

## KOMBINASI *FEATURE SELECTION FISHER SCORE* DAN *PRINCIPAL COMPONENT ANALYSIS* UNTUK KLASIFIKASI *CERVIX DYSPLASIA* (PCA)

Krisan Aprian Widagdo<sup>\*1</sup>, Kusworo Adi<sup>2</sup>, Rahmat Gernowo<sup>3</sup>

<sup>1</sup>Program Studi Magister Sistem Informasi, Universitas Diponegoro,

<sup>2,3</sup>Departemen Fisika, Universitas Diponegoro

Email: <sup>1</sup>krisan.aprian.widagdo@gmail.com, <sup>2</sup>kusworoadi@undip.ac.id, <sup>3</sup>gernowo@yahoo.com

<sup>\*</sup>Penulis Korespondensi

(Naskah masuk: 01 Januari 2020, diterima untuk diterbitkan: 27 April 2020)

### Abstrak

Pengamatan citra Pap Smear merupakan langkah penting dalam mendiagnosis awal terhadap gangguan servik. klasifikasi citra pap smear merupakan tahapan penting dalam analisis citra pap smear. isu fitur irrelevant dan redundant sering muncul dalam proses klasifikasi. isu tersebut mengakibatkan overfitting model serta komputasi cenderung lama. Pada penelitian ini, kombinasi *Fisher Score* PCA berfungsi mengatasi fitur irrelevant dan redundant. Kombinasi *Fisher Score* PCA mampu meningkatkan kinerja klasifikasi. Pertama *Fisher Score* memilih fitur kandidat berdasarkan peringkat *Fisher Score*. selanjutnya PCA mentransformasikan kandidat fitur menjadi dataset baru yang tidak saling berkorelasi. data set transformasi PCA digunakan sebagai input algoritma pembelajaran. Metode *JST Backpropagation* digunakan sebagai metode pembelajaran. Model dievaluasi dengan mekanisme *5 fold cross validation*. Kombinasi ini dibandingkan dengan model fitur asli dan model fitur hasil *FScore*. *fisher Score* dan PCA menghasilkan akurasi 0.964, Sensitivity 0.990 dan Specificity 0.889. *FScore*-PCA menunjukkan performa terbaik dibandingkan dua model lainnya. *Fisher Score* dan PCA membutuhkan waktu 15,49 detik dalam proses pembelajaran.

**Kata kunci:** *Fisher Score, PCA, Pap Smear, Cervix Dysplasia, Backpropagation*

## CLASSIFICATION *CERVIX DYSPLASIA* BASED ON COMBINING *FEATURE SELECTION FISHER SCORE* AND *PRINCIPAL COMPONENT ANALYSIS*

### Abstract

Observation of the Pap Smear image is an important step early diagnosis of cervical dysplasia. The stages of classification of a pap smear image are important stages in the Analysis. The issue of irrelevant and redundant features often arises in the classification process. These issues lead to over fitting models and computation tends to be long. In this study, the combination of *Fisher Score* PCA works to eliminate the irrelevant features. The combination of *Fisher Score* PCA can improve classification performance. First, the *Fisher Score* selects the candidate features based on the *FScore* ranking. then PCA transforms feature candidates into new datasets uncorrelated with each other. PCA transformation data sets are used as learning algorithm input. ANN Back propagation method is used as a learning method. The model is evaluated by the 5 fold cross-validation mechanism. This combination is compared with the original feature model and the *FScore* feature model. *Fisher Score* and PCA produce an accuracy 0.964, sensitivity 0.990 and specificity 0.889. *FScore*-PCA shows the best performance compared to the other two models. *Fisher Score* and PCA takes 15.49 seconds in the learning process.

**Keywords:** *Fisher Score, PCA, Pap Smear, Cervix Dysplasia, Backpropagation*

### 1. PENDAHULUAN

Kanker servik / leher rahim merupakan salah satu ancaman bagi wanita usia subur. Penyakit ini terjadi akibat dari infeksi *Human Papilloma Virus* (Basu dkk., 2018). Penyakit ini sering disebut *silent killer* di kalangan medis. Pada stadium awal atau

prainvasif tidak menunjukkan gejala yang nyata. Penyakit tersebut hanya akan terdeteksi saat dilakukan tes pengamatan pada serviks. (Pandey dkk., 2015). Perlunya tes pengamatan mengurangi resiko kematian. Pap smear tes menjadi populer serta menjadi pilihan untuk mendeteksi gangguan pada

servik. Proses analisis tes pap smear membutuhkan waktu serta ketelitian laboran. Untuk mengatasi permasalahan tersebut diperlukan *Computer Aided Diagnosis*. CAD menggunakan dimensi fitur yang tinggi untuk menggambarkan *region of interest*. Fitur yang dihasilkan dapat bersifat *irrelevant* dan *redundant*. Kondisi tersebut dapat mengakibatkan permasalahan seperti *overfitting* serta komputasi yang tinggi.

Peningkatan kinerja komputasi dapat dicapai dengan penerapan teknik *Feature selection* dan *feature extraction* sebagai tahapan awal algoritma pembelajaran. Teknik ini dapat meminimalisir fitur *irrelevant* dan *redundant* serta mempertahankan fitur yang relevan. *Feature selection* dapat menyusun fitur relevan yang berasal dari fitur asli sehingga mudah diinterpretasikan. Selain itu *Feature selection* mampu mengurangi waktu komputasi (Cai dkk., 2018). *Feature selection* terbagi menjadi tiga bagian antara lain *filter*, *wrapper* dan *embeded method*. *Filter method* membutuhkan waktu komputasi yang cepat dibandingkan *wrapper* dan *embeded method* (Drotár, Gazda dan Vokorokos, 2019). Adapun contoh *filter method* yang sering digunakan dalam antara lain *Correlation Feature selection*, *Consistency*, *Information Gain*, *ReliefF*, *mRMR* (Remeseiro dan Bolon-Canedo, 2019) dan *Fisher Score*. Beberapa penelitian yang sudah dilakukan terkait dengan *Feature selection* dalam cervix dysplasia. Chen dkk. menggunakan algoritma *Boruta* untuk memilih fitur yang relevan. Model yang digunakan untuk mengevaluasi berdasarkan *Random Forest*. Hasil yang diperoleh menunjukkan bahwa performa klasifikasi mengalami peningkatan dengan penggunaan fitur sebesar 1,2% dari jumlah fitur yang ada (Chen dkk., 2019). Wang dkk membandingkan performa metode *filter* dan *wrapper* antara lain *P-value*, *Maximum Relevance Minimum Multicollinearity* (MRmMC) serta *Chain-Like Agent Genetic Algorithm* (CAGA). Fitur yang dihasilkan *Chain-like Agent Genetic Algorithm* (CAGA) terbukti memberikan peningkatan kinerja metode klasifikasi dibandingkan dengan kedua metode yang lain (Wang dkk., 2019).

Teknik selanjutnya, *dimensionality reduction* membentuk fitur set baru yang berdimensi kecil dengan cara mentransformasikan fitur asli. *Principal Component Analysis* sering digunakan dalam mereduksi data. Akan tetapi, PCA menjadi tidak optimal jika terdapat banyak fitur yang tidak relevan dalam dataset (Bharti & Singh, 2015). Wu dan Zhou membandingkan PCA dengan RFE untuk mendiagnosa sample *malignant cancer*. Dataset terdiri dari 32 faktor resiko dan 4 target variabel. Metode kombinasi *Principal Component Analysis* dan SVM memberikan hasil klasifikasi terbaik (Wu & Zhou, 2017). Zheng dkk mengkombinasikan PCA dengan SVM untuk membedakan *cervical squamous cell carcinoma* dan. Nilai akurasi yang *cervical adenocarcinoma* berdasarkan *Raman Spectroscopy*

diperoleh dari model tersebut sebesar 93,125% (Zheng dkk., 2019).

Penggunaan *feature selection* dan *dimensionality reduction* secara bersamaan dapat meningkatkan kinerja klasifikasi. Kombinasi *Information Gain* dengan *Principal Component Analysis* mampu menghilangkan *irrelevant features* dan mempertahankan *features* secara optimal. Hasil eksperimen menunjukkan kombinasi IG-PCA memberikan pengaruh signifikan terhadap kinerja klasifikasi. Kombinasi tersebut meningkatkan akurasi dan menurunkan *False Alarm Rate* (Salo dkk, 2019). Implementasi serupa juga dilakukan (Jain dan Singh, 2018), *Feature selection ReliefF* diintegrasikan dengan PCA. Gabungan metode tersebut menunjukkan performa terbaik apabila dibandingkan dengan *Correlation Based feature Selection* (CBS), *Fast Correlation Based Feature Selection* (FCBF), *Mutual Information Based Feature Selection* (MIFS), *MODTree Filtering Approach* dan *ReliefF Feature Selection*. Selain itu *ReliefF-PCA* mampu menghilangkan *irrelevant features* hingga 25% (data penyakit diabetes) dan 33% (data penyakit kanker payudara). Berdasarkan keunggulan *Feature selection* dan *dimensionality reduction*, penelitian ini mengkombinasikan *Fisher Score* dengan *Principal Component Analysis*. *Fisher Score* memiliki kemampuan dalam melakukan diskriminasi objek antar kelas. *Fisher Score* mampu memilih kandidat fitur berdasarkan perangkikan (Hsu dkk., 2011). *Principal Component Analysis* membentuk data baru dengan fitur tidak saling berkorelasi. Data set hasil ekstraksi *Principal Component Analysis* akan digunakan sebagai fitur metode jaringan syaraf tiruan *Backpropagation*.

## 2. METODE PENELITIAN

Klasifikasi merupakan salah satu fokus dalam penelitian *cervix dysplasia*. Model klasifikasi berkaitan dengan *features* yang digunakan. Tidak jarang *irrelevant* dan *redundant features* terkandung dalam set *features*. Kedua hal tersebut berdampak negatif terhadap performa klasifikasi. Adapun beberapa teknik untuk mengatasi masalah tersebut. Metode *filter* (Saha, Bajger & Lee, 2019), *filter-wrapper* (Kurniawan dkk., 2018) dan *wrapper* (Riana, Hidayanto and Fitriyani, 2017; William dkk., 2019) mampu menangani permasalahan tersebut.

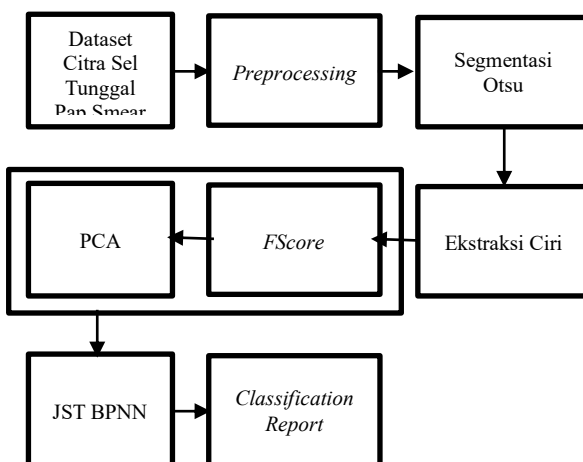
Saha, Bajger & Lee memilih metode fuzzy *entropy* sebagai metode *feature selection*. Ide dasarnya menyaring *features* berdistribusi relatif acak. Metode ini berdasarkan ukuran fuzzy. Semakin kecil nilai *entropy* semakin informatif *features* tersebut. *features* dengan nilai *entropy* terkecil, dipilih untuk digunakan dalam klasifikasi. Lima metode klasifikasi digunakan untuk menguji performa klasifikasi abnormal sel. SVM-Linier menghasilkan kinerja terbaik dibandingkan metode klasifikasi lain. Performa klasifikasi SVM linier

sebagai berikut accuracy 0,96, sensitivity 0,98, specificity 0,91.

Riana, Hidayanto & Fitriyani mengatasi irrelevant *features* dengan metode *Greedy Forward Selection* (GFS). Konsep *Information Gain* digunakan untuk menentukan *features* terbaik. Selain itu, GFS menerapkan dua langkah dalam memilih *features* subset yaitu *Stepwise Forward Selection* dan *Stepwise Backward Elimination*. Algoritma *Naive Bayes* dipilih sebagai metode klasifikasi. Kinerja *Naive Bayes* GFS menghasilkan akurasi tertinggi (92,15%).

Kurniawan dkk mengimplementasikan dua tahap dalam memilih *relevant features* pada klasifikasi normal *cervical cell*. Pada tahap pertama menggunakan *Feature Subset Selection* (FSS) sebagai *filter method*. FSS berfungsi menentukan tingkat signifikansi suatu *features*. FSS melakukan perangkaian pada semua *features* berdasarkan tingkat signifikansi. Semakin tinggi peringkat maka semakin tinggi tingkat signifikansi. Pada tahap kedua menggunakan *wrapper* KNN ( $k = 3$ ,  $k = 5$ ,  $k = 7$ , dan  $k = 9$ ) dan *Backpropagation* (29 hidden layers dengan 1,000 epochs). KNN dan *Backpropagation* menghasilkan 12 *features* untuk membedakan sel cervix normal. William dkk menerapkan simulated annealing sebagai *wrapper*. Metode ini sudah banyak diterapkan. Akan tetapi, William dkk mengevaluasi pemilihan fitur dengan skema random forest ganda. Hasil menunjukkan bahwa Simulated Annealing memberikan dampak positif terhadap kinerja klasifikasi.

Pada penelitian ini diusulkan *hybrid dimensional reduction* dengan kombinasi *Fisher Score*-PCA. *Fisher Score* memiliki keunggulan dalam hal diskriminasi *features* antar kelas. Selain itu dalam segi komputasi tergolong cepat. *Principal Component Analysis* mampu menghasilkan data set baru yang tidak berkorelasi antar fitur. Gambar 1 mengilustrasikan diagram sistem yang dikembangkan. Sistem ini dibangun dengan program Matlab R2015a.



Gambar 1. Blok Diagram Sistem

## 2.1. Pap Smear Tes

Metode deteksi dini dapat mencegah penyakit kanker servik. Ada beberapa metode deteksi yang terkenal dan sering digunakan antara lain *Pap Smear* tes, VIA tes dan VILI tes. *Pap Smear* merupakan metode sitologi yang ditujukan untuk *screening* kanker servik. Metode ini sering dijadikan sebagai program pencegahan *screening* massal. Metode ini sederhana, aman dan efektif. Selain itu *Pap Smear* dipilih sebagai metode *screening* karena lebih efisien dibanding metode yang lain. Cara kerja *Pap Smear* mengambil sel yang tergores dari bagian servik terluar untuk dilakukan pemeriksaan mikroskopis. Tujuan tes ini mendeteksi pra kanker atau perubahan yang berpotensi ke arah pra kanker (*Neoplasia Intrapitel Cervical* atau *Cervical Dysplasia*).



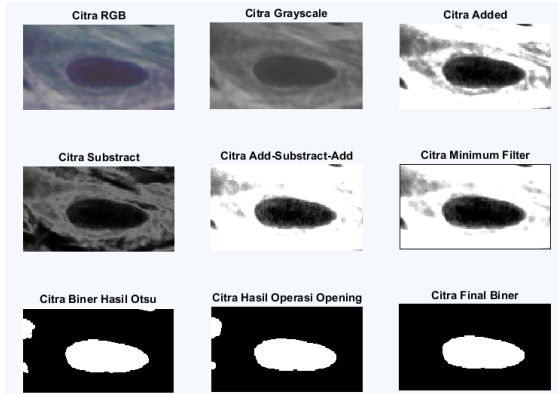
Gambar 2. citra sel tunggal pap smear

## 2.2. Dataset

Dataset yang digunakan berupa citra sel tunggal pap smear berasal dari Rumah Sakit Universitas Herlev (Denmark) (<http://mde-lab.aegean.gr>). Citra Pap Smear tes ini terdiri dari 917 citra sel tunggal. Sel tunggal ini terdiri dari 7 kelas antara lain Normal Columnar, *Carcinoma In Situ*, *Light Dysplastic*, *Moderate Dysplastic*, *Normal Superficial*, *Severe Dysplastic*, dan *Normal Intermediate*. Dari ke tujuh kelas tersebut dikelompokkan menjadi dua kelas yaitu kelas normal (*Normal Columnar*, *Normal Superficial* dan *Normal Intermediate*) dan kelas abnormal (*Severe Dysplastic*, *Light Dysplastic*, *Moderate Dysplastic* dan *Carcinoma In Situ*). Pengelompokan ini berdasarkan pandangan medis, dimana kata *dysplasia* berarti pertumbuhan yang tidak semestinya. Kelas *light*, *moderate* dan *severe* berisiko menjadi *malignant cancer cell* (Norup, 2005).

## 2.3. Segmentasi Nucleus

Segmentasi merupakan langkah yang penting untuk mendapatkan *region of interest* (*nucleus*). Terdapat banyak metode yang dikembangkan dalam untuk melakukan segmentasi. Beberapa diantaranya metode *clustering mean shift* (Wang dkk., 2019), metode *watershed* dan metode otsu *thresholding*. Pada penelitian ini menggunakan metode segmentasi otsu untuk memisahkan *nucleus* dari citra asli. Proses segmentasi otsu ditunjukkan pada gambar (1).



Gambar 2. Proses Segmentasi Nucleus

## 2.4. Ekstraksi Ciri

Citra hasil segmentasi perlu ditransformasikan ke dalam bentuk parameter kualitatif. Parameter kualitatif ini sangat berperan dalam membedakan suatu objek normal dengan objek abnormal. Parameter/Fitur nucleus yang digunakan antara lain fitur tekstur *Gray Level Co-occurrence Matrix*, *Area* dan warna HSV. Fitur *nucleus* disajikan pada tabel 1.

Tabel 1. Fitur Nucleus Sel Tunggal Pap Smear

Fitur	Fitur
X1 <i>Energy</i>	X13 <i>Sum Entropy</i>
X2 <i>Contrast</i>	X14 <i>Sum variance</i>
X3 <i>Correlation</i>	X15 <i>Difference variance</i>
X4 <i>Homogeneity</i>	X16 <i>Difference Entropy</i>
X5 <i>Entropy</i>	X17 <i>Information measure of correlation1</i>
X6 <i>Autocorrelation</i>	X18 <i>Information measure of correlation2</i>
X7 <i>Dissimilarity</i>	X19 <i>Area Ratio 1</i>
X8 <i>Cluster shade</i>	X20 <i>Area Ratio 2</i>
X9 <i>Cluster Prominence</i>	X21 <i>Mean channel H</i>
X10 <i>Maximum probability</i>	X22 <i>Mean channel S</i>
X11 <i>Sum of Squares: Variance</i>	X23 <i>Mean channel V</i>
X12 <i>Sum average</i>	

## 2.5. Feature Selection

### 2.5.1. Fisher Score

*Fisher Score* merupakan *Feature selection* yang berbasis pada ranking dari Ratio masing-masing feature. *Fisher Score* bertujuan menghilangkan irrelevant dan redundant fitur. Nilai *fisher Score* berdasarkan nilai mean dan varian fitur tiap kelas. Semakin tinggi nilai *Fisher Score* suatu fitur, semakin baik dalam membedakan objek antar kelas (Hsu dkk., 2011). Persamaan (1) menggambarkan *Fisher Score*.

$$FScore_r = \frac{\sum_{i=1}^c n_i (\mu_r^i - \mu_r)^2}{\sum_{i=1}^c n_i (\sigma_r^i)^2} \quad (1)$$

$n_i$  menunjukkan jumlah sampel pada kelas ke  $i$ .  $\mu_r^i$  merupakan nilai rata-rata fitur ke  $r$  pada kelas ke  $i$ .  $\sigma_r^i$  merupakan nilai varian fitur ke  $r$  pada kelas ke  $i$ .  $\mu_r$  merupakan nilai rata-rata fitur ke  $r$  (Boonthong dkk., 2015).

Tabel 2. Algoritma FISHER SCORE PCA

Algoritma Fisher Discrimination ScorePCA	
1.	Input dataset X dan Y
2.	Prosedur Hitung <i>FISHER SCORE</i>
3..	Untuk setiap X,
4.	Hitung nilai rata-rata fitur masing-masing kelas
5.	Hitung nilai varian fitur masing-masing kelas
6.	Hitung <i>FISHER SCORE</i> berdasarkan rumus 1.
7.	$R \leftarrow n$ fitur dengan nilai <i>FISHER SCORE</i> tertinggi
8.	Prosedur PCA (R)
9.	Hitung kovarian matrik
10	Hitung vektor eigen dan eigen value dari kovarian matrik
11	$S \leftarrow m$ eigen vector dengan eigen value tertinggi.

### 2.5.2. Principal Component Analysis

PCA sering digunakan dalam dalam mengekstrak informasi dari suatu data. Pca tergolong dalam prosedur statistik yang bertujuan membangun dataset baru yang tidak berkorelasi antar variable tanpa menghilangkan informasi. Hasil rekonstruksi tersebut sering disebut komponen utama. *Principal* komponen disusun dari komponen varian maksimum berurutan setelahnya nilai varian menurun. Sederhananya komponen pertama memiliki varian tertinggi, komponen seterusnya memiliki varian yang lebih kecil.

Berikut ini langkah-langkah PCA dalam mentransformasikan dataset X berdimensi  $m$  menjadi dataset baru Y berdimensi  $n$  dimana  $n < m$  (Zhu dkk., 2019).

- a. Hitung nilai mean

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} \quad (2)$$

- b. Hitung nilai varian

$$s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} \quad (3)$$

- c. Hitung nilai kovarian

$$X^{n \times n} = (x_{i,j}, x_{i,j} = \text{cov}(\text{Dim}_i, \text{Dim}_j)) \quad (4)$$

- d. Hitung nilai eigen value dan eigen vektor

Jika A merupakan matrik  $n \times n$ , vektor  $x$  dalam  $R^n$  disebut vektor eigen dari A. Jika  $Ax$  merupakan skalar dari  $x$ , maka

$$Ax = \lambda x \quad (5)$$

$\lambda$  disebut eigen value dari A dan  $x$  dapat dikatakan sebagai eigen vektor yang berkorespondensi dengan  $\lambda$ . Karena eigen

vektor berkorespondensi dengan eigen value matrik A vektor bukan nol, maka memenuhi persamaan

$$(\lambda I - A)x = 0 \quad (6)$$

Dengan demikian dapat didefinisikan E ke semua vektor x yang memenuhi persamaan 7 sebagai ruang vektor yang bersesuaian.

$$E = \{x; (A - \lambda I)x = 0\} \quad (7)$$

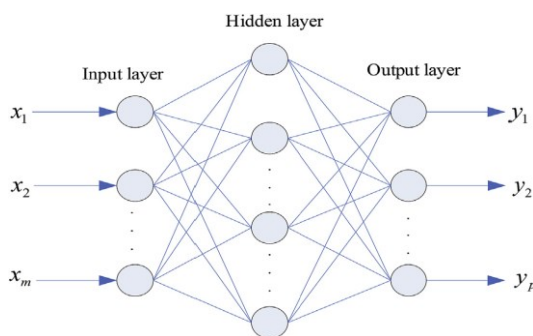
- e. Setelah ruang eigen didapatkan dari kovarian matrik langkah selanjutnya menyusun eigen vektor berdasarkan eigen value dari nilai tertinggi menuju terendah.

## 2.6. Algoritma Backpropagation

Jaringan syaraf tiruan merupakan algoritma pembelajaran yang mengimplementasikan jaringan sederhana yang terkoneksi dengan neuron dan unit. Kinerja jaringan syaraf tiruan mirip dengan kinerja otak manusia dalam mengenali sebuah pola tertentu. JST mampu memberikan hasil klasifikasi yang efektif meskipun data input mengandung *noise* dan tidak lengkap (Dongmei dkk., 2017). Salah satu ragam JST yang sering digunakan adalah *Backpropagation*. Arsitektur *Backpropagation* tersusun atas tiga bagian lapisan antara lain lapisan input, lapisan *hidden* dan lapisan output (Li dkk., 2018). Gambar 3. menunjukkan visualisasi sederhana dari struktur *Backpropagation*. Formulasi *Backpropagation* dapat diformulasikan pada persamaan 8.

$$y = f\left(\sum_{i=1}^L w_i x + b\right) \quad (8)$$

Dengan f merupakan fungsi aktivasi, w menunjukkan vektor bobot, x sebagai vektor input dan b melambangkan nilai bias. Parameter yang digunakan antara lain 1 *hidden layer* (10 *neuron*), 1 *output layer*, fungsi aktivasi *logsig purelin* dan 1000 *epoch*.



Gambar 3. Struktur *Backpropagation* (Li dkk., 2018)

## 2.7. K Fold Cross Validation

Metode *k fold cross validation* merupakan pengembangan metode *hold out validation*. Dalam *k*

*fold cross validation*, keseluruhan data dipartisi sama besar ke dalam K bagian. Sejumlah K bagian tersebut dibagi menjadi dua bagian. K-1 bagian digunakan sebagai pembelajaran, sisanya digunakan sebagai validasi. Metode *k fold cross validation*. Pada gambar 3 diilustrasikan *k=5 fold cross validation*. Kotak yang berwarna putih pada setiap baris menunjukkan data *training*, kotak yang berwarna merah menunjukkan data *testing*.

K1	K2	K3	K4	K5
K1	K2	K3	K4	K5
K1	K2	K3	K4	K5
K1	K2	K3	K4	K5
K1	K2	K3	K4	K5

Gambar 3. Ilustrasi *5 fold cross validation*

## 2.8. Ukuran Kinerja Klasifikasi

Ukuran validasi digunakan untuk mengevaluasi performa klasifikasi. Ada beberapa ukuran validasi antara lain *accuracy*, *sensitivity* dan *specificity* (Wang dkk., 2019). Akurasi merupakan proporsi jumlah sel yang terklasifikasi tepat dalam kelas ke i dengan jumlah keseluruhan sel yang ada. Akurasi ditunjukkan sebagai berikut

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (9)$$

*Sensitivity* merupakan persentase dari data abnormal secara tepat diklasifikasikan sebagai data abnormal (*true positive*) sebagai berikut

$$sensitivity = \frac{TP}{TP+FN} \quad (10)$$

TP adalah jumlah data abnormal yang tepat digolongkan sebagai kelas abnormal dan FN adalah jumlah data abnormal diklasifikasikan sebagai kelas normal. *Specificity* merupakan persentase data normal diklasifikasikan secara tepat ke dalam kelas normal (*true negative*). *Specificity* dijelaskan pada persamaan 11. TN adalah jumlah data normal tepat diklasifikasikan sebagai kelas normal dan FP adalah jumlah data normal diklasifikasikan sebagai kelas abnormal.

$$specificity = \frac{TN}{TN+FP} \quad (11)$$

## 3. HASIL DAN PEMBAHASAN

Pada bagian ini dievaluasi kinerja klasifikasi tiga percobaan yang berbeda dengan metode klasifikasi Neural Network. Percobaan pertama merupakan baseline atau tanpa perlakuan *Feature selection*. Percobaan kedua menggunakan *filter Fisher Score*. Percobaan terakhir menggabungkan *filter Fisher Score* dan *Principal Component Analysis*. Dari ketiga



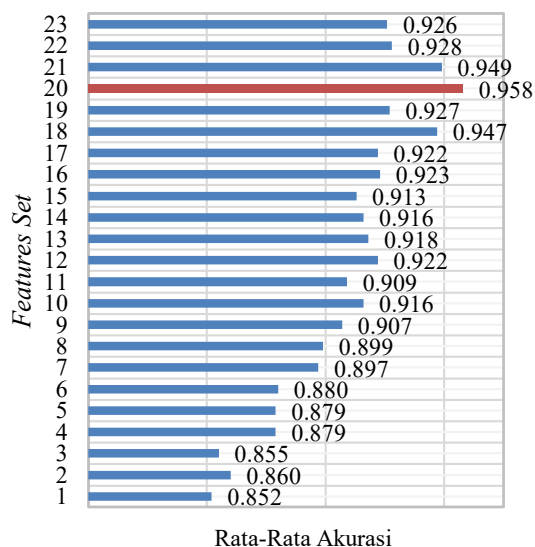
percobaan akan dibandingkan berdasarkan kinerja klasifikasi dengan *5 Fold Cross Validation* dan waktu komputasi.

Percobaan pertama dengan JST *Backpropagation* tanpa *Feature selection* menunjukkan rata-rata akurasi sebesar 0,926. Percobaan dengan *Fisher Score*, langkah pertama melakukan perhitungan nilai *Fisher Score* masing-masing fitur. Hasil perankingan fitur berdasarkan *Fisher Score* ditunjukkan pada tabel 3.

Tabel 3. peringkat fitur berdasarkan *Fisher Score*

Rank	Fitur	FScore	Rank	Fitur	FScore
1	X1	0.4375	13	X7	0.1797
2	X13	0.4255	14	X4	0.1661
3	X5	0.3998	15	X18	0.1661
4	X19	0.3978	16	X15	0.1322
5	X10	0.3928	17	X2	0.1322
6	X17	0.3928	18	X3	0.0754
7	X12	0.3657	19	X8	0.0173
8	X20	0.3302	20	X21	0.0153
9	X6	0.2738	21	X22	0.0135
10	X11	0.2696	22	X9	0.0013
11	X14	0.1874	23	X23	0.0001
12	X16	0.1856			

Dua puluh tiga *features* set (X1, X1 X13, ... dan seterusnya) terbentuk berdasarkan peringkat *fisher Score* (tabel 3). Efektifitas dari kedua puluh tiga *features* set dievaluasi dengan metode *Backpropagation*. Performa klasifikasi setiap *features* set dianalisis berdasarkan nilai akurasi. Sebagai contoh *features* set kedua puluh, terdiri dari 20 *features* tertinggi. *Features* set kedua puluh ini menghasilkan nilai akurasi sebesar 0,958. Begitu pula selanjutnya pada *features* set selanjutnya. Gambar 4 menunjukkan nilai rata-rata akurasi dengan menggunakan *features* set berdasarkan *Fisher Score*. Berdasarkan nilai akurasi tertinggi, *features* set dengan 20 *features* dengan *FScore* tertinggi

Gambar 4. rata-rata akurasi dengan *Fisher Score*

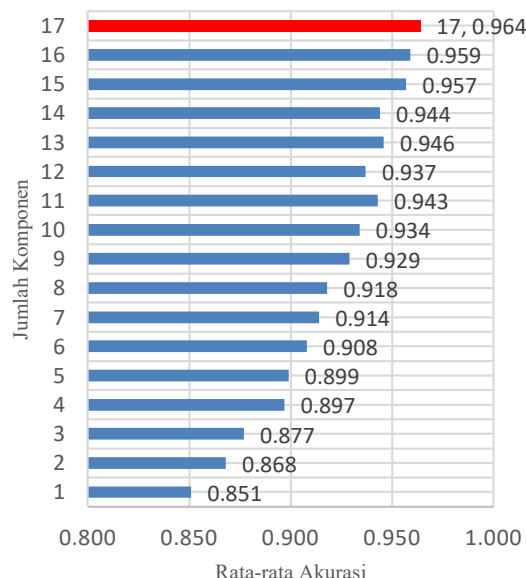
Pada percobaan ketiga, 20 fitur terbaik akan ditransormasikan ke dalam dataset baru berdasarkan

metode PCA. 17 set komponen utama (PC1, PC1 PC2, PC1 PC2 PC3, ... dan seterusnya) terbentuk berdasarkan transformasi PCA. Ketujuh belas set komponen utama akan dianalisis berdasarkan nilai akurasi. Sebagai contoh set komponen ke 17 terdiri dari PC1, PC2, ..., PC17. ketujuh belas komponen utama ini menghasilkan kinerja klasifikasi sebesar 0,964. Gambar 5 menunjukkan hasil klasifikasi JST *Backpropagation* berdasarkan 17 set komponen utama. Set komponen ketujuh belas memiliki nilai akurasi tertinggi dibandingkan 16 set komponen utama lainnya.

Tabel 4. Hasil Kinerja Klasifikasi

Metode	Akurasi	Sensitivity	Specificity
NN	0.926±0.031	0.980±0.013	0.774±0.108
NN + <i>Fisher Score</i>	0.958±0.007	<b>0.992±0.005</b>	0.862±0.024
NN+ <i>Fisher Score</i> PCA	<b>0.964±0.006</b>	0.990±0.005	<b>0.889±0.009</b>

Pada tabel 4 menunjukkan secara lengkap hasil kinerja klasifikasi tiga percobaan yang disimulasikan. Hasil kinerja klasifikasi dengan penambahan metode *Fisher Score* dan PCA mengalami peningkatan. Percobaan ketiga menampilkan kinerja terbaik dibandingkan dua percobaan yang lain. Akan tetapi, pada ukuran *sensitivity* mengalami penurunan sebesar 0.002. Pada *supervised learning*, transformasi PCA tidak menjamin konsistensi dalam membedakan objek antar kelas khususnya kelas abnormal sel.

Gambar 5. Rata-rata akurasi dengan *Fisher Score* PCA

Selanjutnya ketiga percobaan tersebut dibandingkan waktu komputasi klasifikasi. Percobaan pertama membutuhkan waktu komputasi rata-rata 19,21 detik. Percobaan kedua dan ketiga mengalami percepatan waktu komputasi. Percobaan kedua dengan penggunaan *Fisher Score*

membutuhhkan waktu komputasi 18,20 detik. dan kombinasi *Fisher Score* PCA membutuhkan waktu komputasi 15,49 detik. Percobaan ketiga lebih efisien dalam hal komputasi dibandingkan kedua percobaan lain. Hal ini dikarenakan berkurangnya fitur yang digunakan dalam pemodelan klasifikasi. Hal ini menunjukkan bahwa penggunaan *Feature selection* dan *feature extraction* mampu mengurangi atau menghemat waktu komputasi.

Tabel 5. Perbandingan Performa Klasifikasi Pada Data set Herlev

Metode	Akurasi	Sensitivity	Specificity
<b>NB+GFS (Riana, Hidayanto dan Fitriyani, 2017)</b>	0.9215	0.9733	-
<b>FSS + KNN (Kurniawan dkk., 2018)</b>	0.9229	0.9661	0.8020
<b>Fuzzy Entropy + SVM linier (Saha, Bajger dan Lee, 2019)</b>	0.96	0.98	0.91
<b>Enhaced Fuzzy C Means + Simulated Annealing (William dkk., 2019)</b>	<b>0.9888</b>	<b>0.9928</b>	<b>0.9747</b>
<b>NN+ Fisher Score PCA</b>	0.964±0.00 6	0.990±0.00 5	0.889±0.00 9

Perbandingan kinerja klasifikasi penelitian sebelumnya pada data set Herlev disajikan pada tabel 5. Ukuran klasifikasi yang digunakan antara lain akurasi, *sensitivity* dan *specificity*. *Enhaced Fuzzy C means + Simulated annealing* (William dkk., 2019) menghasilkan Kinerja klasifikasi terbaik. Nilai akurasi dan *sensitivity* *Fisher Score* PCA lebih tinggi dibandingkan dengan GFS (Riana, Hidayanto dan Fitriyani, 2017), FSS+KNN (Kurniawan dkk., 2018) dan *Fuzzy Entropy* (Saha, Bajger dan Lee, 2019). Nilai *Specificity* hanya mampu unggul atas FSS +KNN.

#### 4. KESIMPULAN

Penggunaan *Feature selection* dan *Feature extraction* terbukti signifikan memberikan peningkatan dalam hal mengklasifikasikan *cervix dysplasia*. Dilihat dari kinerja klasifikasi, penggunaan *Fisher Score* dan PCA menghasilkan kinerja akurasi 0,964, Sensitivity 0,990 dan specificity 0,889. Hasil tersebut merupakan hasil terbaik dari semua eksperimen. Selain itu dari segi waktu komputasi, kombinasi *Fisher Score* dan PCA mampu mempercepat komputasi algoritma *Backpropagation*. Waktu yang dibutuhkan dalam proses klasifikasi sebesar 15,49 detik.

#### DAFTAR PUSTAKA

- BASU, P., MITTAL, S., BHADRA VALE, D. dan CHAMI KHARAJI, Y., 2018. Secondary prevention of cervical cancer. *Best Practice & Research Clinical Obstetrics & Gynaecology*, [online] 47, pp.73–85. Available at: <<http://dx.doi.org/10.1016/j.bpobgyn.2017.08.012>>.
- BHARTI, K.K. dan SINGH, P.K., 2015. Hybrid dimension reduction by integrating *feature selection* with feature extraction method for text clustering. *Expert Systems with Applications*, [online] 42(6), pp.3105–3114. Available at: <<http://dx.doi.org/10.1016/j.eswa.2014.11.038>>.
- BOONTHONG, P., KULKASEM, P., RASMEQUAN, S., RODTOOK, A. dan CHINNASARN, K., 2015. *Fisher feature selection* for emotion recognition. In: *2015 International Computer Science and Engineering Conference (ICSEC)*. [online] IEEE.pp.1–6. Available at: <<http://ieeexplore.ieee.org/document/7401452>>.
- CAI, J., LUO, J., WANG, S. dan YANG, S., 2018. *Feature selection* in machine learning: A new perspective. *Neurocomputing*, [online] 300, pp.70–79. Available at: <<https://doi.org/10.1016/j.neucom.2017.11.077>>.
- CHEN, H., YANG, L., LI, L., LI, M. dan CHEN, Z., 2019. An efficient cervical disease diagnosis approach using segmented images and cytology reporting. *Cognitive Systems Research*, [online] 58, pp.265–277. Available at: <<https://doi.org/10.1016/j.cogsys.2019.07.008>>.
- DONGMEI, H., SHIQING, H., XUHUI, H. dan XUE, Z., 2017. Prediction of wind loads on high-rise building using a BP neural network combined with POD. *Journal of Wind Engineering and Industrial Aerodynamics*, [online] 170(January), pp.1–17. Available at: <<http://dx.doi.org/10.1016/j.jweia.2017.07.021>>.
- DROTÁR, P., GAZDA, M. dan VOKOROKOS, L., 2019. Ensemble *feature selection* using election methods and ranker clustering. *Information Sciences*, [online] 480, pp.365–380. Available at: <<https://linkinghub.elsevier.com/retrieve/pii/S0020025518309812>>.
- Hsu, H.H., Hsieh, C.W. and Lu, M. Da, 2011. Hybrid *feature selection* by combining filters and

- wrappers. *Expert Systems with Applications*, [online] 38(7), pp.8144–8150. Available at: <<http://dx.doi.org/10.1016/j.eswa.2010.12.156>>.
- JAIN, D. dan SINGH, V., 2018. Diagnosis of Breast Cancer and Diabetes using Hybrid *Feature selection* Method. In: *2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC)*. [online] IEEE, pp.64–69. Available at: <<https://ieeexplore.ieee.org/document/8745830>>.
- KURNIAWAN, R., FUDHOLI, D.H., MUHIMMAH, I., KURNIAWARDHANI, A. dan INDRAYANTI, 2018. Feature *Analysis* of Normal Epithelial Cervical Cell Characteristics in Pap Smear Images. In: *2018 2nd International Conference on Informatics and Computational Sciences (ICICoS)*. [online] IEEE, pp.1–5. Available at: <<https://ieeexplore.ieee.org/document/8621670>>.
- LI, S., CHEN, T., WANG, L. dan MING, C., 2018. Effective tourist volume forecasting supported by PCA and improved BPNN using Baidu index. *Tourism Management*, [online] 68, pp.116–126. Available at: <<https://doi.org/10.1016/j.tourman.2018.03.006>>.
- NORUP, J., 2005. *Classification of Pap-Smear Data by Transductive Neuro-Fuzzy Methods*. Technical University of Denmark.
- PANDEY, D., SHETTY, J., SAMBHAJI, C., SAXENA, P., MISHRA, D. dan CHAWLA, A., 2015. Cervical Cancer as a silent killer: A rare case report with review of literature. *Journal of Cancer Research and Therapeutics*, [online] 11(3), p.653. Available at: <<http://www.cancerjournal.net/text.asp?2015/11/3/653/137997>>.
- REMESEIRO, B. DAN BOLON-CANEDO, V., 2019. A review of *feature selection* methods in medical applications. *Computers in Biology and Medicine*, [online] 112(February), p.103375. Available at: <<https://doi.org/10.1016/j.combiomed.2019.103375>>.
- RIANA, D., HIDAYANTO, A.N. dan FITRIYANI, 2017. Integration of Bagging and greedy forward selection on image pap smear classification using Naïve Bayes. In: *2017 5th International Conference on Cyber and IT Service Management (CITSM)*. [online] IEEE, pp.1–7. Available at: <<http://ieeexplore.ieee.org/document/8089320>>.
- SAHA, R., BAJGER, M. dan LEE, G., 2019. Prior Guided Segmentation and Nuclei Feature Based Abnormality Detection in Cervical Cells. In: *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*. [online] IEEE, pp.742–746. Available at: <<https://ieeexplore.ieee.org/document/8941814>>.
- SALO, F., NASSIF, A.B. dan ESSEX, A., 2019. Dimensionality reduction with IG-PCA and ensemble classifier for network intrusion detection. *Computer Networks*, [online] 148, pp.164–175. Available at: <<https://doi.org/10.1016/j.comnet.2018.11.010>>.
- WANG, P., WANG, L., LI, Y., SONG, Q., LV, S. dan HU, X., 2019. Automatic cell nuclei segmentation and classification of cervical Pap smear images. *Biomedical Signal Processing and Control*, [online] 48, pp.93–103. Available at: <<https://doi.org/10.1016/j.bspc.2018.09.008>>.
- WILLIAM, W., WARE, A., BASAZA-EJIRI, A.H. dan OBUNGOLOCH, J., 2019. Cervical cancer classification from Pap-smears using an enhanced fuzzy C-means algorithm. *Informatics in Medicine Unlocked*, [online] 14(February), pp.23–33. Available at: <<https://doi.org/10.1016/j.imu.2019.02.001>>.
- WU, W. dan ZHOU, H., 2017. Data-Driven Diagnosis of Cervical Cancer With Support Vector Machine-Based Approaches. *IEEE Access*, [online] 5, pp.25189–25195. Available at: <<http://ieeexplore.ieee.org/document/8070120>>.
- ZHENG, C., QING, S., WANG, J., LÜ, G., LI, H., LÜ, X., MA, C., TANG, J. dan YUE, X., 2019. Diagnosis of cervical squamous cell carcinoma and cervical adenocarcinoma based on Raman spectroscopy and support vector machine. *Photodiagnosis and Photodynamic Therapy*, [online] 27(May), pp.156–161. Available at: <<https://doi.org/10.1016/j.pdpdt.2019.05.029>>.
- ZHU, C., IDEMUDIA, C.U. dan FENG, W., 2019. Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. *Informatics in Medicine Unlocked*, [online] 17(March), p.100179. Available at: <<https://doi.org/10.1016/j.imu.2019.100179>>.