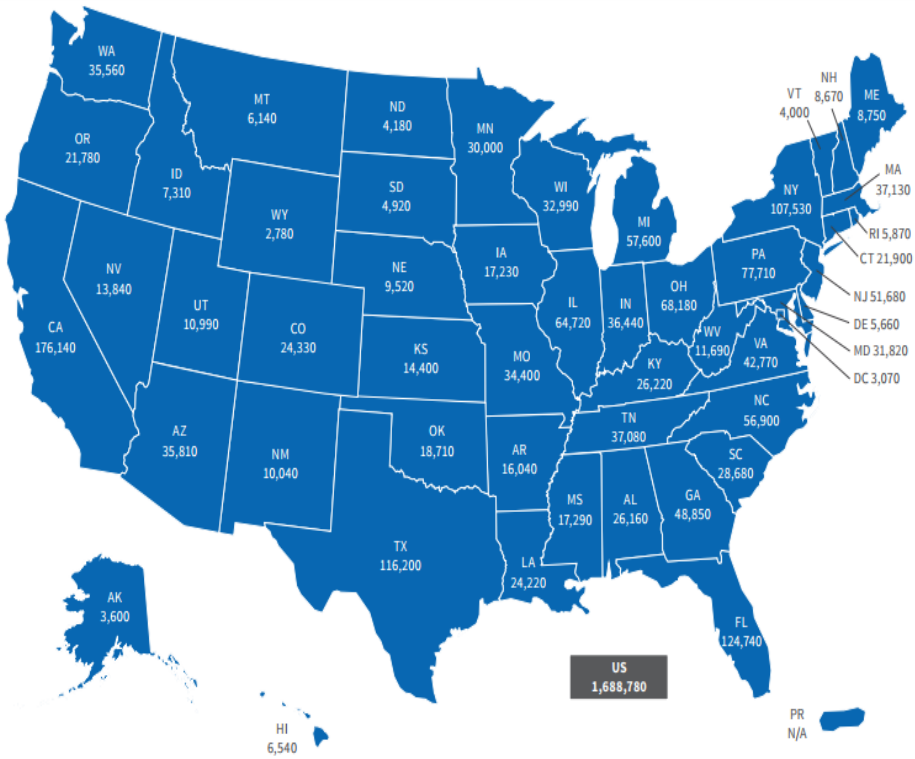
Robert, dkk (2007) mengatan tumor pada kanker dapat dikelompokan berdasar tipe jaringan, keberadaan histologi, dan tingka keganasaanya. Terdapat tiga bentuk utama dari kanker yaitu :

1. *Sarkoma* adalah suatu tumor yang muncul pada jaringan mensenkim seperti tulang, otot atau jaringan ikat.
2. *Karsinoma* adalah kanker yang berasal dari jaringan epitel seperti pada sel usus, bronkus dan saluran mamaria.
3. *Hematopoietik* dan *limfoid* adalah kanker yang menyerang pada bagian sumsum tulang belakang seperti leukimia dan limpoma atau pada sistem limfatik.

Faktor yang dapat meningkatkan resiko kanker pada seseorangadalah penggunaan rokok, mengkonsumsi alkohol berlebihan dan paparan karsinogen.

Pengkomsumsian tembakau akan menyebabkan tumor pada laring, pangkreas, ginjal, dan kandung kemih. pada negara berkembang merokok memberikan resiko 30% dari semua tumor ganas (IARC,2003).

Kanker adalah penyebab kematian terbesar di dunia, terhitung pada tahun 2015 terdapat 8,8 juta kenatian yang dikarenakan kanker dan tercatat 1,68 juta lainya adalah kaker paru-paru (WHO,2017). Menurut American Cancer Society (2017) Mengatakan bahwa estimasi penderita kanker pada tahun 2017 di Amerika mencapai 1.688.780 jiwa.



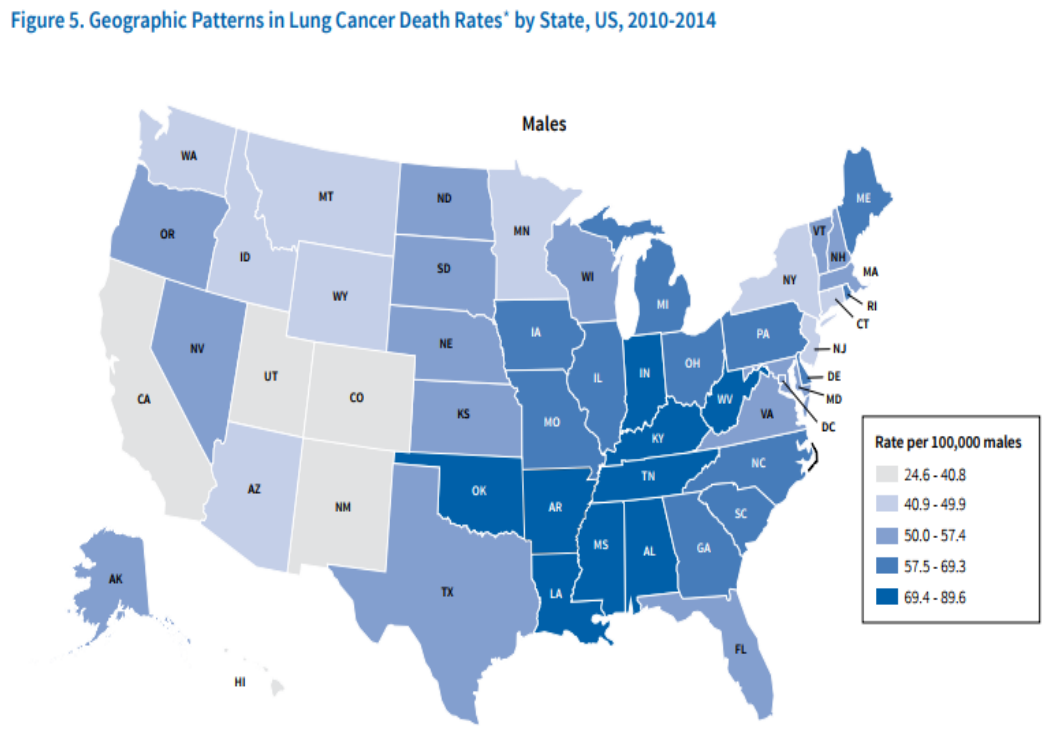

Sumber : American Cancer Society (2017)

**Gambar 3.6** *Kanker di Amerika*

### 3.9 *Kanker Paru -paru*

Kanker paru-paru biasanya meliputi karsinoma yang timbul dari trankea epitel, broknkus dan terdapat beberapa tipe histogi diantaranya adalah karsinoma sek skuamosa, andekarsinoma dan sel kecil karsinoma.

Estimasi tingkat kematian kanker paru-paru di Amerika Serikat pada tahun 2017 yaitu 155.870 jiwa. Tingkat kematian pria berdasarkan kanker paru-paru di negara bagian Amerika pada tahun 2010-2014 adalah sebagai berikut :




Sumber : American Cancer Society (2017)

**Gambar 3.7** *Kanker paru-paru pada pria di Amerika*

*American Joint Commitee onCancer* menetapkan suatu sistem untuk menentukan stadium pada kanker paru-paru dan dibagi mejadi beberapa stadium I – IV dan pada stadium I dan II di sebutstadium awalpada stadium ini kesempatan penderita untuk bertahan akan lebih besar. Penilaian dari setiap stadium diperoleh dari ukuran tumor (T), terdapatnya *lymph node* (N), dan terjadinya *metastasis* (M) (AJCC,2017).

Kanker paru – paru dibedakan menjadi dua tipe yaitu *non-small cell lung cancer (NSCLC) small cell lung cancer (SCLC)* dan dibagi menjadi beberapa subtipe sebagai berikut (Klamerus dkk, 2009) :

***Tabel 3.2****Type* dan *subtype* kanker paru-paru

| Type | Subtype |
|---|---|
| *Non-Small Cell Lung Cancer*<br><br>( 85% dari kasus Kanker yang terjadi ) | 1. *Adenocarcinomas* ( 50-60% dari NSCLC)<br>2. *Squamous Cell Carcinomas* ( 20 - 25% dari NSCLC)<br>3. *Large Cell Carcinomas* ( 15% dari NSCLC)<br>4. Subtipe yang jarang ditemui : *Neuroedocrine atau Carcinoid Tumor, Carcinocarsinomas, Cystic Adenoid Crcinomas.* |
| *Small Cell Lung Cancer*<br><br>Terjadi 15 % dari kasus kanker paru-paru | *1. Oat Cell ( lymphohacyte-like)*<br>*2. Polygonal Cell*<br>*3. Fusiform*<br>4. *Mixed* |

Faktor resiko pada kanker paru-paru terbagi atas tujuh faktor yaitu (NCCN,2016 ) :

1. Merokok

   Merokok menjadi faktor yang paling mempengaruhi kanker paru – paru dimana 50 komponen yang ada pada rokok merupakan penyebab kanker. Semakin banyak komsumsi rokok setiap harinya maka akan semakin tinngi pula resiko kanker paru-paru.

2. Umur

   Resiko kanker pada paru-paru akan meningkat apabila semakin tuanya umur. Rata-rata pasien penderita kanker didiagnosa memiliki kanker paru-paru diusian 70 tahun, hanya 12 dari 100 kasus kanker paru-paru penderita berumur dibawah 55 tahun.

3. Memiliki kanker

   Maksud dari memiliki kanker disini adalah apabila seseorag memiliki penyakit kanker otak atau kanker leher yang memiliki hubungan dengan merokok maka akan semaking meningkatkan resiko terkena kanker paru. Biasanya resiko ini akan meningka setelah adanya terapi radiasi.

4. Memiliki anggota keluarga penderita kanker

   Apabila seseorang memiliki anggota keluarga seorang penderita kanker dengan umur yang masih muda atau memiliki keluarga penderita kanker lebih dari satu maka akan meningkatkan resiko terkena kanker.

5. Memiliki kontak langsung pada penyebab kanker

   Penyebab kanker disini adalah berupa suatu zat seperti uranium yang yang terdapat pada bebatuan dan tanah ataupun zat lainya seperti arsenic, berilium,cadmimum, nikel, asbestos, coal smoke,soot,silic, dan diesel.

6. Infeksi paru-paru

   Infeksi pada paru-paru terkadang memberikan pengaruh terjadinya kanker paru-paru seperti, seseorang pernah infeksi fungal maka resiko terkena kanker paru-parunya akan lebih besar dari pada orang yang tidak pernah terkena infeksi fungal.

7. Penyakit lainya pada paru-paru

   Terdapat dua jenis penyakit yang dapat menigkatkan faktor resiko pada kanker paru-paru yaitu COPD (*chronic obstructive pulmonary disease)* dan *pulmonary fibrosis.*

Klamerus dkk (2009) menyatakan bahwa terdapat 1 dari 5 wanita yang penderita kanker dengan status tidak pernah merokok dan 1 dari 10 pria yang tidak pernah merokok.

### 3.10*K- Nearest Neighbor*

*K-nearest neighbor* adalah sebuah algoritma pada *supervised mechine learning* yang mengklasifikasikan suatu kategori kedalam jarak kelas ketertanggaan terdekat. Nilai *k* disebut dengan jumlah ketertangganya yang biasanya lebih dari satu dan digunakan untuk penentuan kelas.K-NN adalah metode yang paling simpel dan teknik yang paling intutif pada statistik

deskriminasi. *K-nearest neighbor* disebut sebagai non-parametrik klasifikasi karena tidak terdapat asumsi yang mengikuti metode. K-NN mengklasifikasikan suatu pengamatan baru kedalam kelas yang sama dengan suatu pengamatan dari *training set* yang paling dekat dengan suatu pengamatan baru (Cover dan Hart,1967).

Jarak terdekat pada pada K-NN dengan algoritma *euclidean* dengan persamaan sebagai berikut :

$$d\ (a,\ b) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}$$

dimana :

*a*adalah vektor observasi dari objek a

*b* adalah vektor observasi dari objek b

K-NN telah berhasil diterapkan pada klasifikasi dengan data yang besar seperti analisis pola tulisan tangan, gene expression, pola EKG, atau pencitraan satelit.

Klasifikasi dengan menggunakan K-NN melewati beberapa tahap yaitu (Santoso,2007):

1. Hitung semua jarak antara data *testing* dan data training
2. Menentukan k data *training* yang jaraknya paling dekat dengan data testing
3. Periksa label dengan jumlah paling banyak
4. Masukan data *testing* kedalam jumlah label paling banyak

### 3.11*Cross Validation*

*Cross validation* adalah teknik yang digunakan untuk mengestimasi suatu akurasi dan biasanya*cross validation*menggunakan sejumlah data random dari *test set* atau *training set,* atau menggunakan *K-fold cross validation* (Mullin dan Suthankar, 2000).

*K-fold cross validation* adalah suatu pendekatan yang akan mensegmentasi data*K* partisi yang berukuran sama, selama proses ini salah satu partisi akan dipilih menjadi training, sedangkan sisanya akan menjadi testing. Metode ini akan diulang sebanyak *K* kali sehingga partisi digunakan untuk *testing* tepat satu kali, dengan

kata lain metode ini akan membagi *testing* sebanyak satu kali dan *training set* sebanyak K kali perulang (Razaq, 2015). Pada penelitian tersebut digunakan *10 – fold cross validation* yang dapat dijelaskan sebagi berikut :

1. Tahap pertama akan dibagi *K* subset dengan ukuran yang sama.
2. Tahap kedua adalah menggunakan setiap subset menjadi *testing* data dan sisanya akan dijadikan data training. Pada 10 *fold* maka data akan dibagi menjadi 10 subset dengan ukuran yang sama pada iterasi pertama satu bagian akan digunakan sebagai test data dan bagian lainya adalah train data dan pada iterasi kedua seubset kedua akan dijadikan data test dan data laiya akan menjadi data train.

### 3.12 *Confusion Matrix*

*Confusion matrix* adalah sebuah matrik yang digunakan untuk mengevasluasi hasil dari suatu prediksi.*N* nilai pada *confusion matrix* akan menujukan jumlah prediksi yang benar dan salah dalam suatu tabel dan akan dibandingkan dengan data real sebelum prediksi. Menurut Han dan Kamber (2006) nilai tabel *confusion matrix* yaitu :

***Tabel 3.3*** Table *cofusion matrix*

| | *Prediction class 1* | *Prediction class 2* |
|---|---|---|
| *Actual class 1* | *True Positive* | *False Negative* |
| *Actual class 2* | *False Positive* | *True Negative* |

Nilai yang ada pada tabel diatas menjelaskan indentifikasi dari suatu prediksi dimana TP dan TN merupakan hasil klasifikasi yang benar dari masing-masing kelas, sedangkan FP merupakan hasil identifikasi yang salah dimana seharusnya data masuk pada *class 1* tapi diidentifikasi masuk kedalam *class 2*, dan pada FN data yang seharusnya masuk pada *class 2* di identifikasi masuk kedalam *class 1*.

Adapun nilai evaluasi yang dihasilkan dengan menggunakan tabel *cofussion matrix* adalah :

1. Akurasi (*accuration*) adalah nilai akurasi dari suatu model.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

2. Precision, yang merupakan hasil persentase dari pelebelan yang benar dari kasus positif.

$$precison = \frac{TP}{TP + FP} \quad (2)$$

3. Recall (Sensitivity) adalah persentasekasus positif yang diiidentifikasi benar

$$recall = \frac{TP}{TP + FN} \quad (3)$$

Menentukan suatu model klasifikasi baik, buruk, atau sedang dapat dilakukan dengan menggunakan nilai kappa dengan skala nilai sebagai berikut (Landis dan Koch, 1977 ) :

**Tabel 3.4***Skala nilai Kappa*

| Skala | Kekuatan |
|---|---|
| < 0 | Buruk |
| 0.01 - 0.2 | Kurang bagus |
| 0.21 - 0.4 | Cukup bagus |
| 0.41 - 0.6 | Bagus |
| 0.61 - 0.8 | Sangat bagus |
| 0.81 - 1 | Mendekati sempurna |

### 3. 13*Receiver Operating Characteristic* (ROC)

*Receiver Operating Characteristic* atau ROC adalah suatu alat yang digunakan untuk mengevaluasi hasil klasifikasi dan visualisasi. ROC mampu

mengukur kinerja klasifikasi seperti akurasi, standar *error* dan juga memiliki kelebihan seperti grafik *recall* dan *liftcurve* (Fawcett,2005). Mengukur prediksi yang baik dengan menggunakan kurva ROC dari suatu model maka dilihat nilai *area under curve* (AUC) pada kurva ROC dengan skala sebagai berikut (Gorunescu,2011) :

***Tabel 3.5*** kualitas klasifikasi berdasarkan AUC

| Interval AUC | Kualitas model |
|---|---|
| 0.5 – 0.6 | Salah |
| 0.6 - 0.7 | Buruk |
| 0.7 – 0.8 | Cukup |
| 0.8 – 0.9 | Baik |
| 0.9 – 1 | Sangat baik |