Handwritten digit recognition using discriminant analysis

Kamil Błoński, Oliwia Brożek
Faculty of Applied Mathematics, Sielsian University of Technology Gliwice, Poland
Email: kamiblo932@student.polsl.pl, oliwbro901@student.polsl.pl

***Abstract*—The main task of pattern analysis is to find and learn types of relations (clusters, correlations, classifications and so on) in datasets. The article describes discriminant analysis and one of its many applications. This statistics and machine learning method can be used in the problem of handwritten digit recognition. The mathematics part is divided into two sections. The first one refers to Linear Discriminant Analysis (LDA), while the second corncerns its extended version - Kernel Discriminant Analysis (KDA), also known as generalized discriminant analysis or Kernel Fisher discriminant analysis (KFD). There is an overview of kernel functions and kernel trick tool. Next, the proposed digits recognition system is described. That is the handwritten digit recognition application using "KernelDiscriminantAnalysis Class" from Accord.NET Framework. Training data set was subjected to the "Learn Method" and after analysis, testing data set was processed by "Classify Method". The methods belongs to the above mentioned class. The application and its accuracy was tested according to the values of Gaussian kernel function parameters.**

I. Introduction

Discriminant analysis allows to examine differences between groups of objects based on a set of independent variables – predictors. It is a statistical method. The author of this concept was Ronald Fisher, an outstanding mathematician and geneticist. He is known to social researchers more as the author of another statistical method – the variance analysis model. An example illustrating the operation of the discriminant analysis might come from pedagogical research. You can think about what data will allow you to separate secondary school leavers who decide to: (1) go to the university, (2) go to the work and stop learning, (3) take one-year break in learning [1]. In order to use the described method, the researcher must collect appropriate data from students (before graduation and making decision), which will be assigned to the relevant variables. Students from whom data was collected, after graduation, they will make a life decision in a natural way, which will automatically assign them to one of the mentioned groups. Thanks to the data collected in this way, the researcher, in the future, is able to determine whether any of the variables has a significiant impact on the students who are making a decision regarding to their future. If such variable (or variables) exists, it may be used to predict decisions, which will be made by future generations of graduates. Such research could be used to help at work, for example, a school career counselor.

Discriminant analysis finds effective application in many fields of science. In the case of a professional group of psychologists, it can be used to select employees or recruit university students. Economists are able to determine the risk associated with a loan or clearly explain the economic differenced between various regions of the world. Doctors, on the other hand, can predict whether a patient has a chance to recover or even survive a certain treatment. There are many examples of discriminant analysis applications and it has a wide range of applications. This is due to the fact that the mathematical model, on the basis of discriminant analysis was created is relatively simple [2].

Discriminant analysis has also found its application in the field of science, which is an artificial intelligence. It is used to designate groups to which, for example, recognized objects will be assigned. Classification method – LDA (Linear Discriminant Analysis) - looks for a linear transformation by maximizing variance between classes and minimizing variances within classes. It turned out to be the right technique to distinguish between the classes of patterns. However, this is a linear method that may be less accurate when dealing with non-linearity problems. Sometimes in this situations we also use neural networks models composed with other as complex prediction in technical systems [3] or medical diagnosis [4]. The KDA (Kernel Discriminant Analysis) can be used to isolate non-linear discriminant features. It is a non-linear method based on kernel techniques.

II. Discriminant analysis

Discriminant analysis is a set of discriminatory and classification methods. Discriminatory methods are aimed at determining which of the available variables differentiate (discriminate) groups of objects created due to some known identifiers (characteristics) of these groups. Discriminatory variables are not correlated with each other, and thus they do not duplicate information about the examined objects, transferring at the same time the information contained in the input variables. Discriminatory functions are determined in such a way to maximize the ratio of intergroup diversity of input variables to their intra-group diversity, i.e. they strive for the optimal division of objects into groups. At the same time, an assessment is made of which variables most strongly differentiate (discriminate) groups of objects. Classification methods are used to determine to which of the created groups should be

assigned a given object, using for this purpose those variables that had the greatest discriminatory power.

A. LDA - Linear Discriminant Analysis

This is the method used to find a linear combination of features that best distinguish between two (or more) object classes or events. Discriminant analysis aims to make each observation $x \in X$ assign a class to which this observation belongs. The discriminant rule divides the set X into g of disjoint subsets, called classes. This can be saved as a function $d(x) : X \longrightarrow G$, where G is a finite g-element set of class labels to which observations belong. Fisher's discriminatory analysis is defined as the historically first approach to classification under supervision. He assumes that the vectors of observation are vectors in the p - dimensional space $X \in R^p$ and leads to a discriminatory rule based on a linear function. This rule for $g = 2$ tries to find the direction a in X, which best separates both learning attempts, while constructing a measure of the distance between classes, taking into account intra-group variability. The intra-group dispersion should be characterized by sub-samples based on covariance matrices. The Fisher's method of constructing the LDA method requires aggregating information about the classes to which new observations should be classified, using position indicators and scattering g subsamples of the learning sample. If we only have two classes, then try $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ can be divided into a sub-sample of observations from the first class and for the attempt of observations from the second class:

$$x_{11}, x_{12}, \ldots, x_{1n_1}$$

$$x_{21}, x_{22}, \ldots, x_{2n_2}$$

Where $n = n_1 + n_2$. Then the avarage of classes can be saved as:

$$\overline{x_k} = \frac{1}{n_k} \sum_{i=1}^{n_k} x_{ki} \tag{1}$$

If a $\overline{x_k}$ will be defined as a group mean assuming that the variance of the X vector will be the same in all populations (here in two, $k = 1, 2$), this common covariance matrix will be determined by calculating the intra-group covariance matrix:

$$W = \frac{1}{n-2} \sum_{k=1}^{2} (n_k - 1) S_k \tag{2}$$

$$W = \frac{1}{n-2} \sum_{k=1}^{2} \sum_{l=1}^{n_k} (x_{kl} - \overline{x_k})(x_{kl} - \overline{x_k})' \tag{3}$$

Where S_k means the covariance matrix of the k^{th} population sample, and x' is the transposition of the x vector. Fisher defined the task of discriminant analysis in the following way - find the direction a, which maximizes the distance between the dropped averages of both samples, taking into account the variance of the drop.

$$argmax_a = \left(\frac{a'\overline{x_2} - a'\overline{x_1}}{\sqrt{a'Wa(\frac{1}{n_1} + \frac{1}{n_2})}}\right)^2 \tag{4}$$

$$argmax_a = \frac{(a'\overline{x_2} - a'\overline{x_1})^2}{a'Wa} \tag{5}$$

The solution is:

$$a = W^{-1}(\overline{x_2} - \overline{x_1}) \tag{6}$$

The designation a does not create a discriminatory rule yet. To build it, you need to drop x observations into the direction a. Moreover, the observation is classified into the first or second class, depending on whether the drop was closer to the center of the sample of the first or second class.

$$d = 2 \Longleftrightarrow a'(x - (\frac{\overline{x_1} + \overline{x_2}}{2})) > 0 \tag{7}$$

The idea of discriminant analysis is illustrated very well by the following illustration.

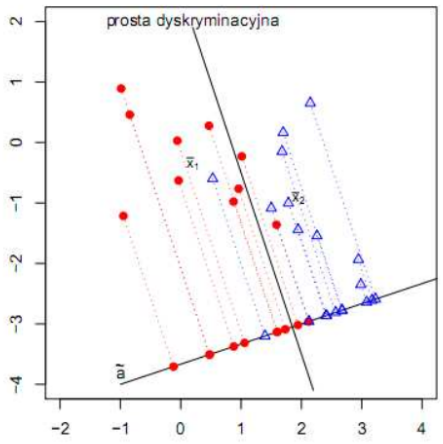


Figure 1: PhD Kornacki "Statystyczne systemy uczące się". Observation chart for classes - triangles and circles. [5]

It can be seen that observations from two classes (circles and triangles) are presented as points in two-dimensional space. Points $\overline{x_1}, \overline{x_2}$ represent respectively the average of the first and second groups. The drawing shows the observation and averages of samples for each class on the line described as $\tilde{a}$. To this line, another straight line was led - named discriminatory line.[6]

B. KDA

Kernel Fisher discriminant analysis (KFDA), also known as Generalized Discriminant Analysis (GDA) and Kernel Discriminant Analysis (KDA) is named as "general" discriminant analysis because it uses general linear model to analysis of discriminant function issues. The matter of discriminant function analysis was formed to general multidimensional linear model. In this model, analysing dependent variables are encoded (zero-ones) vectors that reprezent each of the cases belonging to specific group [1].

One of the advantages of applying geneal linear model to discriminant analysis issues is ability to define complex

models for predictors set. For example, for continous predictors set we can define polynomial regression model, response surface model, fractional regression model or response surface regression model for mixture (without constant term). Therefore, an experiment with limited mixture (exemplary limitation could be values of predictors summing up to constant value) could be analysied. In the experiment the dependent variable is inherently qualitative.

Besides stepwise discriminant analysis for predictors with one degree of freedom executed in discriminant analysis, General Discriminant Analysis enables applying stepwise analysis and best subset analysis too. In particular, there is a choice of predictors or a set of predictors (regarding the efects of many degrees of freedom that reflects qualitative predictors), stepwise method and best subset method on the basis of stathistics F. Moreover, if the test of cross-validation is specified, then selection using best subset method can be carried out in accordance with erroneous classification indicators for this test. It means that after approximation, discriminant functions for given set of predictiors erroneous classification indicators are caluculated and the model (subset of predicators) which gives the lowest value of erroneous classification indicator in the test indented for cross-validation should be chosen. Therefore, it is powerful technique which enables to select models characterized by good predictive accuracy and simultaneously allowing to prevent excessive adjustment of model to data.

The other feature of discriminant analysis general models is an access to usability/response profiles creation. The predicted values are calculated for every dependent variable and single usability values are assigned. Next, the graphic conclusion can be created. It shows how predicted responses and usability values "behave" against range of predictors values. In General Discriminant Analysis we can create profiles whether streight lines of predicted values for encoded dependend values or create a posteriori probabilities classification profiles.

The last feature enables assesment of impact of various predictors values on predicted case classification and it is esspecialy useful in interpreting results for complex models considering qualitative and continious predictors and their mutual interactions. General Discriminant Analysis contains methods, which are making it very effective tool to classification issues and data mining techniques.

However, the vast majority of handbooks in which the analysis of discriminatory functions is discussed are limited to the description of simple and stepwise analysis, and only for continuous predictors with one degree of freedom. There are no studies on the issue of resilience and efficiency of these techniques, in the case of their generalization to the form presented in the General Discriminant Analysis. The use of the best subset method (in particular in relation to qualitative predicators or in the case of erroneous classification indicators in cross-validation) to choose the best subset of predictors should be treated more as a heuristic search method than a statistical analysis technique.

From a statistical point of view, the use of qualitative predictors or their effects in the discriminant analysis model may raise doubts. For example, we can use General Discriminant Analysis to analyze the 2x2 cardinality table, assuming one of the variables in such table for the dependent variable and the other for the predictor. It is clear that usege of General Discriminant Analysis would be unreasonable (although in most cases we will get results in line with those that we would receive as a result of the usual Chi-square test for the 2x2 table). On the other hand, if we treat the assessment of parameters calculated in General Discriminant Analysis as a solution to the system of linear equations obtained by the least squares method, then the use of qualitative predictor in General Discriminant Analysis is fully justified. What's more, in applied research there is often a situation in which we deal with a combination of continuous and qualitative predictors (e.g. income and age, which are continuous variables and professional status, which is a qualitative variable) on the basis of which we want to predict values for a dependent variable of a qualitative nature. In such cases, it can be very instructive to examine the case classification for specific models that include quality predictors and models that take into account possible interactions between qualitative and continuous variables. However, it should be emphasized again that the use of qualitative predictors in the analysis of discriminatory functions is not well documented in the literature and therefore it is necessary to take particular care when accepting the results of statistical significance tests as well as to draw final conclusions from the analysis.

C. Application of Kernels in discriminant analysis

There are two groups (classes) of data and we want to set a discriminant function that would allow us to set the boundary between both classes. The simplest function is the linear function $g(x) = w \cdot x + b$ but it turns out that this function may be ineffective when the boundary between the two groups considered in the space R^d is non-linear. If we expect a non-linear boundary between data groups, we can do the following: map the data by transforming ϕ into Hilbert space $\mathcal{F}$ with the scalar product defined as:

$$K(x, y) \doteq \phi(x) \cdot \phi(y). \tag{8}$$

The space $\mathcal{F}$ to which the mapping occurs is called the space of the transformed Feature Space variables. In this way Kernel K induces explicit mapping - when mapping ϕ is known, and implicit mapping when ϕ is unknown. So ϕ is a mapping function:

$$\phi : R^n \mapsto \mathcal{F}. \tag{9}$$

This property is the basis of the "kernel trick" used in the General Discriminant Analysis, Support-vector machine and other methods [7].

The Kernel trick mapping does not need to be computed. If the algorithm can be expressed only as an inner product between two vectors. This inner product must be swaped with the inner product from other matching space. That is the "trick": a dot product is alwayes replaced with a Kernel function.

The problem can be solved in the space $\mathcal{F}$ using simple linear algorithms. When using kernels, it is important to select the appropriate type of kernel. We choose a mapping $\phi(x)$ for which direct calculations of mapped data are not necessary, but for which kernels can be easily calculated in $\mathcal{F}$, which are expressions of form (1). The widely used transformations with this property are: linar kernel, polynomial kernel and Gaussian kernel.

The linear kernel is a basic kernel function. Kernel algorithms that use it are often equivalent to their non-kernel counterparts. The linear has the form:

$$K(x, y) = x^T y + c. \tag{10}$$

This kernel has one sigma parameter.

The polynomial kernel has the form:

$$K(x, y) = (\alpha x^T y + c)^d. \tag{11}$$

Polynomial kernels are proper for applications where the training data is normalized. The polynomial kernel has two parameters and the constant $c > 0$ (constant term of polynomial). If the parameter c is not specified, it is assumed $c = 0$ by default.

Gaussian Kernel has the form:

$$K(x, y) = \exp\left(-\frac{(x-y)^2}{2\sigma^2}\right). \tag{12}$$

The parameter sigma affects on the efficiency of the kernel, and must be precisely adjusted. If it is too high in value, the exponential will behave almost linearly. If it is too low, there would be a lack of regularization and the decision boundary would be very sensitive [8].

D. Classification

The main purpose of applying discriminant analysis is to create predictions of case classification for specific groups. If the model is known and the discrimination function has been entered, we may wonder how accurate we can predict to which group the case belongs.[1]

1) Classifying functions: It is a big mistake to confuse the discriminating function with the classifying function. Amount of classification function is exactly the same as number of groups. They are used to decide to which group the cases belong.

$$K_i = c_i + w_{i1}x1 + w_{i2}x_2 + \cdots + w_{ip}x_p \tag{13}$$

Where: i defines the group, $1, 2, 3, \ldots, p$ determines the number of variables, c_i it's constant for the i^{th} group, w_{ij} is the weight of the j$^{\text{th}}$ variable in the i$^{\text{th}}$ group, x_j is observed value for a given case of the j$^{\text{th}}$ variable.

2) Probability a priori and a posteriori: Probability a priori concerns the situation in which a given case is assigned to a given group without using knowledge about the values of discriminatory variables. This probability can strongly influence the correctness of the classification process. In many cases, it is used to improve prediction accuracy or minimize errors. In a situation where groups are not well separated and many cases are close to the boundary lines, the application of a priori probability is particularly important. The probability in which the values of discriminatory variables are used to determine that a case belongs to a given group is called a posteriori probability. This is a conditional probability. They can be used to increase the accuracy of precision using the Bayes formula:

$$P(G_k|X) = \frac{p_k \cdot P(X|G_k)}{\sum_{i=1}^{g} p_i \cdot P(X|G_i)} \tag{14}$$

Where: p_i is a priori probability, $P(G_k|X)$ is the conditional probability that a given case belongs to the group k (conditioned by the knowledge of discriminatory variables), $P(X|G_i)$ is the conditional probability of receiving a variable vector if it is known that the case belongs to the group k. Therefore, the expected probability of incorrect classification is form

$$\sum_{i}^{g} p_i \cdot \sum_{k=1; i \neq k}^{g} P(k|i) \tag{15}$$

Where: $P(k|i)$ is the probability of erroneously classifying the i^{th} group's object into the k^{th} group.

3) Mahalanobis distance: It is a measure of the distance between two points in the space defined by two or more correlated variables. This classification method consists in determining the distance measures of the individual case from the centroid of the group. In the two-dimensional space, the Mahalanobis distance is equal to the Euclidean distance (e.g. the distance measured by a ruler). In the case of a number of variables greater than three, they can not be represented on the chart. In such cases the distance of Euclid is not a proper measure, it is the distance of Mahalanobis.

III. Description of the proposed system

The proposed solution enables the recognition of handwritten digits, has been implemented in C# using the Accord.NET Framework. This section contains descriptions of methods used.

A. The method loading training and testing set from database.

This method supports loading the contents of a data file into a *DataGridView* component. It supports Menu and the *Open* element after selecting the file. The 32x32 matrix is sent to the buffer variable using the *ReadBlock* method from *StreamReader* class. If the number of readed characters is smaller than the length of the buffer reserved for the bitmap, or the variable that holds the line of text behind the matrix label is empty, loop execution is interrupted. From 1 to 500 loop interactions, extracted text data into a bitmap are placed in the *DataGridView* in the training section. From 1000 to 1500 iterations in the test section. The *Extract* function taking a bitmap as an argument, converts it to an array of numbers and this data is also entered in the *DataGridView*.

B. The trainig method using Kernel Fisher Discriminant Analysis.

The method supports the *Run Analysis* button. If in the *DataGridView component*, the number of rows in the training data section is 0, the user is asked to load the data. Otherwise, the input and output data are taken from the *DataGridView* to the input and output tables. Then the entire learning process is performed using KDA (Kernel Discriminant Analysis). The Accord.NET Framework has been used for this purpose. A Gaussian kernel (Accord Statistics package) was created with the specified parameter σ (selected in *NumericUpDown*). A Kernel Discriminant Analysis object is created with threshold and reguralization parameter specified by the user. Next, the *Learn* method from *KernelDiscriminantAnalysis* class is invoked, which takes input and output tables.

C. The classifying method.

This method is responsible for classifying data after learning by KDA analysis. If the rows in the *DataGridView* in the testing section are empty, the user is asked to load the data. If the data is not analysed yet, the user is asked to it before the classification starts. The test data and expected values are extracted from the *DataGridView* to the corresponding input and expected tables. Next, the result of the *Decide* method from the *Classifier class* that takes the input data as an argument is assigned to the output array. It takes decisions on the classification of data, correct or incorrect recognition. Then, if the output data from the output array is equal to the expected value, the row with the bitmap corresponding to this data is highlighted in green. Otherwise, the row color remains white. The progress of the classification is shown on the progress bar.

D. The method extracting a bitmap from text data

The Extract method takes the text as an argument (zero-one matrix with dimensions 32x32). The "0" character changes to a white pixel, "1" changes on a black pixel. In this way, it returns a bitmap.

In the figure 2 there is a number representation in a text file.

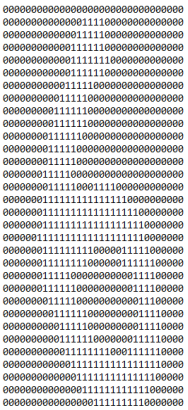


Figure 2: Caption

Bitmap representation of a number is shown in the figure 3 below.

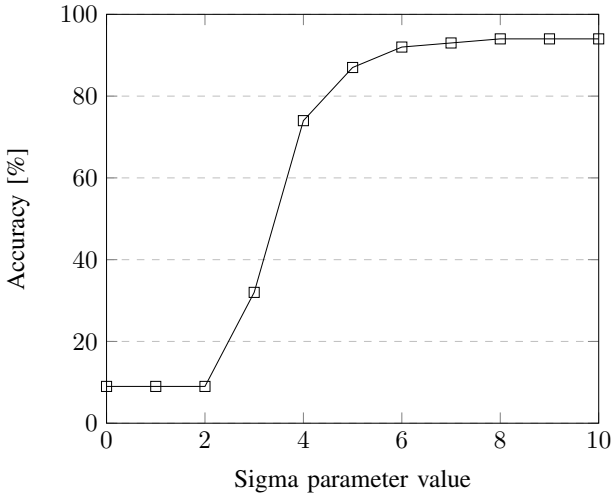
Figure 3: Caption

E. The method extracting text data from a bitmap.

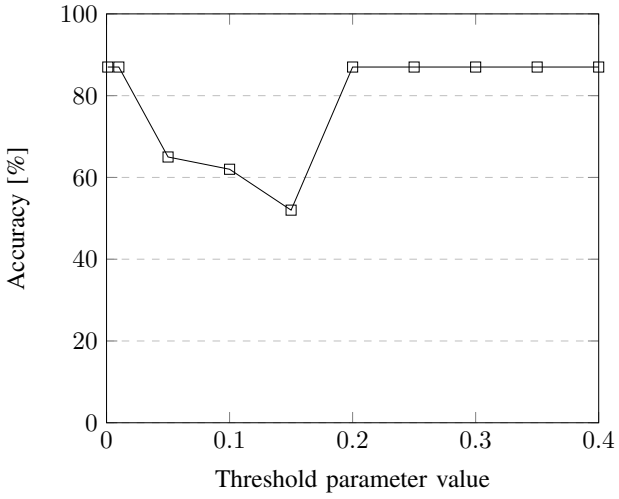
The Extract method that accepts the bitmap as an argument and returns an array of floating point numbers is an overload of the above method. Gets every pixel from a bitmap and checks if it is white. If so it types into the table 1, otherwise 0.

IV. Experiments

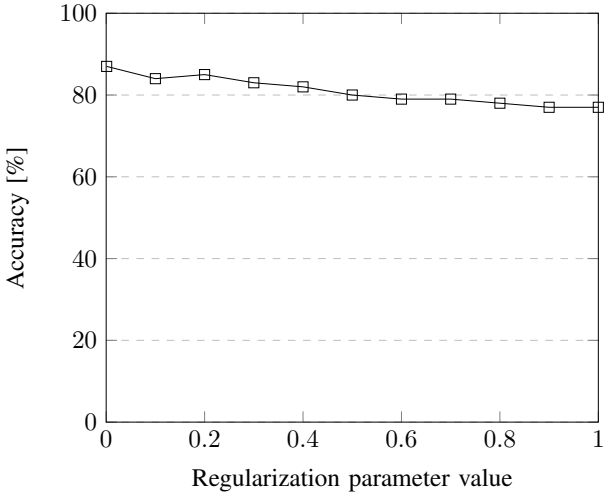
Classification accuracy depending on the sigma parameter



Classification accuracy depending on the threshold parameter



Classification accuracy depending on the regularization parameter



V. Conclusions

In this article it is shown how to recognize handwritten digits thanks to discriminant analysis application. The accuracy in recognizing images of digits databases taken out from UCI Machine Learning repository was checked up. It was presented that parameters tuning has a huge impact on algorithm accuracy. However, method used in described application had some limits. One of the obstacles is selection of appropriate discriminant function and its parameters which could give the best solutions. It is shown on the charts, how much impact parameters have on accuracy of application. Conversely, described method has superiority over other methods, for example neural networks. The is no problem of local minimas. So, there is no worry that obtain training result is not a valid result.

References

[1] Statsoft, "Elektroniczny podręcznik statystyki pl," http://www.statsoft.pl/textbook/stathome.html, 2006.

[2] P. Radkiewicz, "Analiza dyskryminacyjna, podstawowe założenia i zastosowania w badaniach społecznych." *Psychologia Społeczna 2010, tom5.*

[3] M. Woźniak and D. Połap, "Hybrid neuro-heuristic methodology for simulation and control of dynamic systems over time interval," *Neural Networks*, vol. 93, pp. 45–56, 2017.

[4] D. Połap, M. Woźniak, R. Damaševičius, and R. Maskeliūnas, "Bio-inspired voice evaluation mechanism," *Applied Soft Computing*, vol. 80, pp. 342–357, 2019.

[5] J. Kornacki, *Statystyczne systemy uczące się*. Exit, 2008.

[6] A. Nowak-Brzezińska, "Statystyczne metody analizy danych," http://zsi.tech.us.edu.pl/ñowak/smad/SMAD_lda.pdf, March 2010.

[7] A. Bartkowiak, "Kernele, sieci svm i sieci gda," https://www.ii.uni.wroc.pl/ aba/teach/NN/w11svm.pdf, January 2010.

[8] C. Souza, "Kernel functions for machine learning applications," http://crsouza.com/2010/03/17/kernel-functions-for-machine-learning-applications, March 2010.

[9] B. A. Nowak, R. K. Nowicki, M. Woźniak, and C. Napoli, "Multi-class nearest neighbour classifier for incomplete data handling," in *International Conference on Artificial Intelligence and Soft Computing*. Springer, 2015, pp. 469–480.

[10] M. Wróbel, J. T. Starczewski, and C. Napoli, "Handwriting recognition with extraction of letter fragments," in *International Conference on Artificial Intelligence and Soft Computing*. Springer, 2017, pp. 183–192.