Indonesian Journal of Pure and Applied Mathematics
http://journal.uinjkt.ac.id/index.php/inprime
Vol. 1, No. 1 (2019), pp. 8-16
ISSN 2686-5335
doi: 10.15408/inprime.v1i1.12827

Application of Fuzzy K-Nearest Neighbor (FKNN) To Detect the Parkinson's Disease

L. N. Desinaini, Azizatul Mualimah, Dian C. R. Novitasari, and Moh Hafiyusholeh
Department of Mathematics, Faculty of Sciences and Technology
Universitas Islam Negeri Sunan Ampel Surabaya
Email: nadya.desinaini@gmail.com, {diancrini, hafiyusholeh}@uinsby.ac.id

Abstract

Parkinson's disease is a neurological disorder in which there is a gradual loss of brain cells that make and store dopamine. Researchers estimate that four to six million people worldwide, are living with Parkinson's. The average age of patients is 60 years old, but some are diagnosed at age 40 or even younger and the worst thing is some patients are late to find out that they have Parkinson's disease. In this paper, we present a diagnosis system based on Fuzzy K-Nearest Neighbor (FKNN) to detect Parkinson's disease. We use Parkinson's disease dataset taken from UCI Machine Learning Repository. The first step is normalize the Parkinson's disease dataset and analyze using Principal Component Analysis (PCA). The result shows that there are four new factors that influence Parkinson's disease with total variance is 85.719%. In classification step, we use several percentage of training data to classify (detect) the Parkinson's disease i.e. 50%, 60%, 70%, 75%, 80% and 90%. We also use k = 3, 5, 7, and 9. The classification result shows that the highest accuracy obtained for the percentage of training data is 90% and k = 5, where 19 are correctly classified i.e. 14 positive data and 5 negative data, while 1 positive data is classified incorrectly.
Keywords: Parkinson's disease; Fuzzy K-Nearest Neighbor; Principal Component Analysis.

Abstrak

Penyakit Parkinson merupakan kelainan sel saraf pada otak yang menyebabkan hilangnya dopamin pada otak. Para peneliti mengestimasi bahwa, empat sampai enam juta orang di dunia, menderita Parkinson. Penyakit ini rata-rata diderita oleh pasien berusia 60 tahun, namun beberapa orang terdeteksi saat berusia 40 tahun atau lebih muda dan hal terburuk adalah seseorang terlambat untuk mendeteksinya. Di dalam artikel ini, kami menyajikan sistem diagnosa penyakit Parkinson menggunakan metode Fuzzy K-Nearest Neighbor (FKNN). Kami menggunakan Data uji yang diperoleh dari UCI Machine Learning Repository yang telah banyak diterapkan pada masalah klasifikasi. Tahapan pertama yang kami lakukan adalah menormalisasi data kemudian menganalisisnya menggunakan Analisis Komponen Utama (*Principal Component Analysis*). Hasil Analisis Komponen Utama menunjukkan bahwa terdapat empat factor baru yang mempengaruhi penyakit Parkinson dengan variansi total 87,719%. Pada tahap klasifikasi, kami menggunakan beberapa prosentase data latih untuk mendeteksi penyakit yaitu 50%, 60%, 70%, 75%, 80% and 90%. Selain itu, kami menggunakan beberapa nilai k yaitu 3, 5, 7, and 9. Hasil menunjukkan bahwa klasifikasi dengan akurasi tertinggi diperoleh untuk 90% data latih dengan k = 5, dimana 19 diklasifikasikan secara tepat yaitu 14 data positif dan 5 data negatif, sedangkan satu data positif tidak diklasifikasikan dengan tepat.
Kata kunci: penyakit Parkinson; Fuzzy K-Nearest Neighbor; Analisis Komponen Utama.

1. INTRODUCTION

Parkinson's disease is a neurological disease, where the disease causes loss of brain cells that make and store dopamine which is useful for sending messages to control movement in the body. Researchers estimate that 4-6 million people worldwide live with Parkinson's disease. Usually people suffer Parkinson average on 60 years old, but there are some Parkinson's sufferers aged 40 or even younger. Some sufferers find out too late that they have Parkinson's disease [1]. Parkinson's disease was first discovered by Dr. James Parkinson in 1817. Parkinson's disease is a neurological disease, in which the disease causes loss of brain cells that make and store dopamine. Common symptoms of Parkinson's disease are muscle weakness, slow and stiff movements, blood pressure problems, tremors and loss of balance. The cause is still unknown, although researchers believe that Parkinson's disease can be caused by a combination of environmental factors and genetic factors. Until now there has been no treatment that can cure Parkinson's disease, it's just found therapies and drugs to inhibit cell damage [2]. Therefore, we need a program that can detect Parkinson's disease earlier.

The use of computer-based systems as an analytical technique in diagnosing disease is important. Machine learning is an analytical method that helps deal with big data by developing computer algorithms. Machine Learning is broadly classified into supervised learning and unsupervised learning [3]. Some Fuzzy methods used for clustering include Fuzzy C-Means (FCM) and Subtractive Clustering, while those used for classification methods include Sugeno, Tsukamoto, Mandani, and several hybrid methods with fuzzy are Adaptive Neuro Fuzzy Inference System (ANFIS), Fuzzy K-Nearest Neighbor (FKNN), Fuzzy Neural Network (FNN) and others.

In several studies there have been many researchers who use either classification or clustering. First research example conducted by Novitasari et al. [4], they using Fast Fourier Transform (FFT) and ANFIS for classify Epilepsy disease, the results of this research indicate the EEG signal classification system using ANFIS with two classes (Normal-Epilepsy) states accuracy, sensitivity, and precision of 100%. And the classification systems with three class division (Normal- Not Seizure Epilepsy - Epilepsy) resulted in an accuracy of 89.33% sensitivity of 89.37% and precision of 89.33% [4]. Second research example conducted by Novitasari et al. [5], they using fuzzy c-mean, gray level co-occurrence matrix and support vector machine for classify Alzheimer disease, the results of this research give accuracy 93.33% [5]. Third research example conducted by Afifah et al. [6], they using Fuzzy C-means for clustering of rice field in Indonesia as an evaluation of the availability of food production, this research give results the most potential rice field in Indonesia is East Java, Central Java and West Java [6]. Fourth research example conducted by Novitasari et al. [5], using Fuzzy C-means and Adaptive Neighborhood Modified Backpropagation (ANMBP) for classify EEG signals. This research give the temporary result system accuracy 74.37% [7]. Fifth example conducted by Novitasari et al. [8], using Fuzzy C-means and Adaptive Neuro Fuzzy Inference System (ANFIS) for classify EEG signals. This research give the accuracy 89.19% using 2 level wavelet and FCM with 3 clusters [8]. Last example conducted by Febrianti et al. [9], they compare K-means method with Fuzzy C-means for clustering iris data. This research give RMSE value 2.2122E-14 for Fuzzy C-means in 80 training data and 70 checking data. From this research we known that Fuzzy C-means method has a higher level accuracy than the K-means method [9].

In this research, we use FKNN to classify health and Parkinson patient based on 22 attributes. Fuzzy K-Nearest Neighbor (FKNN) method is a combination of Fuzzy logic and KNN method. The advantage of FKNN, compared to the KNN method, is that the FKNN algorithm classifies test data based on metric similarity [10] [11]. In some studies, feature extraction has been widely used to reduce datasets that have a very large number of attributes, so that the dataset can be simplified. There are many methodologies that can be used to perform feature extraction, one of them is the Principal

Component Analysis (PCA) method. PCA is the oldest and most widely used multivariate statistical analysis technique [12].

In this paper, we develop diagnosis system based on Fuzzy K-Nearest Neighbor (FKNN) to detect the Parkinson's disease with k = 3, 5, 7, and 9 and PCA as feature extraction. We use the Parkinson's disease dataset taken from UCI Machine Learning Repository. Data divided into training and testing with percentage of training data are 50%, 60%, 70%, 75%, 80% and 90%. In the classification step, we use the confusion matrix to compare the accuracy.

2. METHOD

We use the Parkinson dataset from the UCI machine learning repository. The purpose of this dataset is to distinguish healthy people from those suffering from Parkinson's with various medical tests conducted. The Parkinson dataset has 22 attributes and consists of 195 data samples divided into 2 classes, namely 147 positive Parkinson data indicated by label 1 and 48 negative Parkinson (healthy/normal) data indicated by label 0. Table 1 is the example of the Parkinson dataset.

Table 1. Sample of Dataset Parkinson

	Data 1	**Data 2**	**Data 3**	**Data 4**	**Data 5**
MDVP: Fo(Hz)	119.992	122.4	122.4	198.764	214.289
Fhi(Hz)	157.302	148.65	148.65	396.961	260.277
Flo(Hz)	74.997	113.819	113.819	74.9040	77.9730
Jitter (%)	0.00784	0.00968	0.00968	0.0074	0.00567
Jitter(Abs)	0.00007	0.00008	0.00008	0.00004	0.00003
RAP	0.0037	0.00465	0.00465	0.0037	0.00295
PPQ	0.00554	0.00696	0.00696	0.0039	0.00317
Jitter: DDP	0.01109	0.01394	0.01394	0.01109	0.00885
Shimmer	0.04374	0.06134	0.06134	0.02296	0.01884
Shimmer(dB)	0.426	0.626	0.626	0.24100	0.19000
Shimmer: APQ3	0.02182	0.03134	0.03134	0.01265	0.01026
Shimmer: APQ5	0.0313	0.04518	0.04518	0.01321	0.01161
APQ	0.02971	0.04368	0.04368	0.01588	0.01373
Shimmer: DDA	0.06545	0.09403	0.09403	0.03794	0.03078
NHR	0.02211	0.01929	0.01929	0.07223	0.04398
HNR	21.033	19.085	19.085	19.0200	21.2090
RPDE	0.414783	0.458359	0.458359	0.451221	0.462803
D2	0.815285	0.819521	0.819521	0.643956	0.664357
DFA	-4.81303	-4.07519	-4.07519	-6.74458	-5.72406
Spread1	0.266482	0.33559	0.33559	0.207454	0.190667
Spread2	2.301442	2.486855	2.486855	2.138608	2.555477
PPE	0.284654	0.368674	0.368674	0.123306	0.148569
Class	Positive Parkinson	Positive Parkinson	Positive Parkinson	Negative Parkinson (Normal)	Negative Parkinson (Normal)

2.1. Pre-processing Data

There are 3 steps i.e. data normalization, variable reduction using PCA, and divide data into training and testing. PCA is the oldest technique and most widely used multivariate statistics to find out which variables are most influential on a data and to reduce datasets that have a very large number of variables, so that the dataset can be simplified. For example, the dataset is a matrix of size (n × D) where n represents observation x_i for $i \in \{1, 2, 3, ..., n\}$ and D represents the variable in the dataset. The general PCA algorithm is [13]:

1. KMO test:

$$KMO = \frac{\sum_i^P \sum_{j \neq i}^P r_{ij}^2}{\sum_i^P \sum_{j \neq i}^P r_{ij}^2 + \sum_i^P \sum_{j \neq i}^P a_{ij}^2},$$

where r_{ij} is the correlation coefficient between variables and a_{ij} is the partial correlation coefficient between variables.

2. Calculate the covariance matrix using $C = \frac{\sum_{i=1}^n (x_{ik} - \bar{x})(x_{ij} - \bar{x})}{n-1}$.
3. Calculate the eigen value $det(C - \lambda I)v = 0$, and eigenvector v using $(C - \lambda I)v = 0$.
4. The obtained eigenvector is the main component that will be used to form a new variable based on the product between the eigenvector v and the normalized dataset matrix.
5. Calculate the variance using $\frac{\lambda_j}{\sum_{j=1}^D \lambda_j} \times 100\%$.
6. The number of the new variables determined based on the percentage of cumulative contributions that calculated using $pk_r = \frac{\sum_{j=1}^r \lambda_j}{\sum_{j=1}^D \lambda_j} \times 100\%$, where $\lambda_1 > \lambda_2 > \ldots > \lambda_D$.

2.2. The FKNN algorithm

The steps in FKNN are [11] [10]:

1. Determine k (the number of the nearest neighbor where $1 \leq k \leq n$) and n (the number of training data).
2. Calculate the membership function using:

$$u_{ij} = \begin{cases} 0.51 + \left(\frac{n_j}{K}\right) * 0.49, & j = i \\ \left(\frac{n_j}{K}\right) * 0.49, & j \neq i \end{cases}, \quad (7)$$

where $\sum u_{ij} = 1$ and n_j is the number of member of class j in the training data n, j is the class data, and K is the number of training data.

3. Calculate the *Euclidean* distance of training data to the test data.
4. Sort the Euclidean values from the smallest values.
5. Determine k nearest neighbors and refer it as new data.
6. Calculate the membership value for the new data from each class:

$$u_i(x) = \frac{\sum_{j=1}^K u_{ij}\left(1/\|x - x_j\|^{\frac{2}{m-1}}\right)}{\sum_{j=1}^K \left(1/\|x - x_j\|^{\frac{2}{m-1}}\right)} \quad (8)$$

where $u_i(x)$ is the membership value of data x to class i, K is the number of the closest neighbor, $\|x - x_j\|$ is distance between data x to x_j in K closest neighbor, $m > 1$ is weight exponent.

7. Select the class that has the largest membership value as output.

These steps are illustrated in Figure 1.

2.3. Performance calculate using Confusion Matrix

The next step is performance calculate test using confusion matrix. Confusion matrix is used to check the performance of a classification model on a set of test data for which the true (real) values are known. Most performance measures such as precision, recall (sensitivity), accuracy and specificity are calculated from the confusion matrix

1. Accuracy (ACC) is calculated as the number of all correct predictions divided by the total number of the dataset.

$$\text{Accuracy (ACC)} = \frac{\text{TP} + \text{TN}}{\text{TP+TN+FP+FN}} \times 100\%, \tag{9}$$

2. Recall/Sensitivity (SN) is calculated as the number of correct positive predictions divided by the total number of positives. It is also called recall (REC).

$$\text{Sensitivity (SN)} = \frac{\text{TP}}{\text{FP+FN}} \times 100\%, \tag{10}$$

3. Specificity (SP) is calculated as the number of correct negative predictions divided by the total number of negatives.

$$\text{Specificity (SP)} = \frac{\text{TN}}{\text{FP+TN}} \times 100\%, \tag{11}$$

4. Precision (PREC) is calculated as the number of correct positive predictions divided by the total number of positive predictions.

$$\text{Precision (PREC)} = \frac{\text{TP}}{\text{TP+FP}} \times 100\%, \tag{12}$$

where TP and TN are explained in Table 2 [14]:

Table 2. Confusion Matrix

Class	Negative	Positive
Negative	TN (True Negative)	FN (False Negative)
Positive	FP (False Positive)	TP (True Positive)

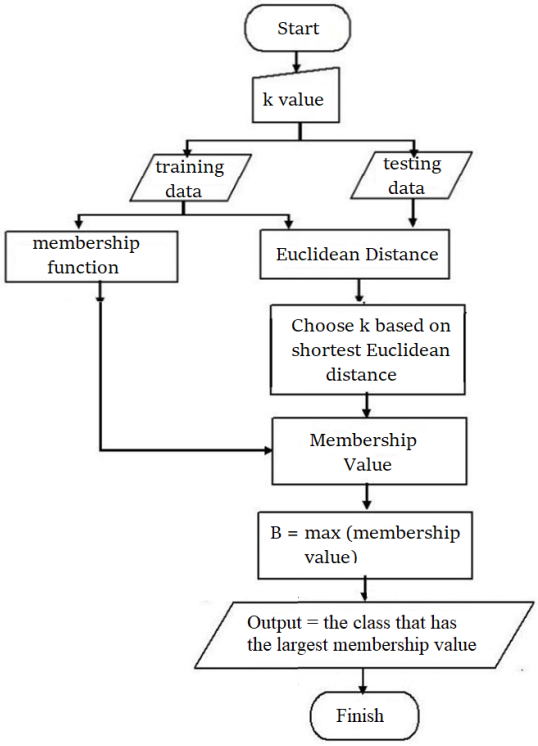

Figure 1. The FKNN Algorithm.

3. RESULTS AND DISCUSSIONS

3.1 The Result of PCA

The first result of PCA obtained MSA value for the D2 variable is 0.491. This MSA value is less than 0.5, therefore the D2 variable is more appropriate to be reduced and the KMO value increases to 0.892 with significance value 0,000 and a Bartlett Test of Sphericity value is 13244,618. This means that the data are meets the requirements to be analyzed using PCA. The next step is choosing a variable that will be entered into the factor i.e. variable which has loading value more than 0.5. The results of this analysis can be seen in Table 3. In Table 4, the variance values for each component are 61.610%, 11.176%, 7.089%, and 5.844%, respectively. This means that the four new components able to explain 85.719% the diversity of data. Factor scores for each component can be seen in Table 3. This factor score will be used in classification process.

Table 3. Factor Score for PCA

	Data 1	Data 2	Data 3	Data 4	Data 5
FAC1	0.50560	1.73694	0.96339	-0.84585	-0.66422
FAC2	0.20348	-0.04479	0.55143	0.77976	0.14006
FAC3	0.48432	0.54947	0.38101	-0.59160	-0.18394
FAC4	-0.74449	-0.58318	-0.89564	1.20900	1.13566
Class	Positive Parkinson	Positive Parkinson	Positive Parkinson	Negative Parkinson	Negative Parkinson

Table 4. The Main Principal Components in Parkinson's Disease

Component	Variables	Variance (%)
1 (FAC 1)	1. Shimmer	61.610%
	2. Shimmer(dB)	
	3. Shimmer: APQ3	
	4. Shimmer: APQ5	
	5. APQ	
	6. Shimmer: DDA	
	7. HNR	
2 (FAC 1)	1. Jitter(%)	11.176%
	2. Jitter(Abs)	
	3. RAP	
	4. PPQ	
	5. Jitter: DDP	
	6. NHR	
3 (FAC 3)	1. MDVP: Fo(Hz)	7.089%
	2. Flo(Hz)	
	3. RPDE	
	4. DFA	
	5. Spread1	
4 (FAC 4)	1. Fhi(Hz)	5.844%
	2. Spread2	

3.2. Classification Results

To detect Parkinson's disease, we will classify data based on factors produced by PCA using FKNN method. In the classification step, we use the training data by percentage: 50%, 60%, 70%, 75%, 80% and 90%. The results of the classification using FKNN can be seen in Table 5 and Figure 2. From Figure 2, it can be seen that the classification results to detect Parkinson's disease using the Fuzzy K-Nearest Neighbor method obtain the highest accuracy of 95% with the percentage of training data is 90% and k = 5, where 19 are correctly classified i.e. 14 positive data and 5 negative data , while 1 positive data is classified incorrectly.

Table 5. The Classification Results for the Training Data

k-value	Testing Data Percentage	Testing Data Percentage	Accuration (%)	Recall/Sensitivity (%)	Specificity (%)	Precision (%)
3	50%	50%	85	90	68	89
5			87	92	72	90
7			85	89	70	90
9			85	89	70	90
3	60%	40%	86	94	67	86
5			94	95	89	97
7			94	95	89	97
9			94	95	89	97
3	70%	30%	84	95	63	84
5			82	90	61	79
7			81	92	58	82
9			81	92	58	82
3	75%	25%	88	97	69	86
5			84	94	63	84
7			84	94	63	84
9			84	94	63	84
3	80%	20%	87	96	69	86
5			82	92	62	83
7			82	92	62	83
9			82	92	62	83
3	90%	10%	85	100	63	80
5			95	100	83	93
7			95	100	83	93
9			95	100	83	93

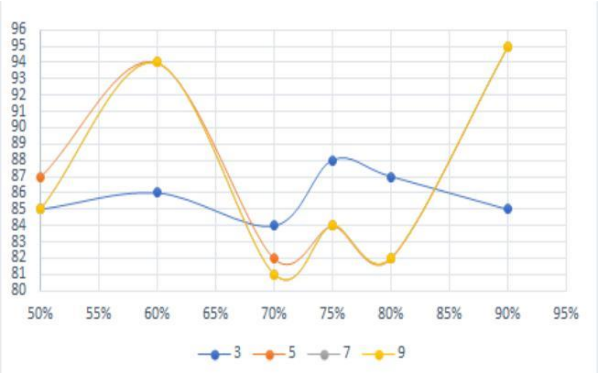

Figure 2. Classification Result for k = 3, 5, 7, and 9.

CONCLUSSION

The main factors in detection Parkinson's disease are factor 1 contains Shimmer, Shimmer (dB), Shimmer : APQ3, Shimmer APQ5, APQ, Shimmer : DDA and HNR; factor 2 contains Jitter (%), Jitter (abs), RAP, PPQ, Jitter: DDP and NHR; factor 3 contains MDVP: Fo(Hz), Flo(Hz), RPDE, DFA and Spread1; factor 4 contains Fhi(Hz) and Spread2. From the several percentage of training data to detect the Parkinson's disease i.e. 50%, 60%, 70%, 75%, 80% and 90%, the highest accuracy for detection is when the 90% of training data with k = 5 i.e. 19 are correctly classified where 14 positive data and 5 negative data, while 1 positive data is classified incorrectly.

REFERENCES

[1] Partners in Parkinson, Inc, "Partners In Parkinson's," AbbVie, Chicago, 2014.

[2] J. S. Purba, Penyakit Parkinson, Jakarta: Badan Penerbit FKUI, 2012.

[3] Tech Differences, "Tech Differences," 2 January 2018. [Online]. Available: https://techdifferences.com/. [Accessed 2 11 2019].

[4] D. C. R. Novitasari, Suwanto, M. H. Bisri and A. H. Asyhar, "Classification of EEG Signals using Fast Fourier Transform (FFT) and Adaptive Neuro Fuzzy Inference System (ANFIS)," *Jurnal Matematika "MANTIK",* vol. 5, no. 1, pp. 35-44, 2019.

[5] D. C. R. Novitasari, W. T. Puspitasari, P. Wulandari, A. Z. Foeady and M. F. Rozi, "Klasifikasi Alzheimer dan Non Alzheimer Menggunakan Fuzzy C-Mean, Gray Level Co-Occurence Matrix dan Support Vector Machine," *Jurnal Matematika "MANTIK",* vol. 4, no. 2, pp. 83-89, 2018.

[6] N. Afifah, D. C. R. Novitasari and A. Lubab, "Pengklasteran Lahan Sawah Di Indonesia Sebagai Evaluasi Ketersediaan Produksi Pangan Menggunakan Fuzzy C-Means," *Jurnal Matematika "MANTIK",* vol. 2, no. 1, pp. 40-45, 2016.

[7] D. C. R. Novitasari, "Klasifikasi Sinyal EEG Menggunakan Metode Fuzzy C-Means Clustering (FCM) Dan Adaptive Neighborhood Modified Backpropagation (ANMBP)," *Jurnal Matematika "MANTIK",* vol. 1, no. 1, pp. 31-36, 2015.

[8] D. C. R. Novitasari, "Klasifikasi Sinyal Eeg Menggunakan Metode Fuzzy C-Means Clustering (FCM) dan Adaptive Neuro Fuzzy Inference System (ANFIS)," ITS, 2013.

[9] F. Febrianti, M. Hafiyusholeh and A. H. Asyhar, "Perbandingan Pengklusteran Data Iris Menggunakan Metode K-Means dan Fuzzy C-Means," *Jurnal Matematika "MANTIK",* vol. 2, no. 1, pp. 7-13, 2016.

[10] A. A. I. Wiratmaka, I. F. R. Asmara and R. Andrie, "Klasifikasi Kualitas Tanaman Cabai Menggunakan Metode Fuzzy K-nearst Neighbor (Fknn)," *Jurnal Informatika Polinema,* vol. 3, no. 3, p. 1, Maret 2017.

[11] S. A. Dudani, "The Distance-Weighted k-Nearest-Neighbor Rule," *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS,* pp. 325-327, 1976.

[12] I. Jolliffe, Principal Component Analysis, 2nd ed., New York: Springer-Verlag, 2002.

[13] R. Susetyoko and E. Purwantini, Teknik Reduksi Dimensi Menggunakan Komponen Utama Data Partisi Pada Pengklasifikasian Data Berdimensi Tinggi dengan Ukuran Sample Kecil, Surabaya: EEPIS Repository, 2011.

[14] E. Prasetyo, Data Mining, Mengolah Data menjadi Informasi menggunakan Matlab 1st Ed., Yogyakarta: ANDI , 2014.

[15] A. S. P. A. Dewi and I. Candra, "Implementasi Algoritma Fuzzy K-Nearest Neighbor untuk Penentuan Lulus Tepat Waktu (Studi Kasus : Fakultas Ilmu Komputer Universitas Brawijaya)," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer,* vol. 3, no. 3, pp. 1726-1732, April 2018.

[16] K. Toduka and Y. Endo, "Fuzzy K-nearest Neighbor and Its Application to Recognize of te Driving Environtment," *IEEE International Conference on Fuzzy Systems,* 2006.