ABSTRACT

Cervical cancer and breast cancer is a disease with the highest prevalence in Indonesia. Annually about 1,500 cases of cervical cancer are found in Indonesia, which made Indonesia as the country with the highest number of cervical cancer cases in the world. Cervical cancer screening and HPV testing is done with a Pap smear test. However, this examination requires a lot of time, costly and highly susceptible bias of the observer during the process of investigation and analysis.

To overcome these problems, several studies have modeled the machine learning with a variety of approaches have been made. However, these studies are constrained by the limitation of the data amounts and the imbalanced data that caused by the different ratio of each case. This can lead to errors in the classification of the minority due to the tendency of the classification results that focus on the majority class.

This study addresses the handling imbalance data on classification of cases Pap test results using the method of *over-sampling*. ADASYN-N and ADASYN-KNN algorithms are proposed as development of ADASYN algorithm to handle datasets with nominal data types. This study include SMOTE-N algorithm to deal with the issues ss a comparison algorithm. As the results, ADASYN-KNN with the preference “0” gives the highest accuracy, precision, recall, and f-score of 95.38%; 95.583%; 95.383%; and 95.283%. The highest ROC area value is obtained with the ADASYN-KNN with preference “1” of 99.183%.

Keywords – ADASYN, cervical cancer, imbalance class, nominal, over-sampling, Pap smear, SMOTE-N

INTISARI

Penyakit kanker serviks dan payudara merupakan penyakit kanker dengan prevalensi tertinggi di Indonesia. Setiap tahunnya sekitar 1.500 kasus kanker serviks ditemukan di Indonesia yang menjadi-kan Indonesia sebagai negara dengan jumlah kasus kanker serviks tertinggi di dunia. Pemeriksaan kanker serviks dilakukan dengan tes HPV dan tes Pap smear. Namun pemeriksaan ini membutuhkan waktu yang lama, biaya yang mahal dan sangat mudah terpengaruh bias dari pengamat saat proses pemeriksaan dan analisisnya.

Untuk mengatasi masalah tersebut, sejumlah penelitian yang memodelkan *machine learning* dengan berbagai pendekatan telah dilakukan. Namun penelitian tersebut terkendala pada terbatasnya data yang tersedia dan *imbalanced* (ketidakseimbangan) data yang disebabkan oleh rasio antar kemunculan kasus yang berbeda-beda. Hal ini dapat mengakibatkan *error* pada klasifikasi kelas minoritas akibat kecenderungan hasil klasifikasi yang berfokus pada kelas mayoritas.

Penelitian ini membahas penanganan ketidakseimbangan data pada kasus klasifikasi hasil tes Pap Smear menggunakan metode *over-sampling*. Algoritme ADASYN-N dan ADASYN-KNN diajukan sebagai pengembangan dari algoritme ADASYN untuk menangani *dataset* dengan tipe data nominal. Sebagai bahan perbandingan, penelitian ini mengikutsertakan algoritme SMOTE-N untuk menangani permasalahan yang diangkat. Hasilnya algoritme ADASYN-KNN dengan preferensi “0” memberikan hasil akurasi, presisi, *recall*, dan *f-score* tertinggi sebesar 95,38%; 95,583%; 95,383%; dan 95,283%. Nilai ROC area tertinggi diperoleh dengan algoritme ADASYN-KNN dengan preferensi “1” sebesar 99,183%.

Kata kunci – ADASYN, kanker serviks, ketidakseimbangan kelas, nominal, *over-sampling*, pap smear, SMOTE-N