1. Pendahuluan

Penyakit Parkinson atau Parkinson *Disease* (PD) pertama kali diperkenalkan oleh Dr. James Parkinson pada tahun 1817 sebagai *shaking palsy* [1]. Penyakit tersebut merupakan penyakit neurodegeneratif yang mengakibatkan sel-sel saraf di otak atau sistem saraf tepi mengecil dan kehilangan fungsinya seiring berjalannya waktu dan akhirnya sel tersebut mati [2]. Penyakit ini menyerang saraf motorik dengan menyebabkan kondisi tremor tak disengaja pada saat beraktivitas maupun saat beristirahat. Saat ini tidak ada cara untuk memperlambat perkembangan penyakit serta belum ditemukan obat yang dapat menyembuhkannya. Metode pengobatan saat ini hanya untuk meringankan gejala fisik dan mental pasien, di antaranya adalah fisioterapi, *occupational therapy*, terapi wicara dan bahasa, penerapan diet yang baik, dll. [3].

Tidak dapat disembuhkannya PD menjadi salah satu faktor yang menjadikan PD sebagai penyakit neurodegeneratif yang paling umum ditemui. Di antara gangguan neurologis yang diteliti dalam *Global Burden of Disease, Injuries, and Risk Factors Study* 2015, PD merupakan penyakit yang paling cepat berkembang dalam hal kecacatan dan kematian. Tingkat kejadian secara spesifik meningkat tajam dimulai pada usia 60 tahun, lalu memuncak pada mereka yang berusia 85-89 tahun, dan menurun mulai pada usia 90 tahun [4][5]. Meskipun fenomena tersebut menunjukkan bahwa insiden PD meningkat seiring bertambahnya usia, beberapa individu telah dinyatakan mengidap penyakit parkinson di usia 30-an dan 40-an [6].

Berdasarkan penjelasan di atas, maka diperlukan suatu penelitian untuk mengidentifikasi PD secara cepat dan akurat. Prioritas tinggi ditempatkan untuk menemukan pengidentifikasi biologis atau penanda biologis dari fase awal. Sehingga orang yang berisiko tinggi untuk maju ke fase klinis PD dapat segera diberikan perawatan. Adapun solusi alternatif untuk mengidentifikasi PD adalah memanfaatkan pendekatan pembelajaran mesin.

Beberapa studi terbaru membuktikan bahwa metode pembelajaran mesin menggunakan data *microarray* memiliki peluang besar dalam mendeteksi penyakit. B. Jeya dkk. (2020) melakukan penelitian tentang penerapan beberapa metode pembelajaran mesin yaitu *Logistic Regression, Support Vector Machine* (SVM)*, Naive Bayes, dan Random Forest-Decision Tree C4.5* dan membandingkan dengan metode klasifikasi *Ensemble Bayesian Rule Learning* (EBRL). Metode tersebut digunakan pada 25 data *microarray* dengan salah satu datasetnya yaitu GSE6613 [9]. Model EBRL yang telah diagregasi menggunakan *uniform combination* (Bagged-BRL-UC) menghasilkan performa prediksi yang lebih baik daripada Bagged-C4.5 dan Boosted-C4.5 dengan masing-masing hasil *Area Under Receiver Operator* (AUROC) 88%, 85%, dan 83% untuk gabungan 25 data *microarray*. Namun, metode yang paling baik dalam memprediksi data GSE6613 adalah Boosted-C4.5 dengan nilai AUROC 61% [9].

Selanjutnya, Falchetti M. dkk. (2020), melakukan penelitian tentang penerapan algoritma klasifikasi pada 4 data *microarray blood-based transcriptome* untuk memprediksi PD [11]. 2 di antara 9 algoritma klasifikasi yang digunakan adalah XGBoost dan RF. Terdapat 2 tahap seleksi fitur yang digunakan yaitu *collinearity recognition* dan *feature relevance-ranking* yang berbasis algoritma RF. Metode klasifikasi *XGBoost dan RF* masing-masing menghasilkan nilai AUROC 79% dan 76%. Shafila G. (2020) melakukan klasifikasi pasien pengidap penyakit Ebola menggunakan kombinasi metode XGBoost dan ANOVA sebagai teknik seleksi fitur pada data *microarray multi-class*. Dengan *hyperparameter tunning*, penelitian tersebut berhasil memperoleh akurasi yang cukup baik di angka 76,92% [12].

Selain itu, Adiwijaya (2018) meneliti metode deteksi kanker menggunakan beberapa teknik klasifikasi *ensemble method* (yaitu *Random Forest, Bayesian Network, Naive Bayes*), SVM dan *Artificial Neural Network* pada data *microarray*. Metode seleksi fitur yang digunakan adalah *Principal Component Analysis (PCA), Mutual Information (MI), dan Relief Method*. Adapun *dataset* yang digunakan merupakan gabungan dari 7 data kanker *microarray (Breast Cancer, Lung Cancer, Lymphoma, Leukemia, Colon, Ovarian,* dan *Prostate)* yang diperoleh dari *Kent-Ridge Biomedical Data Repository*. Kombinasi *Modified Back Propagation* dan PCA sebagai seleksi fitur menghasilkan akurasi 96.07%. Di sisi lain, metode klasifikasi RF dengan teknik seleksi fitur *Relief Method* menghasilkan akurasi sebesar 75% [10].

Pada penelitian ini, kami melakukan pengujian performa *ensemble method* yaitu *Random Forest (RF), Adaptive Boosting (AdaBoost),* dan *Extreme Gradient Boosting (XGBoost)*. Selanjutnya teknik seleksi fitur yang digunakan adalah *analysis of variance* (ANOVA) dan *Mutual Information* (MI) untuk deteksi dini PD pada *microarray*. Adapun *dataset* yang digunakan berasal dari *National Center for Biotechnology Information* dengan nomor data GSE6613.

Batasan

Batasan masalah untuk penelitian ini yaitu mendeteksi PD berdasarkan *dataset* GSE6613 jaringan *whole blood*. Dilakukan penggabungan pemindaian ekspresi gen dalam darah terkait memberi penanda calon *biomarkers* seperti yang ditunjukkan pada fitur *dataset* ini. Pada penelitian ini, hanya ada 2 tahap teknik seleksi fitur yang digunakan di antaranya adalah reduksi fitur yang memiliki nilai *variance* lebih besar dari 0.5 dan pemanfaatan parameter statistik

ANOVA dan MI sebagai teknik seleksi fitur. Terakhir, untuk setiap model yang dibangkitkan hanya memiliki maksimal 82 fitur dengan mempertimbangkan tren nilai *log loss.*