Deteksi Kanker pada Data *Microarray* Menggunakan Metode *Naïve Bayes* dengan *Hybrid Feature Selection*

Bintang Peryoga[1], Adiwijaya[2], Widi Astuti[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[1]bintangperyoga@students.telkomuniversity.ac.id, [2]adiwijaya@telkomuniversity.ac.id,
[3]wididwu@telkomuniversity.ac.id

1. Pendahuluan

Kanker merupakan penyakit mematikan yang dapat menyerang bagian tubuh mana pun. Menurut World Health Organization[1], kanker merupakan penyakit mematikan kedua dan bertanggung jawab atas 9.6 juta kematian pada tahun 2018 di dunia dengan kasus kanker yang banyak terjadi yaitu kanker paru-paru (2.09 juta kasus) dan kanker payudara (2.09 juta kasus) sehingga diperlukan pendeteksian kanker sejak dini agar dapat penanganan segera dan tingkat kematian akibat kanker dapat dikurangi. Salah satu teknologi yang dapat dimanfaatkan untuk mendeteksi kanker yaitu *microarray*. Microarray mampu membantu peneliti untuk memantau dan menganalisis perubahan ekspresi gen dalam suatu organisme[2]. Teknologi *Microarray* pada data kanker mempelajari identifikasi ekspresi dan karakteristik yang berbeda pada gen pasien kanker yang hasilnya dapat diaplikasikan untuk memprediksi keadaan pasien tersebut[3]. Akan tetapi, data *microarray* memiliki dua masalah penting yaitu *high-dimensional* dan *high-complexity*[4]. Data *microarray* bersifat *high-dimensional* karena memiliki fitur yang mencapai ribuan lebih. Dimensi data yang tinggi akan berdampak pada *learning algorithm* karena akan menurunkan kinerja program ketika fitur yang tidak terlalu penting menambah ruang pencarian dan membuat generalisasi menjadi lebih sulit[5]. Oleh karena itu, dibutuhkan proses reduksi dimensi untuk mengurangi kompleksitas data tersebut[6].

Reduksi dimensi dapat mengurangi penggunaan fitur yang dianggap tidak penting untuk proses klasifikasi. Pemilihan reduksi dimensi yang tepat dapat mengoptimalkan waktu pengklasifikasian dan akurasi[7]. Seleksi fitur merupakan salah satu cara untuk mereduksi dimensi. Menurut Pengyi Yang pada penelitiannya[8], seleksi fitur dibagi menjadi 3 yaitu *Filter*, *Wrapper*, dan *Embedded(Hybrid)*. Metode *Filter* bekerja tanpa pengaruh dari teknik klasifikasi yang dipakai sehingga secara komputasi akan lebih efisien[9]. Cara kerja metode *Filter* yaitu dengan menghitung nilai peringkat dari tiap fitur. *Information Gain* merupakan salah satu metode *Filter*. Metode *Wrapper* memiliki kelemahan yaitu komputasi yang tidak efisien karena ia mengambil hipotesis model ke dalam *training* dan *testing* pada ruang fitur yang dipakai, juga mengambil lebih banyak *CPU time* dan memori untuk *running program*[9]. Kelebihan dari *Wrapper* adalah ia dapat mendeteksi sifat ketergantungan antar fitur. *Genetic Algorithm* merupakan salah satu metode *Wrapper* dengan jenis *Randomize* yang paling sering dipakai[9]. Di antara semua metode *Wrapper*, *Genetic Algorithm* mendapatkan akurasi tertinggi dengan jumlah gen yang dipilih paling sedikit[4].

Hybrid Feature Selection merupakan salah satu metode seleksi fitur. Metode *Hybrid* dapat menggabungkan metode *Filter* dan *Wrapper* menjadi suatu kesatuan sehingga secara *computational time* lebih cepat dan secara performansi lebih baik[4]. Pada penelitian Nada Almugren[4] tahun 2019 yang berisi tabel komparasi penelitian sebelumnya tentang penggunaan metode *Hybrid* yang beragam, hasil dari banyaknya penelitian tersebut mendapatkan tingkat akurasi diatas 83% untuk data *microarray Colon*, *Leukemia*, *Prostate*, *Lung*, dan *Breast* sehingga terbukti bahwa metode *Hybrid* dapat mengurangi penggunaan fitur gen pada saat klasifikasi tanpa mengurangi tingkat akurasi. Pada penelitian ini, peneliti menggunakan seleksi fitur *Hybrid* dengan menggabungkan *Information Gain* dan *Genetic Algorithm* serta menggunakan metode klasifikasi *Gaussian Naïve Bayes* yang bertujuan data kanker yang dipakai mendapatkan akurasi diatas 95% dengan fitur yang dipakai kurang dari 50 fitur.