

BAB I

PENDAHULUAN

Bab ini merupakan penjelasan mengenai latar belakang, rumusan masalah, tujuan dan manfaat, ruang lingkup, serta sistematika penulisan dari implementasi metode seleksi fitur *Chi-Square* dan *Gradient Boosting Machines Fusion* pada diagnosis penyakit kanker payudara.

1.1 Latar Belakang

Kanker payudara merupakan penyakit yang memiliki pertumbuhan yang tinggi pada penderitanya untuk setiap tahunnya. Kanker payudara terjadi karena adanya tumor ganas yang terbentuk disel-sel payudara. Di mana gen yang bertanggung jawab mengatur pertumbuhan sel di dalam payudara bermutasi (Sari & Bahrum, 2022). Kematian pada tahun 2020 yang diakibatkan oleh kanker payudara pada wanita berkontribusi sebanyak 30,8% yang diambil dari sumber data *Observatory* (Angkasa & Pangaribuan, 2022).

Machine learning (ML) merupakan salah satu aplikasi dari kecerdasan buatan atau *Artificial Intelligence* (AI) yang fokus dalam mengembangkan sebuah sistem yang mampu belajar sendiri tanpa harus diprogram lagi. ML memerlukan sebuah data *training* sebagai proses *learning* sebelum algoritma tersebut menghasilkan sebuah hasil (Chazar & Widhiaputra, 2020). Jadi dapat disimpulkan bahwa ML merupakan suatu pemrograman komputer yang memiliki tujuan untuk mencapai kriteria (performa) tertentu dengan menggunakan sekumpulan data *training* atau *past experience* (Primartha, 2018).

Penelitian terhadap penyakit kanker payudara telah beberapa kali dilakukan oleh para peneliti. Penelitian yang menggunakan metode *Support Vector Machine* (SVM) memiliki hasil angka akurasi sebesar 98,24%. Penelitian tersebut diterapkan dalam bentuk *web base* dengan *framework Django* agar dapat digunakan oleh dokter dalam menentukan keputusan dengan cepat (Adiningrum dkk., 2023).

SVM merupakan suatu teknik klasifikasi untuk masalah *nonlinier* yang dapat bekerja secara akurat dan efisien pada data dengan fitur berdimensi tinggi, selain itu SVM memiliki landasan dasar matematika yang kuat. Hasil dari penelitian tersebut

yang dilakukan terhadap *Breast Cancer Wisconsin (Diagnostic) Dataset* diperoleh bahwa algoritma SVM dapat menghasilkan akurasi sebesar 88,733% pada pengaplikasian dalam bentuk *web* (Chazar & Widhiaputra, 2020).

Wisconsin Breast Cancer (WBC) Dataset memiliki hasil bahwa algoritma SVM lebih akurat, tingkat kesalahan lebih rendah, dan berkinerja lebih baik. Penelitian ini dilakukan dengan membandingkan 6 algoritma yaitu *Decision Tree*, *Neural Networks*, *Naive Bayes* (NB), *Logistic Regression* (LR), SVM, *K-Nearest Neighbor* (K-NN) (Solikin, 2021).

Penelitian yang menggunakan data *Breast Cancer Wisconsin (Diagnostic) Dataset* dengan pengolahan data menggunakan aplikasi *RapidMiner*, menghasilkan kesimpulan bahwa algoritma LR, *Decision Tree*, NB dan K-NN memiliki nilai akurasi yaitu 95,00% sedangkan *Random Forest* dan SVM memiliki nilai akurasi 94,75% dan *Neural Network* 94,25%. Penelitian ini dilakukan dengan membandingkan 7 metode algoritma yaitu *Neural Network*, *Decision Tree*, NB, K-NN, LR, *Random Forest*, dan SVM (Muntiari & Hanif, 2022).

Elastic SCAD adalah metode yang mengombinasikan antara metode *Ridge Penalty* dan SCAD. SCAD sendiri merupakan pengembangan dari metode LASSO. *Elastic SCAD* dapat digabungkan menggunakan algoritma SVM. Penelitian menggunakan *elastic SCAD SVM* memiliki hasil akurasi yaitu 95,4%. Hasil tersebut memiliki kelebihan dibandingkan penelitian sebelumnya yang juga mengombinasikan antara algoritma *elastic SCAD* dan SVM yaitu fitur yang terpakai pun berkurang dari 1213 atribut menjadi 1193 atribut (Firmansyah dkk., 2012).

Penelitian dengan menggunakan data sekunder yang diperoleh dari *Kaggle*, memperoleh hasil yaitu dari algoritma K-NN mendapatkan nilai akurasi 73,6% dan digabungkan dengan arsitektur *Artificial neural network* (ANN) diperoleh hasil rata-rata kesalahan atau *mean (cv.error)* yaitu sebesar 0,1799. Penelitian tersebut menerapkan algoritma K-NN yang ditambah dengan membuat arsitektur ANN. Penggabungan metode tersebut mendapatkan kesimpulan yaitu tingkat akurasi yang dihasilkan dari perancangan sistem ini sebesar 100% (Yuhana, 2018).

Penelitian lain yang menggabungkan dari 3 algoritma yaitu K-NN, *Bagging* dan *Random Forest*. Penelitian tersebut memiliki hasil yaitu K-NN lebih unggul daripada algoritma yang lain, yaitu memiliki nilai akurasi sebesar 74,37 % (Widodo dkk., 2021). Penelitian tersebut dilakukan terhadap data yang diambil dari UCI Edu,

dengan data latih sebanyak 546 data dan data uji sebanyak 137. Penelitian tersebut memiliki hasil yaitu *Deep Learning* menghasilkan nilai akurasi sebesar 98,5401% dan ML 97,0803%. Hal ini menunjukkan bahwa perbandingan antara metode *Deep Learning* dengan algoritma RBM lebih baik dibandingkan ML dengan algoritma FNN (Amelia dkk., 2018). Penelitian lain yang dilakukan oleh Prahartiwi & Dari (2021) yaitu membandingkan performa antara NB, *Decision Tree* dan SVM, menghasilkan kesimpulan bahwa algoritma SVM memiliki tingkat akurasi tertinggi yaitu sebesar 74,29%. Penelitian tersebut dilakukan pada data sekunder *Breast Cancer Coimbra* yang diambil dari *UCI Repository*.

Receiver Operating Characteristic (ROC) Curve merupakan metode analisis yang menampilkan kinerja dari sebuah model klasifikasi pada semua *thresholds* klasifikasi dalam bentuk graf (Angkasa & Pangaribuan, 2022). Penelitian terkait mengaplikasikan metode ROC pernah dilakukan oleh (Angkasa & Pangaribuan, 2022). Penelitian tersebut membandingkan algoritma *Random Forest* dan K-NN dengan ROC. Penelitian tersebut memperoleh kesimpulan yaitu dapat membuktikan bahwa algoritma K-NN memiliki nilai *ROC Curve* 0,9599 lebih besar dari nilai algoritma *Random Forest* yaitu sebesar 0,9551.

Chi-Square merupakan metode uji numerik yang mengukur penyimpangan dari distribusi yang diharapkan. Distribusi yang diharapkan merupakan distribusi fitur-fitur yang menunjukkan sebuah hubungan *independent* pada *class value* (Ikram & Cherukuri, 2017). Berdasarkan penelitian yang dilakukan oleh Purnamawati dkk. (2023) fitur *Chi-Square* dapat memberikan nilai akurasi sebesar 83,33%, dengan metode BM25 sebagai pembobotan kata, dan *improved* K-NN sebagai penentu dalam memilih sentimen.

Gradient Boosting Machine merupakan algoritma *boosting* yang mengaplikasikan pohon regresi sebagai *base learner*-nya. Algoritma ini dapat diaplikasikan ke dalam permasalahan klasifikasi ataupun regresi. Pada masalah klasifikasi, algoritma ini menggunakan *base learner*-nya untuk melatih *classifier* secara berulang menggunakan gradien negatif dari fungsi kerugian yang diturunkan. *Classifier* diharapkan dapat mengurangi fungsi kerugian berdasarkan data *training* (Sunaryono dkk., 2022).

Dasar dari penggunaan metode *Gradient Boosting Machines Fusion* ini mengacu pada penelitian yang dilakukan oleh Sunaryono dkk. 2022). Penelitian

tersebut menggunakan metode *Gradient Boosting Machines Fusion* untuk mendeteksi epilepsi secara otomatis menggunakan *public EEG signals dataset* yang disediakan oleh *Department of Epileptology University of Bonn (UoB)*, Jerman. Hasil dari penelitian tersebut dapat memberikan nilai akurasi sebesar 100% dalam mendeteksi epilepsi.

Gradient Boosting Machines Fusion pada penelitian ini dapat digunakan untuk meningkatkan performa klasifikasi metode *Gradient Boosting Machine* (Sunaryono dkk., 2022). Penelitian ini menggunakan data sekunder dari *Breast Cancer Wisconsin (Diagnostic) Dataset* yang berasal dari *Kaggle*. Data ini dilakukan seleksi fitur menggunakan metode *Chi-Square*. Implementasi metode *Chi-Square* dan *Gradient Boosting Machines Fusion* pada *dataset* tersebut diharapkan dapat menghasilkan nilai akurasi yang lebih besar dari penelitian-penelitian sebelumnya, sehingga dapat menghasilkan metode baru dalam mendiagnosis penyakit kanker payudara. Hal tersebut berdasarkan pada penelitian sebelumnya yang dilakukan oleh Sunaryono dkk. (2022). Penelitian tersebut menyatakan, bahwa hasil yang mereka dapatkan yaitu sebesar 100%, akan berbeda juga jika diaplikasikan ke *dataset* EEG lainya (Sunaryono dkk., 2022). Hal ini juga mungkin akan berlaku pada penelitian yang penulis lakukan dikarenakan data yang digunakan juga berbeda.

1.2 Rumusan Masalah

Berdasarkan latar belakang di atas, rumusan masalah pada penelitian ini adalah bagaimana mengimplementasi metode seleksi fitur *Chi-Square* dan *Gradient Boosting Machines Fusion* pada prediksi penyakit kanker payudara.

1.3 Tujuan dan Manfaat

Tujuan dari penelitian ini adalah mengimplementasikan metode seleksi fitur *Chi-Squared* dan *Gradient Boosting Machines Fusion* pada prediksi penyakit kanker payudara agar menghasilkan model yang dapat memprediksi penyakit kanker payudara dengan mencapai nilai akurasi yang tinggi dan andal.

Manfaat dari penelitian ini adalah diperolehnya model prediksi penyakit kanker payudara menggunakan metode seleksi fitur *Chi-Squared* dan *Gradient Boosting Machines Fusion* yang dapat memberikan akurasi yang signifikan dalam mendeteksi kanker payudara.

1.4 Ruang Lingkup

Ruang lingkup pada penelitian ini yaitu:

1. Implementasi dilakukan untuk memperoleh nilai akurasi tinggi dari pengimplementasian metode *Chi-Squared* dan *Gradient Boosting Machines Fusion* pada prediksi penyakit kanker payudara.
2. Data yang digunakan dalam penelitian ini merupakan data sekunder yang didapatkan dari *Kaggle*, merupakan data *Breast Cancer Wisconsin (Diagnostic) Dataset* dengan diagnosis 'M' (*malignant*) dan 'B' (*benign*). Terdiri dari 569 data dan 32 kolom. Fitur-fitur yang ada didapatkan dari gambar digital pada *Fine Needle Aspirate (FNA)* massa payudara.

1.5 Sistematika Penelitian

Sistematika penulisan pada laporan ini terbagi dalam lima bab pokok bahasan mengenai penelitian “Implementasi Metode Seleksi Fitur *Chi-Squared* dan *Gradient Boosting Machines Fusion* pada Prediksi Penyakit Kanker Payudara”. Sistematika penulisan ini ditujukan agar lebih sistematis dan mudah dipahami. Uraian singkatan mengenai masing-masing pokok bahasan adalah sebagai berikut:

BAB I PENDAHULUAN

Bab ini merupakan penjelasan mengenai latar belakang, rumusan masalah, tujuan, manfaat, ruang lingkup, serta sistematika penulisan dari Implementasi Metode seleksi fitur *Chi-Square* dan *Gradient Boosting Machines Fusion* pada diagnosis penyakit kanker payudara.

BAB II LANDASAN TEORI

Bab ini menyajikan *state of the art*, *breast cancer wisconsin (Diagnostic) dataset*, *pre-processing*, *Chi-Square*, *Gradient Boosting Machines Fusion*, *Gradient Boosting Machines*, kinerja model, *Anaconda Navigator* dan *python*.

BAB III METODE PENELITIAN

Bab ini menyajikan pengumpulan data, preprocessing, pembagian data, membangun model *Gradient Boosting Machines Fusion*, menguji model *Gradient Boosting Machines Fusion*, evaluasi model *Gradient Boosting Machines Fusion* dan menentukan model terbaik.

BAB IV HASIL DAN PEMBAHASAN

Bab ini menjabarkan hasil dan pembahasan penelitian.

BAB V PENUTUP

Bab ini membahas kesimpulan dari uraian pada bab-bab sebelumnya dan saran untuk pengembangan penelitian selanjutnya.