IMPLEMENTASI MUTUAL INFORMATION DAN BAYESIAN NETWORK UNTUK KLASIFIKASI DATA MICROARRA

Mohamad Syahrul Mubarok[1], Mahendra Dwifebri Purbolaksono[2], Adiwijaya[3]

[1,2,3]Proram Studi S1 Teknik Informatika, Fakultas Informatika, Universitas Telkom

Email :

Abstract.Kanker merupakan salah satu penyebab utama kematian seseorang diseluruh dunia. Menurut data WHO, ada sekitar 14 juta kasus kanker baru di tahun 2012. Karena hal itu pengawasan sejak dini dibutuhkan guna mencegah pertumbuhan kanker. Selain itu pendeteksian secara dini juga merupakan hal yang dibutuhkan. Salah satu cara mendeteksi yaitu melalui ekspresi gen. Ekspresi gen merupakan metode ekstrasi gen menjadi data yang menjadi data bernama *microarray*. Data *microarray* memungkinkan terjadi proses pengklasifikasi secara langsung namun atribut dalam suatu record sangat besar sehingga memakan waktu komputasi yang lama. Karenanya dibutuhkan sistem yang dapat menyelesaikan masalah tersebut. Pada penelitian ini, sistem menggunakan pendekatan machine learning yaitu Bayesian Network. Sedangkan untuk seleksi fitur menggunakan Mutual information. Hal ini berguna untuk mengurangi attribute yang terlalu banyak. Untuk pengukuran menggunakan F1-score. Sistem yang dibangun mampu mengklasifikasi kanker dengan f1-score tertinggi mencapai 91.06%.

Keyword: Bayesian Network, Mutual Information, Microarray.

1. Pendahuluan

Kesehatan adalah hal yang paling penting dalam kebutuhan hidup setiap insan manusia. Banyak penyakit ataupun wabah yang bisa menyerang sistem kekebalan tubuh kita sewaktu-waktu. Salah satu penyakit menakutkan adalah kanker. Kanker adalah penyebab utama kematian seseorang diseluruh dunia, terhitung 8,2 juta jiwa meninggal dalam tahun 2012 saja [1]. Kematian akibat kanker bisa dicegah jika terdeteksi sejak dini [2]. Untuk itu waspada sejak dini adalah hal yang wajar dilakukan dalam hal mencegah pertumbuhan kanker ini. Mendeteksi adanya kanker perlu dilakukan agar lebih waspada dalam menjaga kesehatan dan konsumsi makanan. Beberapa cara dilakukan untuk mendeteksi adanya kanker dalam tubuh manusia, salah satunya yaitu melalui ekspresi gen yang terdapat setiap manusia.

Pada tubuh manusia mempunyai ribuan gen. Gen tersebut mempunyai keunikan sendiri dalam setiap manusia. Ekspresi gen atau kumpulan gen tersebut bisa diekstrak menjadi data yang mempunyai jutaan elemen dan puluhan ribu atribut. Ekstrasi dari kumpulan gen atau ekspresi gen yang sangat panjang dan bisa mencapai puluhan ribu dimensi disebut *DNA microarray technology.* Dari ekspresi gen itu kita bisa simpan dalam file berupa kumpulan data float. Kumpulan data float itu disebut juga *array of float.* Data yang didapatkan tersebut akan diklasifikasi menurut kelas

masing-masing. Klasifikasi disini akan menentukan apakah gen tersebut terjangkit kanker atau tidak. Dari gen kita bisa menganalisis apakah gen tersebut normal atau mengandung sel kanker.

Dari data *microarray* kita bisa saja langsung mengklasifikasikannya. Tetapi pada penelitian kali kita tidak akan menggunakan semua attribut yang ada dikarenakan banyaknya attribut dalam suatu *record* yang bisa membuat akurasi dan performansinya menurun. Sehingga diperlukan *feature selection* untuk memilih atribut mana saja yang penting dalam mengklasifikasi kedalam kelas tertentu. Pada tugas akhir ini, akan digunakan *Mutal Information* untuk menjadi algoritma *feature selection. Mutual Information* akan menghasilkan inputan yang lebih berpengaruh kepada proses klasifikasi. *Mutual Information* mengukur berapa banyak informasi atau atribut tersebut berperan untuk membuat klasifikasi benar didalam *class* manapun.

Dari proses *feature selection* tersebut akan dihasilkan atribut-atribut yang akan diklasifikasi. Pada tugas akhir ini, akan digunakan algoritma *Bayesian Networks. Bayesian networks* merupakan suatu metode pemodelan data berbasis probabilitas yang merepresentasikan suatu himpunan variabel dan *conditional dependency*-nya melalui suatu *Directed Acyclic Graph* (DAG) [3]. *Bayesian networks* digunakan karena setidaknya ada empat alasan. Pertama *Bayesian Networks* bisa mengatasi dataset yang tidak lengkap, kedua *Bayesian Networks* memungkinan untuk membaca keterkaitan antar atribut atau variabel, ketiga *Bayesian Networks* sejalan dengan teknik *bayesian* statistik yang memfasilitasi kombinasi antara data dan domain knowledge, dan yang terakhir *Bayesian Networks* menyediakan jalan yang efisien untuk menghindari data yang bersifat *over fit* [4].

Data yang akan digunakan dalam tugas akhir ini diambil dari suatu repositori data yaitu Kent Ridge Bio-medical Dataset yang bisa diakses melalui web http://datam.i2r.a-star.edu.sg/datasets/krbd/. Pada repository data tersebut terdapat sembilan kelas utama. Dalam dataset tersebut sudah tersedia data *trainset* dan *testset*. Karena dataset yang bisa dikatakan sedikit maka akan dilakukan proses *cross validation* untuk memenuhi kebutuhan klasifikasi ini. *Cross Validation* atau disebut juga validasi silang adalah membagi data menjadi dua bagian, yaitu data pelatihan (*trainset*) dan data pengujian (*testset*) yang selanjutnya data *trainset* akan digunakan juga untuk data tes juga dan sebaliknya.

2. Dasar Teori

2.1. Penelitian Terkait

Pada penelitian terkait dengan judul *Dimensionilty Reduction using Principal Component Analysis for Cancer Detection Based on Microarray Data Classification* yang menggunakan metode SVM (Support Vector Mechine) dan BPLM (Backpropagation – Levvenberg Marquardt) untuk *classifier*. Dan untuk reduksi dimensi menggunakan metode PCA. Pada penelitian tersebut menggunakan 6 dataset yaitu Leukimia, Ovarian, Central Nervous, Colon, Lung dan Prostate. Hasil klasifikasi rata-rata menggunakan BPLM sebesar 96.07%, sedangkan SVM 94.98%.

2.2. Microarray Data

Gen adalah segmen atau bagian dari DNA yang beri informasi yang penting yang terbuat dari protein di dalam tubuh kita [5]. Ekspresi *microarray* adalah percobaan yang menyimpan data yang terdiri dari ribuan gen secara serentak. Percobaan ini lebih fokus untuk memantau setiap gen berulang-ulang dalam kondisi yang berbeda pula atau mengevaluasi setiap gen dalam sebuah lingkungan yang sama tetapi dalam jaringan yang berbeda terutama fokus ke jaringan kanker [6] [7]. Ekspresi *microarray* adalah percobaan yang sangat berpotensial untuk bagian dari standar diagnosa tes yang dilakukan dikalangan ilmu kesehatan [8].

Dalam kasus *microarray* memiliki masalah dengan banyaknya atribut atau variabel yang dihasilkan. Dalam proses klasifikasi ini akan dicari atribut yang penting-penting saja untuk

dijadikan inputan. Dari inputan itu akan dihasilkan output berupa kelas yang telah ditetapkan. Lalu hasil kelas tersebut dijadikan diagnosis apakah gen tersebut terjangkit penyakit itu atau tidak.

Untuk di penilitian ini kita akan menggunakan lima dataset yaitu:

Tabel 2-1 Distribusi Data pada Dataset Kent Ridge Bio-medical [8]

Data Set	Instances	Genes	Classes
Breast-cancer	97	24481	2 (Relapse & Nonrelapse)
Colon-cancer	62	2000	2 (Positive & Negative)
Leukemia ALL-AML	72	7129	2 (ALL & AML)
Ovarian-cancer	253	15154	2 (Cancer & Normal)
Lung-cancer	181	12533	2 (Mesothelioma & ADCA)

2.3. Mutual Information

Salah satu masalah utama dari teknik analisis ekspresi gen ini adalah dimensi dari gen itu sendiri [9]. Pemilihan gen yang relevan untuk klasifikasi sampel adalah tugas umum dalam kebanyakan ekspresi gen [10]. Dalam penelitian ini, *Mutual Information* (MI) [11] adalah teknik yang digunakan untuk memilih gen informatif dari keseluruhan ekspresi gen asli. Untuk menghitung MI, distribusi probabilitas dari gen yang diperlukan dalam prakteknya belum diketahui, dan yang bisa dilakukan adalah dengan menggunakan histogram dari data. Langkah-langkah yang terlibat dalam menghitung MI dari histogram data pelatihan adalah seperti dibawah ini:

- Data set di urutkan secara *ascending* berdasarkan *output*
- *Output* akan diberikan label (C) yang akan dibagi menjadi dua kelompok dan H(C) akan dihitung menggunakan:

$$H(Y) = -\sum_{j=1}^{Ny} P(Yi).\log(P(Yi)) \quad (2.1)$$

- *Input* akan diberikan label (X) akan dibagi menjadi sepuluh kelompok dan H(C|X) akan dievalusi menggunakan:

$$H(Y|X) = -\sum_{i=1}^{Nx} P(Xi).\sum_{j=1}^{Ny} P(Yj|Xi).\log(P(Yj|Xi)) \quad (2.2)$$

- Selanjutnya, MI dari setiap gen yang berhubungan dengan output akan dihitung menggunakan:

$$I(Y;X) = H(Y) - H(Y|X) \quad (2.3)$$

MI akan diurutkan berdasarkan yang terbaik (*ascending*). Yang pertama adalah atribut dengan nilai MI yang paling tinggi dan akan dipilih sebagai gen yang informatif untuk data latih dengan dukungan vektor [12].

2.4. Klasifikasi

Pada *machine learning* [13] klasifikasi dianggap sebagai sebuah contoh dari pembelajaran *supervised*. Proses pembelajaran tersebut dilakukan dengan memanfaatkan *training set*. Pada prosedur yang tidak memerlukan *training set* adalah proses *clustering*.

Kembali pada klasifikasi, dua tahap utama pada proses klasifikasi dapat dilihat kedalam bidang statistik. Tahap pertama merupakan tahap klasik yang berfokus pada turunan dari teori Fisher mengenai diskriminan linear. Kedua adalah tahap modern yang mengeksploitasi kelas dan model secara lebih fleksibel, dimana usaha untuk menghitung *joint distribution* dari fitur-fitur pada setiap kelas, yang nantinya dapat digunakan pada aturan klasifikasi.

Pada pendekatan statistik pada umumnya digolongkan memiliki model probabilitas yang eksplisit, dimana memungkinkan probabilitas berada pada setiap kelas dari pada hanya satu kelas [14].

Contoh aplikasi dari *classification* antara lain [15]:

a. Klasifikasi *genre*, topik, kategori sebuah dokumen
b. *Spam filtering*
c. *Language identification*
d. Identifikasi umur dan jenis kelamin dari penulis sebuah dokumen
e. Sentimen analisis

Metode klasifikasi yang dapat digunakan adalah [15]:

a. Naive Bayes
b. Logistic Regression
c. K-Nearest Neigbors (KNN)
d. Support-vector Machines (SVM)
e. Max Enthropy

2.5. K-Means

K-means adalah metode *clustering*, *clustering* sendiri adalah proses pembagian kelompok data sesuai dengan distribusinya masing-masing. Komponen utama *K-means* adalah *centroid*, *centroid* adalah nilai yang mempunyai bentuk yang sama dengan data yang ada, akan tetapi mempunyai nilai yang berbeda dengan data, *centroid* merupakan representasi titik tengah dari kelompok data.

Untuk memperoleh *cluster* yang dapat mengelompokkan data sesuai distribusi, *K-means* memiliki persamaan utama sebagai berikut:

$$J = \sum_{i=1}^{k} \sum_{j=1}^{n} ||x_i - c_j||^2 \quad (2.4)$$

Di mana J merepresentasikan kumpulan *centroid* yang selalu di perbarui hingga tercapai *cluster* yang tepat, yang dalam hal ini berdasarkan pada penilaian manusia. Data yang akan di kelompokan direpresentasikan sebagai x, dengan k merepresentasikan banyaknya *cluster*. *Centroid* direpresentasikan sebagai c dengan n adalah banyaknya c. dari persamaan tersebut $||x_i - c_j||^2$ merepresentasikan ukuran jarak antara data x dengan *cluster centroid c.*

2.6. Bayesian Network

Bayesian Networks berasal dari teorema Bayes, teorema Bayes adalah sebuah pendekatan untuk sebuah ketidakpastian yang diukur dengan probabilitas. Teorema ini dikemukakan oleh Thomas Bayes dengan rumus dasar:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \quad (2,5)$$

Berdasarkan Heckerman [4] *bayesian networks* merupakan model *graphical* untuk hubungan probabilitas antara variabel. Menurut Kjaerulff [3] *bayesian networks* merupakan sebuah *Probabilistic Graphical Model* (PGM) sederhana yang dibangun oleh dua teori yaitu teori probabilistik dan teori graf. Teori probabilistik berhubungan langsung dengan data sedangkan teori graf berhubungan langsung dengan bentuk representasi yang ingin didapatkan. Informasi tersebut direpresentasikan secara kualitatif menggunakan struktur graf dan secara kuantitatif menggunakan parameter-parameter numerik [16].

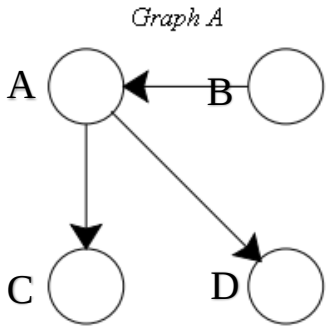
Bayesian networks adalah gambaran yang paling umum atau global dari teorema Bayes. *Naïve Bayes* dan *Hidden Markov Model* adalah turunan dari *Bayesian Network.* Hal ini disebabkan karena DAG (*Directed Acyclic Graph*) dalam *Bayesian Network* tidak ada bentuk khusus seperti *Naïve Bayes* dan *Hidden Markov Model.* Di dalam DAG terdiri dari *node* dan *edge.* Pada setiap *node* terdapat CPD (*Conditionally Probabilistic Distribution*) atau probabilitas dari atribut atau variabel yang ada di *node* tersebut *given* atribut orang tuanya. Untuk *Edge* sendiri adalah untuk menunjukan bahwa *node* tertentu mempunyai anak atau orang tua. Jika *node* A misalnya menunjuk ke arah *node*

B berarti *node* A adalah orang tua dari *node* B atau *node* A mempunyai anak *node* B. dari CPD tersebut bisa untuk mencari JPD (*Joint Probability Distribution*) yang akan bisa melihat dari semua keterkaitan atribut pertama sampai atribut terakhir.

Menurut Pearl [17] *Bayesian networks* terdiri dari dua komponen yaitu:

1. Struktur graf *bayesian networks* disebut dengan *Directed Acyclic Graph* (DAG) yaitu graf berarah tanpa siklus berarah. DAG terdiri dari *node* dan *edge*. *Node*merepresentasikan variabel acak dan *edge* merepresentasikan adanya hubungankebergantungan langsung dan dapat juga diinterpretasikan sebagai pengaruh (sebab-akibat) langsung antara variabel yang dihubungkannya. Tidak adanya *edge* menandakan adanya hubungan kebebasan kondisional di antara variabel.

   Struktur grafis *bayesian networks* ini digunakan untuk mewakili pengetahuan tentang sebuah domain yang tidak pasti. Secara khusus, setiap *node* dalam grafik merupakan variabel acak, sedangkan ujung antara *node* mewakili probabilistik yang bergantung di antara variabel-variabel acak yang sesuai. Kondisi ketergantungan ini dalam grafik sering diperkirakan dengan menggunakan statistik yang dikenal dengan metode komputasi. Oleh karena itu, *bayesian networks* menggabungkan prinsip-prinsip dari teori graf, teori probabilitas, ilmu pengetahuan komputer, dan statistik.
2. Himpunan parameter mendefinisikan distribusi probabilitas kondisional untuk setiap variabel. Pada *bayesian network*, *node*s berkorespondensi dengan variabel acak. Tiap *node* diasosiasikan dengan sekumpulan peluang bersyarat, p(xi|Ai) sehingga xi adalah variabel yang diasosiasikan dengan *node* dan Ai adalah set dari parent dalam graf.



Gambar 2-1 Contoh Struktur Sederhana Bayesian Networks [17]

Secara umum, jika terdapat *node* U = U_i…Un, maka fungsi *joint probability* untuk setiap *Bayesian Networks* adalah sebagai berikut:

$$P(U) = \prod_i^n P(P_k|parents(P_k)) \quad (2.6)$$

Berdasarkan rumus 5, dapat diketahui bahwa U adalah representasi suatu *node* yang terdapat pada *Bayesian Networks*, maka P(U) merupakan perkalian dari semua probabilitas bersyarat yang terdapat pada struktur *graph* dan *parents*(P_k) merupakan *parent* dari *node* P_k.

$$P(A,B,C,D) = P(A).P(B|A).P(C|B)..P(C|B) \quad (2.7)$$

Persamaan 6 adalah perhitungan untuk *joint probability* yang terdapat pada Gambar 1 *graph* A.

2.7. F1-Measure

Evaluasi sistem dilakukan setelah *classifier* melakukan klasifikasi terhadap data *testing*. Evaluasi sistem ini dilakukan untuk melihat seberapa bagus performansi yang dihasilkan oleh *classifier*. Pada tugas akhir ini, performansi sistem akan diukur dengan nilai akurasi, *precision, recall*, dan *F1-Measure* [18].

a. Akurasi, *Precision*, dan *Recall*

Akurasi merupakan rasio ketepatan sistem dalam mengklasifikasikan dokumen ke kelasnya yang sesuai. Precision merupakan rasio perbandingan jumlah dokumen relevan yang ditemukan dengan total jumlah dokumen yang ditemukan oleh *classifier*. *Precision* mengindikasikan kesesuaian antara kelas aktual dan kelas prediksi dari *classifier*. *Recall* adalah rasio perbandingan jumlah dokumen relevan yang ditemukan kembali dengan total jumlah dokumen dalam kumpulan dokumen yang dianggap relevan. *Precision* dan *Recall* dapat dihitung dengan membuat tabel *confusion matrix* terlebih dahulu. Contoh tabel *confusion matrix* dapat dilihat pada Tabel 2-1 dan persamaan untuk menghitung nilai akurasi, *precision*, dan *recall* dapat dilihat pada persamaan 8, 9, dan 10.

Tabel 2-2 Confusion Matrix

Classifier \ Actual	Actual Positive	Actual Negative
Classified Positive	True Positive	False Positive
Classified Negative	False Negative	True Negative

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.8)$$

$$Precision = \frac{TP}{TP + FP} \quad (2.9)$$

$$Recall = \frac{TP}{TP + FN} \quad (2.10)$$

Keterangan:

1) *True Positive*(TP) merupakan *positive class* (kelas yang ingin dievaluasi) yang terklasifikasikan dengan benar oleh sistem klasifikasi.
2) *True Negative*(TN) merupakan *negative class* (kelas selain yang ingin dievaluasi) yang terklasfikasikan dengan benar oleh sistem klasifikasi.
3) *False Positive*(FP) merupakan *negative class* yang terklasifikasikan oleh sistem klasifikasi sebagai *positive class*.
4) *False Negative*(FN) merupakan *positive class* yang terklasifikasikan oleh sistem klasifikasi sebagai *negative class*.

b. *F1-Measure*

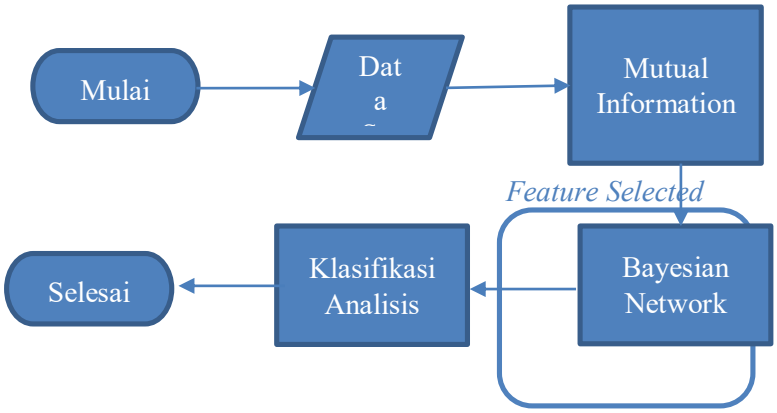
Selain *precision* dan *recall*, performansi sistem juga dapat dihitung dengan menggunakan *F1-Measure*. *F1-Measure* merupakan *harmonic mean* (*weighted harmonic mean*) dari *precision* dan *recall* yang dapat ditulis dengan persamaan 2.9.

$$F1 = \frac{2.Precision.Recall}{Precision + Recall} \quad (2.11)$$

3. Pembahasan

3.1. Deskripsi Sistem

Sistem ini mengimplementasikan pendekatan klasifikasi yakni dengan *Bayesian Network*. Sistem akan mengolah data dari *repository* yang kemudian akan diambil dengan *cross validation* atau *cleansing* data jika diperlukan. Setelah didapat data training akan dicari atribut penting menggunakan *feature selection* dengan algoritma *Mutual Information*. *Mutual Information* akan menghasilkan atribut atau inputan untuk proses klasifikasi. Klasifikasi algoritma akan menggunakan Bayesian Network. Adapun proses rancangan sistem digambarkan pada Gambar 1 berikut.



Gambar 3-1 DAG Rancangan Bayesian Network

Berdasarkan gambar diatas dapat dijelaskan bahwa:

- Data Set, dalam proses ini dilakukan pengumpulan data dari sumber yang sudah tersedia. Data Set terbagi menjadi dua yaitu dataset *training* dan dataset *testing.*
- *Mutual Information*, ini adalah salah satu jenis *feature selection* yang akan menghasilkan atribut atau input untuk dilakukan klasifikasi.
- Bayesian Network, disini akan dilakukan proses klasifikasi menggunakan metode tersebut tetapi sebelumnya akan dilakukan *cross validation* terlebih dahulu. Pada proses *training* akan dilakukan proses *Learning* terhadap datasetnya. Pada proses *testing* akan dilakukan proses pengklasifikasian atau menghasilkan *class* yang sudah ditargetkan.
- Klasifikasi Analisis adalah proses setelah klasifikasi dilakukan yaitu untuk mencari hasil performasi dan akurasi dari semua proses yang sudah dijalankan. Pada proses *training* akan dianalisis hasil *learning*nya dan pada proses *testing* akan menghasilkan kelas-kelas penyakit dan hasil performansi *precision*, *recall*, dan *f-measure.*

3.2. Penjelasan Rancangan Sistem

3.2.1. Mutual Information (MI)

Mutual Information adalah salah satu metode untuk *feature selection* yang akan digunakan dalam penelitian ini. *Feature selection* sendiri berguna untuk mencari fitur atau variabel yang terdapat dalam data *microarray* agar variabel yang nantinya akan jadi input lebih informatif dan efektif. Hal ini dilakukan karena dalam data *microarray* sendiri terdapat ribuan variabel. Sebelum melakukan proses seleksi fitur dibutuhkan data yang berbentuk diskrit atau bulat.

Proses diskritisasi merupakan proses yang perlu dilalui dikarenakan akan menyulitkan perhitungan histogram apabila *parent* dan *child* memiliki distribusi yang berbeda. Dalam hal ini *parent* memiliki distribusi diskrit dan *child* memiliki distribusi continu. Proses diksritisasi dalam sistem ini menggunakan *K-means*, diskritisasi ini akan menghasilkan bilangan antara satu sampai dengan K. K adalah jumlah *cluster* yang diobservasi.

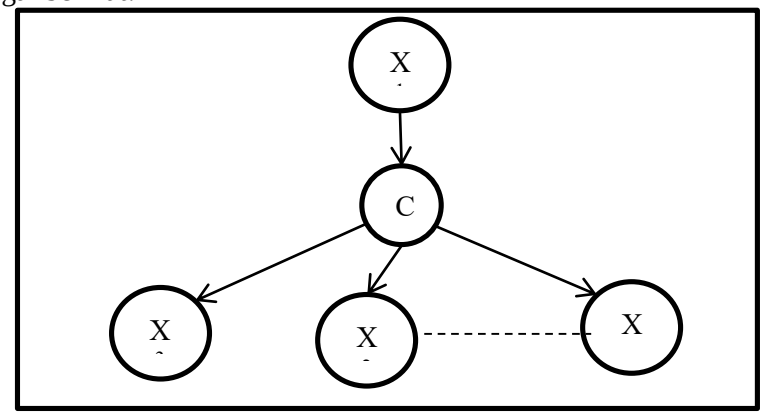
MI akan diproses secara perhitungan histogram. Proses histogram terjadi karena adanya proses pengurutan secara *ascending*. Nilai *output* (C) akan dibagi menjadi dua kelompok dan dihitung menggunakan rumus (2.1) pada bab 2. Sedangkan nilai *input* (X) akan dibagi menjadi sepuluh kelompok dan dievaluasi menggunakan rumus (2.2) pada bab 2. Preses selanjutnya adalah setiap gen atau *input* (X) yang berhubungan dengan *output* (C) akan dihitung menggunakan rumus (2.3) pada bab 2. Lalu nilai yang didapatkan akan diurutkan berdasarkan nilai yang tertinggi. Nilai-nilai yang dibagian atas akan dipilih sebagai variabel *input* pada proses klasifikasi sebagai gen yang informatif.

3.2.2. Bayesian Network (BN)

Pada penelitian kali ini akan digunakan metode Bayes tepatnya adalah *Bayesian Network* (BN). BN bias melakukan proses klasifikasi dengan menggunakan probabilitas. Dari data yang dihasilkan oleh *feature selection* yaitu MI nantinya akan dijadikan inputan atau variabel yang bias mempengaruhi kelas yang akan ditentukan. Nilai pada setiap variabel yang ada merupakan suatu nilai real antara 0 sampai dengan 1.

Bayesian Network (BN) sendiri adalah suatu metode Bayes yang menyeluruh. Semua variasi DAG (*Directed Acyclic Graph*) yang bias bermacam-macam asalkan masih tetap pada aturan DAG itu sendiri. DAG terdiri dari *node* dan *edge*. *Node* adalah variabel-variabel yang terdapat pada model. Pada setiap *node* terdapat CPD (*Conditionally Probabilistic Distribution*) atau nilai probabilitas pada variabel pada *node* yang dipengaruhi oleh variabel orang tuanya. Sedangkan *edge* adalah anak panah yang menunjuk dari *node* ke *node* yang lain, dimana *node* asal adalah orang tua dari *node* yang ditunjuk. DAG ini tidak boleh *Cyclic* atau berputar kembali. Pada perhitungan setiap *node input* nantinya akan mempengaruhi hasil dari kelas yang akan dicari.

Pada penelitian Tugas Akhir ini sendiri terdapat gambaran umum dari DAG yang akan dibangun sebagai berikut:

4. 

Gambar 3-2 DAG Rancangan Bayesian Network

Pada DAG diatas menunjukan C adalah variabel class *output* yang akan dituju dan X1 sampai dengan Xn adalah variabel *input*. Antar inputan bisa saja saling terkait seperti X1 dan X2, jadi semua variabel dapat dipengaruhi oleh variabel lain termasuk variabel *output* (C). Nilai X1 akan

digantikan dengan X2 dan X1 akan berpindah dibawah, begitu seterusnya sampai n atau variable yang terakhir.

3.2.3. Pengukuran Performansi *Classifier*

Pengukuran performansi sistem dilakukan dengan menggunakan metode perhitungan yaitu akurasi *precision, recall* dan *f-measure* [19]. Akurasi adalah ketepatan sistem melakukan proses klasifikasi dengan benar.

3.2.4. Spesifikasi Perangkat Keras

Spesifikasi perangkat keras yang digunakan dalam pembangunan sistem adalah sebagai berikut:

1. Processor Intel (R) Core (TM) i5 CPU M 4210U 2.7 Ghz
2. RAM 12 GB
3. VGA Nvidia 820M

3.2.5. Spesifikasi perangkat Lunak

Spesifikasi perangkat lunak yang digunakan dalam pembangunan sistem adalah sebagai berikut:

1. Sistem operasi Microsoft Windows 10
2. Netbeans 8.0.2
3. Java

4. Hasil Pengujian dan Analisis

4.1. Pengujian Sistem

Dataset yang digunakan adalah Kent Ridge Bio-medical [20] dan berfokus pada distribusi data *Breast Cancer*, disana tersedia data *training* dan dan data *testing.*

4.1.1. Tujuan Pengujian

Pengujian sistem dilakukan dalam rangka memenuhi tujuan penelitian. Adapun tujuan pengujian sistem yang dilakukan adalah sebagai berikut:

1. Menerapkan sistem yang dibangun dengan *Bayesian Network* sebagai *classifier* dan *Mutual Information* sebagai *feature selector* dalam melakukan klasifikasi ekspresi gen untuk mendeteksi kanker dari *Kent Ridge Bio-medical Dataset*.
2. Menganalisis performasi dari algoritma Bayesian Network dalam mengklasifikasi dan metode Mutual Infromation dalam seleksi fitur data microarray menggunakan F1 Measure

4.2. Analisis Hasil Pengujian

4.2.1. Analisis Pengaruh Nilai K PADA

Pengujian dilakukan dengan menggunakan lima nilai k yaitu 2, 8, 10, 12 dan 16. Berikut merupakan perbandingan nilai performansi sistem maksimal. Tujian dari pengujian ini adalah untuk mengetahui nilai k mana yang terbaik pada saat proses diskritisasi. Nilai k akan mempengaruhi besarnya persebaran data pada tiap nilai. Semakin besar nilai k maka jumlah variasi nilai akan semakin banyak. Maka otomatis akan mempengaruhi banyaknya parameter yang terdapat pada tiap *Conditional Probability Table* (CPT). Dengan demikian performansi dari classifier akan dipengaruhi oleh nilai k.

Berikut pada Tabel 4-6 adalah detail hasil dari klasifikasi berdasarkan pengaruh nilai K.

Tabel 4-1 Hasil Rata-rata dari Klasifikasi dengan nilai k yang berbeda

Evaluation		**k = 2**	**k = 8**	**k = 10**	**k = 12**	**k = 16**
Breast	**F1-Score**	0.684	0.631	0.84	0.631	0.631
	Precision	1.00	1.00	1.00	1.00	1.00
	Recall	0.666	0.631	0.8	0.631	0.631
Colon	**F1-Score**	0.733	0.867	0.867	0.8	0.733
	Precision	0.8	0.6	0.6	0.4	0.8
	Recall	0.7	1	1	1	0.7
Leukimia	**F1-Score**	0.882	0.765	0.765	0.882	0.588
	Precision	1	1	1	1	1
	Recall	0.833	0.714	0.714	0.833	0.588
Ovarian	**F1-Score**	0.950	0.967	0.984	0.984	0.598
	Precision	0.974	0.975	1	1	0.375
	Recall	0.95	0.975	0.975	0.975	1
Lung	**F1-Score**	0.973	0.973	0.980	0.980	0.980
	Precision	0.867	0.867	0.933	0.933	0.933
	Recall	0.867	0.867	0.875	0.875	0.875
Avarage	**F1-Score**	0.846	0.842	0.887	0.857	0.705
	Precision	0.942	0.902	0.907	0.880	0.808
	Recall	0.805	0.839	0.873	0.865	0.757

Pada Tabel 4-6 dapat dilihat dari lima nilai k yang berbeda yaitu 2, 8, 10, 12 dan 16 bahwa nilai performansi *F1-measure* dengan nilai tertinggi adalah nilai k = 10. Hal tersebut karena setiap dataset mempunyai ciri khas sendiri, dengan skenario yang sudah di jalankan ternyata turun naiknya pada *F1-measure*. Adapun analisis yang di dapatkan adalah berdasarkan hasil yang sudah didapat adalah:

1. Meningkatkan nilai K tidak memastikan naiknya akurasi yang didapat, seperti pada data Leukimia. Hal ini terjadi karena persebaran data (perbedaan dari nilai maksimal dan nilai minimal) setiap dataset berbeda-beda yang menyebabkan perbedaan nilai k pada setiap dataset dengan hasil yang maksimal.
2. Maksimal rata-rata dari akurasi dengan nilai k adalah 10 yaitu 88.7%. Namun pada data *Leukimia* tidak mendapatkan nilai maksimal pada k sama dengan 10. Data *Leukimia* mendapatkan nilai tertinggi pada k=2 dan k=12.
3. Pada data *Lung-Cancer* terdapat titik jenuh. Dimana hasil tidak meningkat atau menurun setelah k=10. Tidak seperti halnya data yang lain setelah mendapatkan hasil yang maksimal maka akan menurun setelah nilai k ditambahkan.

4.2.2. Perbandingan dan Rata-rata

Setelah melakukan seluruh percobaan didapatkan hasil maksimal untuk setiap dataset sebagai berikut:

Tabel 4-2 Hasil Rata-rata seluruh nilai maksimal setiap data

Evaluation	Breast	Colon	Leukimia	Ovarian	Lung	Avg
Max F1-Score	0.84	0.867	0.882	0.984	0.980	0.9106
Max Precision	1.00	0.6	1	1	0.933	0.9066
Max Recall	0.8	1	0.833	0.975	0.875	0.8966

Pada tabel 4-7 membuktikan bahwa metode *Mutual Information* dan *Bayesian Network* dapat mengklasifikasi data *Microarray* dengan hasil rata-rata F1-Score 91.06%. Hasil ini didapatkan dari jumlah variabel dan nilai k yang berbeda untuk setiap data.

5. Kesimpulan dan Saran

5.1. Kesimpulan

Berdasarkan hasil penelitian yang telah didapatkan, maka kesimpulan yang dapat diambil dari penelitian ini adalah sebagai berikut.

a. Metode klasifikasi *Bayesian Networks* teruji dapat melakukan klasifikasi ekspresi gen untuk mendeteksi kanker dari *Kent Ridge Bio-medical Dataset* dengan nilai performansi rata-rata F1-Score 91.06%, *precision* 90.66%, *recall* 89.66%.
b. Proses seleksi fitur dapat dilakukan setelah melakukan diskritisasi dan *cleansing dataset*. Metode yang dapat digunakan untuk melakukan seleksi fitur adalah *Mutual Information.* Atribut yang terbaik digunakan berbeda-beda pada setiap data.
c. Rata-rata akurasi di percobaan nilai k terbaik pada saat diskritisasi dengan nilai 10. Hal tersebut dapat dilihat pada tabel 4-6 yang menjelaskan bahwa rata-rata performansi nilai *F1-Measure* terbaik adalah k dengan nilai 10

5.2. Saran

Sistem yang telah dibangun sangat mungkin untuk dikembangkan lebih lanjut. Hal yang peneliti sarankan yaitu penggunaan *Structure Learning* untuk meningkatkan efisiensi dan efektifitas dari DAG yang dibangun. Serta menggunakan metode seleksi fitur yang lain akan mendapatkan nilai-nilai baru pada atribut yang didapatkan. Melakukan percobaan pada semua dataset yang ada pada *Kent Ridge Bio-medical Dataset.*

References

[1] W. H. Organization, "Cancer fact sheet," Februari 2015. [Online]. Available: http://www.who.int. [Accessed 1 November 2016].

[2] A. Nurfalah, Adiwijaya and A. A. Suryani, "Cancer is a leading cause of death worldwide,," *Far East Journal of Electronics and Communications,* vol. 16, no. 2, pp. 269-281, 2015.

[3] U. B. Kjaerulff and A. L. Madsen, Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis, New York: Springer, 2012.

[4] D. Heckerman, A Tutorial on Learning With Bayesian Networks, Redmond: Microsoft Corporation, 1996.

[5] C. Sun-Bae and W. Hong-Hee, "Machine learning in DNA microarray analysis for cancer classification," *Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics,* vol. 19, pp. 189-198 , 2003.

[6] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaseenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, D. Caligiuri and Lander E.S, Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, Science, 1999.

[7] U. Alon, Broad patterns of gene expression revealed by clustering of tumor and normal colon tissues probed by oligonucleotide arrays, Proceedings of the National Academy of Sciences of the United States of America, 1999.

[8] T. Furey, Support Vector Machine classification and validation of cancer tissue samples using microarray expression data, Bioinformatics, 2000.

[9] A. Narayanana, E. Keedwell, J. Gamalielsson and S. Tatineni, Single-layer artificial neural networks for gene expression analysis, Neurocomputing, 2004.

[10] J. Lee, J. Lee, M. Park and S. Song, An extensive evaluation of recent classification tools applied to microarray data, Computation Statistics and Data Analysis, 2005.

[11] D. Devaraj, B. Yegnanarayana and K. Ramar, Radial basis function networks for fast contingency ranking, Journal of Electrical Power and Energy Systems, 2002.

[12] D. A. V. C, D. D and V. M, "Gene Expression Data Classification using Support Vector Machine and Mutual Information-based Gene Selection," *Procedia Computer Science,* vol. 47, pp. 13-21, 2015.

[13] E. Alpaydin, Introduction to Machine Learning., MIT Press, 2010.

[14] D. Michie, D. Spiegelhalter and C. Taylor, Machine Learning, Neural and Statistical, 1994.

[15] D. Jurafsky and C. Manning, "Information Retrieval," [Online]. Available: www.class.coursera.org. [Accessed 1 10 2015].

[16] A. Lukman and M. Nadzirin Anshari Nur, "ALGORITMA BAYESIAN NETWORK UNTUK SIMULASI PREDIKSI PEMENANG PILKADA MENGGUNAKAN MSBNX".

[17] J. Pearl, Probabilistic Reasoning in Intelligent Systems, San Francisco: Morgan Kaufmann, 1988.

[18] D. C. Manning, P. Raghavan and H. Schutze, Introduction to Information Retrieval, Cambridge: Cambridge University Press, 2008.

[19] F. Guillet and H. J. Hamilton, Quality Measures in Data Mining, Berlin: Springer, 2007.

[20] Bio-medical, Kent Ridge, "Kent Ridge Bio-medical Dataset," 2003. [Online]. Available: http://datam.i2r.a-star.edu.sg/datasets/krbd/. [Accessed 2016].

[21] Adiwijaya, Aplikasi Matriks dan Ruang Vektor, Yogyakarta: Graha Ilmu, 2014.

[22] Adiwijaya, Matematika Diskrit dan Aplikasi, Bandung: Alfabeta, 2016.

[23] M. S. Mubarok, Adiwijaya and M. D. Aldhi, "Aspect-based sentiment analysis to review products using Naïve Bayes," *AIP Conference Proceedings*, vol. 1867, no. 020060, 2017.

[24] "Degree centrality and eigenvector centrality in twitter," *IEEE*, vol. 8, no. In Telecommunication Systems Services and Applications (TSSA), pp. 1-5, 2014.

[25] V. Effendy, Adiwijaya and Z. A. Baizal, "Handling imbalanced data in customer churn prediction using combined sampling and weighted random forest. In Information and Communication Technology (ICoICT).," *IEEE*, vol. 2, no. Information and Communication Technology (ICoICT), pp. 325-330, 2014.

[26] I. N. Yulita, L. T. Houw and Adiwijaya, "Fuzzy Hidden Markov Models for Indonesian Speech Classification," *JACIII*, vol. 16(3), pp. 381-387, 2012.