BAB II
TINJAUAN PUSTAKA

2.1 Kanker Prostat

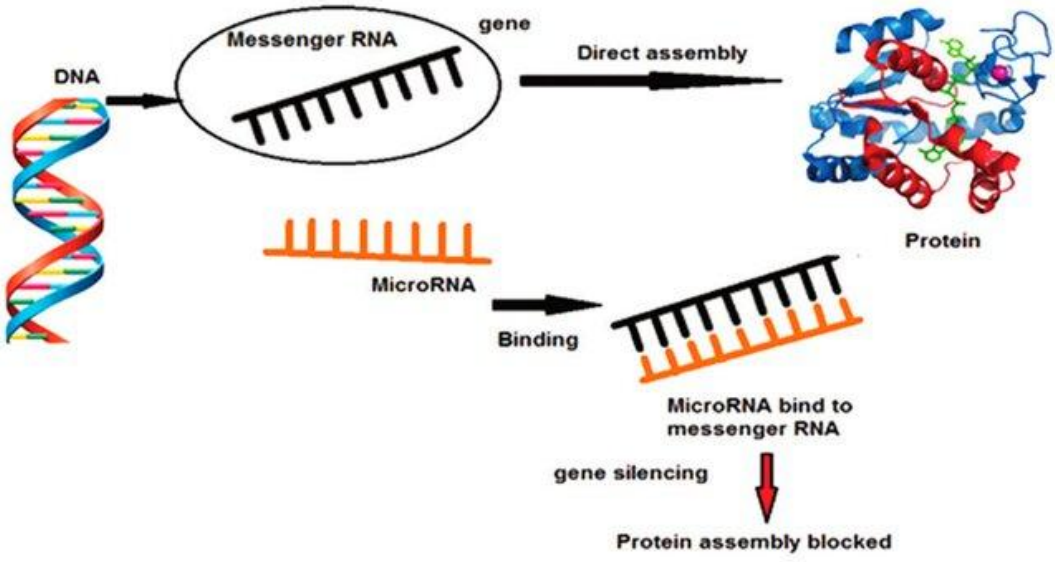
Kanker prostat adalah salah satu penyakit yang menyerang pria dan secara signifikan berkontribusi terhadap peningkatan angka kematian pada pria secara global. Penyakit kanker prostat biasanya menyerang pria dengan umur 45 hingga 60 [8]. Kanker prostat merupakan penyakit heterogen berdasarkan epidemiologi dan genetika. Pengaruh genetika dan lingkungan menjadi penyebab penurunan angka keselamatan pada ras spesifik terhadap terhadap kenker prostat. Salah satu dampak genetika adalah keturunan dari keluarga [9].

Bahaya dari penyakit kanker prostat harus segera ditangani dengan melakukan deteksi dini pada kanker prostat. Deteksi dini dalam kanker prostat dapat membantu pasien mendapatkan pengobatan yang lebih cepat dan akurat sehingga bisa dilakukan proses penyembuhan. Penyakit kanker yang semakin lama tidak ditangani maka akan semakin sulit untuk disembuhkan. *Prostate specific antigen* (PSA) merupakan salah satu cara untuk melakukan deteksi dini pada kanker prostate, namun faktor lainnya seperti kesalahan diagnosis masih menjadi kekhawatiran bagi pasien [10].

Kanker prostat juga memiliki tahap-tahap sendiri yaitu tahap 1 hingga 4. Semakin tinggi tahap pada kanker maka akan semakin bahaya dampaknya. Pada tahap 4 yang paling berbahaya biasa disebut tahap metastatik. Untuk mengetahui tahap pada penyakit kanker, dapat dilihat dari nilai T (tumor), N (*node*), M (Metastasis), PSA dan *Gleason score*. Nilai T menggambarkan penyebaran local tumor dalam penyakit kanker prostat, nilai N menunjukkan penyebaran pada kelenjar getah bening regional, nilai M menunjukkan apakah kanker sudah menyebar ke organ lainnya, PSA menunjukkan protein yang diproduksi oleh prostat apabila PSA semakin tinggi maka semakin berbahaya kanker prostat yang dialami, sama halnya dengan *gleason score* yang semakin tinggi maka semakin ganas kanker yang dialami.

2.2 Data miRNA

MicroRNA (miRNA) adalah kelompok molekul RNA kecil yang berukuran 19-25 nukleotida. Satu miRNA dapat mempengaruhi ekspresi dari setiap miRNA yang sering terlibat dalam jalur interaksi fungsional [11]. miRNA mengontrol proses biologis, seperti pembelahan sel, diferensiasi sel, angiogenesis, migrasi, apoptosis, dan onkogenesis [12]. Gangguan ekspresi miRNA pada sel kanker sering kali berakar pada lokalisasi miRNA yang mengkodekannya. Mereka sering berada di daerah yang secara genetik tidak stabil, situs yang rapuh atau daerah genom terkait kanker (CAGR), yang sering mengakibatkan penghapusannya, yang mengakibatkan kurangnya ekspresi miRNA [13]. miRNA juga memiliki konsep seperti *central dogma* yang penggambaran dapat dilihat pada Gambar 2.1. Konsep dari *central dogma* merupakan tahapan DNA akan dipecah hingga menjadi protein [14]. miRNA sendiri pada *central dogma* merupakan produk transkripsi dari DNA yang tidak dapat berproses untuk menjadi protein, tetapi menjadi pengatur miRNA yang mengatur pertumbuhan sel yang ada ditubuh. Sel yang bertumbuh secara tidak terkontrol akan mengakibatkan sel kanker berkembang didalam tubuh [15].



Gambar 2. 1 Central Dogma miRNA [16]

Dalam proses mendapatkan dataset miRNA per pasien dibutuhkan teknologi *next generation sequencing* (NGS) [17]. Platform yang digunakan untuk

melakukan NGS adalah Illumina Hiseq 2000. Tahap awal dalam melakukan RNA *seq* yaitu miRNA di ekstrak dari sampel biologi seperti *tissue* dan darah. Data yang sudah di ekstrak akan di konversi menjadi RNA yang lebih kecil dengan beberapa *library* seperti *3' Adapter Ligation, 5' Adapter Ligation, Reverse Transcription, PCR Amplification* dan *Size Selection*. Selanjutnya *bridge amplification* akan dilakukan yang bertujuan untuk memperbanyak setiap molekul menjadi kluster yang bisa dideteksi selama proses sekuensing. Pada tahap selanjutnya menggunakan *Sequencing by Synthesis* (SBS) yang merupakan teknologi utama dari illumina. Pada tahap ini dilakukan pada cDNA dengan menambahkan nukleotida satu per satu ke DNA. Selanjutnya output yang digunakan akan beru *FASTQ* format yang berisi *read sequence* dan *quality score*. Terakhir melakukan proses pada data yang sudah dinormalisasi menjadi *read per millions* (RPM) yang dapat digunakan untuk melakukan klasifikasi [18]. Proses normalisasi untuk mendapatkan RPM dapat dilihat pada persamaan (2.1). Langkah yang panjang dalam mendapatkan ekspresi pada miRNA membutuhkan biaya yang cukup mahal. Biaya untuk miRNA *sequencing* sekitar 200-600 tergantung sebarapa dalam *sequencing* yang ingin dilakukan [19].

$$RPM = \frac{Raw\ read\ count\ of\ a\ miRNA}{Total\ reads\ mapped\ in\ the\ sample} \text{ X } 1.000.000 \qquad (2.1)$$

2.3 Xgboost

Ensemble Learning merupakan pendekatan *machine learning* yang menggabungkan beberapa model untuk mendapatkan hasil yang lebih akurat dan optimal [20]. Dalam algoritma *Ensemble Learning* terdapat beberapa jenis seperti *Bagging* dan *Boosting*. *Bagging* bekerja dengan melatih model secara paralel pada subset data yang berbeda dan menggabungkan hasilnya untuk menghasilkan prediksi akhir, salah satu contoh dari *bagging* adalah algoritma Random Forest. Sementara itu, *boosting* membangun model secara berurutan, di mana setiap model baru berfokus pada kesalahan yang dibuat oleh model sebelumnya, salah satu contoh dari Boosting adalah algoritma Xgboost [21].

]Xgboost atau dikenal dengan *eXtreme Gradient Boosting* merupakan bentuk implementasi yang lebih lanjut dari *gradient boosting* [22]. Algoritma Xgboost disebut Extreme karena menggunakan regularisasi untuk mencegah *overfitting*. Tidak seperti algoritma LSTM, Xgboost bukan merupakan ANN melainkan sebuah ensemble dari *decision tree*. Xgboost merupakan salah satu metode yang sering digunakan untuk melakukan model prediksi karena akurasi, efisiensi dan dapat beradaptasi pada data yang ada [23]. Selain itu algoritma Xgboost dapat digunakan untuk klasifikasi menggunakan 2 kelas yang dapat membantu mendapatkan akurasi dari model yang dibuat. Xgboost dapat melakukan klasifikasi pada data miRNA dengan kelas *early stage* dan *late stage*. Dalam beberapa kasus, Ogunleye menunjukkan performa yang bagus menggunakan Xgboost untuk penyakit hati dengan mendapatkan akurasi dan sensitivitas yang tinggi [24].

$$Obj(\theta) = \sum_{i=1}^{n} l?\left(yi, \hat{y}i^{(t)}\right) + \sum_{k=1}^{t} \Omega\,(fk) \qquad (2.2)$$

Xgboost memiliki beberapa prosedur yang dijalankan untuk menentukan model dari *machine learning*. Pada pembuatan model, algoritma Xgboost akan memulai proses dengan cara menghitung prediksi awal pada data yang ada. Setelah itu Xgboost akan menghitung error dari nilai yang di prediksi dan nilai asli dari data yang sudah dihitung. Error yang ditemukan akan dibenarkan dengan membuat pohon keputusan. Proses tersebut akan diulang terus menerus hingga mendapatkan hasil yang baik. Dalam Xgboost cara menghitung model dapat dilihat pada persamaan (2.2). Persamaan tersebut menunjukkan fungsi objektif utama pada Xgboost yang menggabungkan akurasi model melalui loss dan kompleksitas model melalui regulasi. Pseudocode pada XGBoost dapat dilihat pada Algoritma 2.1.

Algortima 2.1 Pseudocode XGBoost

1. *Initialize predictions*:

$$\hat{y}_i^{(0)} = 0, for\ all\ i = 1 \ldots m$$

2. *for t = 1, …, T*:

a. *For each sample i, compute:*

Gradient: $g_i = \frac{\partial L(y_i, \hat{y}_i)}{\partial \hat{y}_i}$

Hessian: $h_i = \frac{\partial^2 L((y_i, \hat{y}_i)}{\partial \hat{y}_i^2}$

b. *Fit a regression tree* $h_t(x)$ *to the data using* $\{g_i, h_i\}$

each plit is chosen to maximize the gain:

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma$$

c. *Assign leaf values:*

$$w_j = \frac{G_j}{H_j+\lambda}$$

(where G_j, H_j *are sums of gradients and hessians in leaf j)*

d. *Update Predictions:*

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta \cdot h_t(x_i)$$

3. *end*

2.4 Seleksi Fitur

2.4.1. Lasso + RFE

Seleksi fitur merupakan sebuah Teknik yang digunakan untuk mendapatkan data miRNA yang lebih signifikan yang akan digunakan dalam model sehingga mendapatkan akurasi yang lebih baik. Dalam seleksi fitur terdapat seleksi fitur dalam bentuk statistik dan bioinformatika, salah satu contoh dari statistik berupa *Lasso* dan *Recursive Feature Elimination* (RFE) dan *bioinformatics* berupa EdgeR [25]. Dalam konteks statistik, seleksi fitur bertujuan untuk mengidentifikasi variabel independen yang memiliki kontribusi signifikan terhadap variabel target. RFE merupakan Teknik seleksi fitur yang menghapus fitur yang paling tidak

signifikan di data, proses tersebut akan diulang terus menerus hingga mendapatkan hasil fitur yang paling signifikan berdasarkan kondisi yang ditentukan. RFE sendiri merupakan Teknik seleksi fitur dengan pendekatan wrapper yang menggunakan algoritma pembelajaran untuk menilai kombinasi fitur [26]. RFE adalah sebuah seleksi fitur yang sering digunakan pada data ekspresi miRNA yang dapat mengurangi jumlah fitur menjadi fitur yang signifikan. Selain itu ada juga seleksi fitur *Lasso* menggunakan pendekatan *embedded*, *lasso* menggunakan *ridge regression* dengan koefisien L2 dan L1 regularisasi [27]. *Lasso* juga dapat digunakan untuk mengurangi fitur dengan menambahkan *penalty* terhadap koefisien regresi sehingga koefisisen menjadi 0 dan dikeluarkan dari fitur. Rumus yang digunakan untuk seleksi fitur *Lasso* dapat dilihat pada persamaan (2).

$$Loss\ Function = RSS + \alpha \sum_{j=1}^{p} \left|\beta_j\right| \tag{2.3}$$

2.4.2. EdgeR

EdgeR merupakan Teknik seleksi fitur bioinformatika yang digunakan dalam data miRNA. Dengan menggunakan edgeR, output yang sering digunakan berupa logFC dan nilai P. LogFC merupakan perubahan ekspresi miRNA yang semakin besar maka semakin tinggi terkena kanker yang dapat digunakan sebagai Batasan untuk memilih fitur. Nilai *Pvalue* merupakan signifikansi dari data miRNA yang semakin kecil maka akan semakin signifikan [28]. Pada pencarian logFC dapat dicari menggunakan rumus pada persamaan (3), sedangkan rumus untuk mencari nilai *Pvalue* dapat dilihat pada persamaan (4).

$$\log\left(\mu_{gi}\right) = \log(N_i) + x_i^T \beta_g \tag{2.4}$$

$$LRT = 2[log\ L_{full} - log\ L_{reduced}] \tag{2.5}$$

2.5 Matrik Evaluasi

Matrik evaluasi merupakan matrik yang digunakan untuk mendapatkan akurasi, presisi, *recall* dan *f1-score*. Matrik tersebut dapat digunakan untuk menentukan seberapa baik model yang sudah digunakan. Setiap matrik yang ada memiliki caranya sendiri untuk mendapatkan hasilnya [29]. Dalam akurasi dapat dilihat pada persamaan (5), akurasi didapat dari *true positive* ditambah dengan *true negative* lalu dibagikan dengan semua matrik yang ada.

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.6}$$

Presisi merupakan salah satu matrik yang perhitungan dapat dilihat pada persamaan (6), presisi dapat dihitung dengan membagikan *true positive* dengan total dari *true positive* ditambah dengan *false positive*.

$$Presisi = \frac{TP}{TP + FP} \tag{2.7}$$

Perhitungan dari *recall* dapat dilihat pada persamaan (7) yang mirip dengan presisi yang membagi *true positive* dengan total dari *true positive* ditambah *false negative*.

$$Recall = \frac{TP}{TP + FN} \tag{2.8}$$

Matrik terakhir yang digunakan adalah *f1-score* yang membutuhkan hasil dari presisi dan *recall*. Semakin bagus hasil dari presisi dan *recall* maka hasil dari f1-score akan semakin baik. Perhitungan untuk *f1-score* dapat dilihat pada persamaan (8).

$$F1\ Score = \frac{2 * Presisi * Recall}{Presisi + Recall} \tag{2.9}$$

Dalam matrik evaluasi, jumlah dari *true positive*, *true negative*, *false positive* dan *false negative* dapat dilihat dari *confusion matrix*. *Confusion matrix* merupakan

table yang digunakan untuk mengevaluasi kinerja model yang sudah diklasifikasi sehingga dapat membantu untuk mendapatkan hasil perhitungan dari akurasi, presisi, *recall* dan *f1-score* lebih jelas.

UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA