Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th October 2023. Vol. 3. No. 1; e-ISSN : 2808-4519

Comparison Of K-Nearest Neighbor And CNN Classification Methods In Diabetic Data Sets

Ajeng Arina Nisa R[1], A M H Pardede[2], Marto Sihombing[3]

[1,2,3] *Information Systems, STMIK Kaputama*
Jl. Veterans No. 4A-9A, Binjai, North Sumatra, Indonesia
ajengarinanisa@gmail.com[1], akimmhp@live.com [2], martosihombing45@gmail.com [3]*

Abstract

The number of diabetics worldwide is projected to increase by 204 million (48%), from 425 million in 2017 to 629 million in 2045. Indonesia ranks sixth out of ten countries with the most number of diabetics in the world or 10 million people. The majority of people with diabetes are between 20 and 64 years old, or 327 million people, compared to 123 million people between 65 and 99 years old. The incidence of diabetes increases by about 4.8% at the age of 55-64 years, and women (1.7%) suffer from diabetes more than men (1.4%). Therefore, the authors will create a program to determine the patient's diabetes. One approach is to use machine learning as a data mining classification technique. The author will do a classification comparison with the two methods, namely the KNN and CNN methods to provide the best results of the two methods for testing. So that the accuracy of the data from the diagnosis and photo images of the disease can be known to provide early treatment before the severity of the disease.

Keywords: *Diabetes, Data Mining, K-Nearest Neighbor, Convolutional Neural Network, Classification, Phyton*

1. Introduction

Diabetes mellitus is a group of metabolic diseases characterized by hyperglycemia caused by defects in insulin secretion, insulin action or both. Diabetes is a chronic disease that lasts a lifetime and can cause complications in the form of organ damage which can lead to various conditions such as blindness, kidney failure, nerve damage, heart damage and diabetic feet, which can lead to amputation [1]. Diabetes is one of the most common chronic diseases and its prevalence is increasing worldwide due to population growth, aging, urbanization, obesity and physical inactivity. The number of diabetics worldwide is projected to increase by 204 million (48%), from 425 million in 2017 to 629 million in 2045. Indonesia ranks sixth out of ten countries with the most number of diabetics in the world or 10 million people. The majority of people with diabetes are between 20 and 64 years old, or 327 million people, compared to 123 million people between 65 and 99 years old. The incidence of diabetes increases by about 4.8% at the age of 55-64 years, and women (1.7%) suffer from diabetes more than men (1.4%). In addition, occupational factors (2.0%) and higher education (2.5%) are indicators of an increase in the incidence of diabetes because it is higher than the prevalence of diabetes (1.5%) Diabetes is not only life threatening, but it can also affect quality of life and cause many other health problems. Diabetes is an incurable disease. Therefore, early detection is the only way to treat and control this disease. Therefore, it is necessary to work hard to reduce the number of people with diabetes, and use diagnostic tests to determine whether a person has diabetes, so that comprehensive diabetes prevention and treatment can be carried out. One approach is to use machine learning as a data mining classification technique. CNN is an efficient recognition algorithm that is widely used in pattern recognition and image processing. It has many features such as simple structure, less training parameters and adaptability. This study will use retinal images of patients, to understand their physiological condition, potential health risks, challenges, and condition of the pancreas. Therefore, this study aims to analyze the comparison of the k-nearest neighbor classification method with CNN in a dataset of people with diabetes.

2. Theoretical basis

2.1. Data Mining

Data mining is the process of managing large data so that these data can provide accurate information and can facilitate problem solving and decision making. The other names for data mining are Pattern Analysis, Knowledge Extraction, Information Harvesting, and so on.

2.2. Classification

Classification comes from the Latin word classis, which means grouping similar objects and separating dissimilar objects. Literally the meaning of classification is classification, grouping. In relation to the library world, classification is defined as the activity of grouping library materials based on their characteristicsthe same, for example author, physical, content and so on.

2.3. K-Nearest Neighbor (KNN)

K-NN is a group that has an instance-based learning system, in conducting group searches by entering k object values into the test with the closest value to other data values. KNN uses the closest distance value to the dataset being tested in carrying out its classification process. This approach is taken in looking for a problem in calculations at the shortest distance between the new problem and the previous one with weighting by equating it with the number of existing features [2], [3].

2.4. Convolutional Neural Network (CNN)

Convolutional Neural Networks (CNN) is a variation of an artificial neural network that has weights and several hidden layers arranged into an architecture [4]. There are several layers in the CNN model, namely the convolution layer, activation function, pooling layer, flatten layer, and fully connected layer [5].

2.5. Image

Image is another term for an image as a multimedia component which plays a very important role as a form of visual information. Images have characteristics that text data does not have, namely images that are rich in information. Literally, an image is an image in a two-dimensional (two-dimensional) field. From a mathematical point of view, the image is a continuous function of the light intensity in the two-dimensional plane. The light source illuminates the object, the object reflects back as a beam of lightThis reflection of light is captured by optical devices, for example the human eye, cameras, scanners, and so on. So that the shadow of the object called the image is captured [6].

2.6. Diabetes

Diabetes is a chronic disease because the pancreas cannot produce more insulin (a hormone that regulates blood sugar) when the body cannot use the insulin it produces effectively. Diabetes Mellitus is one of the four non-communicable diseases, and is a very important public health problem.The number of cases of people with Diabetes Mellitus continues to increase [7], [8].

2.7. Python Programming Language

Python is a programming language that is freeware or freeware in the truest sense of the word, there are no restrictions on copying or distributing it. Complete with source code, debugger and profiler, the interface contained therein for interface services, system functions, GUI (Graphical User Interface), and database.

3. Results and Discussion

The purpose of this study is to make a comparison of the K-Nearest Neighbor (KNN) and Convolutional Neural Network (CNN) methods in detecting and accuracies of diabetes.

3.1. Research Materials

Before continuing to code the system to do a comparison of classification methods K-Nearest Neighborand CNN on the data set of people with diabetes, the first step that needs to be done is to install Python software. The python software used is Anaconda3 with version 2023. The application installation can be seen in the image below:

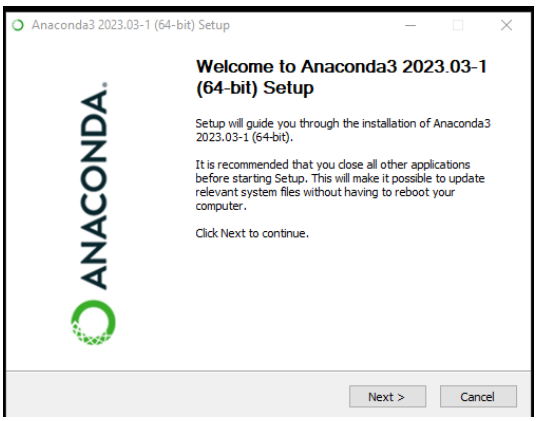

Fig. 1: DisplayInstalling Anaconda3

After the Anaconda3 installation is complete, proceed with installing the required modules so that errors do not occur when running the system later. The modules needed are:*tensorflow*, *scikit-learn*and opencv-python. Module installation can be done in the default anaconda3 application, namely "Anaconda Prompt" which can be seen in the image below:

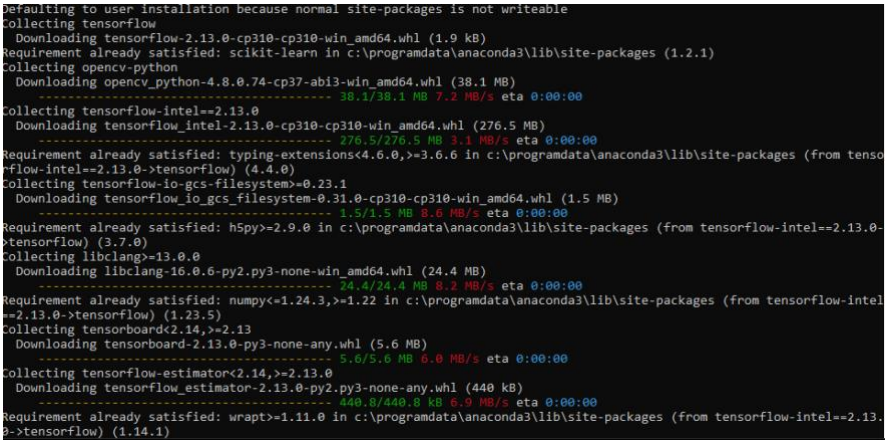

Fig. 2: Display Installation Module

3.2. Implementation

The comparison system for the K-Nearest Neighbor and CNN classification methods for people with diabetes, in the classification process, uses different datasets for diabetics. For classification using the KNN method using an excel file-based dataset and for classification using the CNN method using image file-based datasets. The dataset for classification is as follows:

KNN Method Dataset

Table 1: Diabetes Disease Dataset

No	Variable						Target
	X1	X2	X3	X4	X5	X6	
1	67	85	167	1	1	1	2
2	60	67	150	1	1	2	2
3	53	75	156	1	2	2	2
4	48	80	174	1	1	2	2
5	56	67	156	0	1	2	1
6	47	65	153	1	1	2	2
7	68	70	164	1	2	2	2
8	70	80	173	0	1	2	2
9	57	59	168	0	1	2	2
10	54	56	170	0	2	3	2
11	40	75	165	0	1	3	2
12	65	88	175	1	1	2	2
13	60	68	155	1	2	3	2
14	63	67	168	1	1	3	2
15	59	86	173	1	1	3	2
16	43	56	150	1	1	2	2
17	50	80	155	1	2	2	2
18	45	59	149	1	1	2	2
19	42	64	169	0	1	2	2
20	61	75	157	1	2	1	2
21	65	73	172	0	1	1	2
22	51	59	173	1	1	1	2
23	50	54	169	1	1	2	2
24	55	76	157	0	1	2	2
25	43	68	150	1	1	1	2

CNN Method Dataset

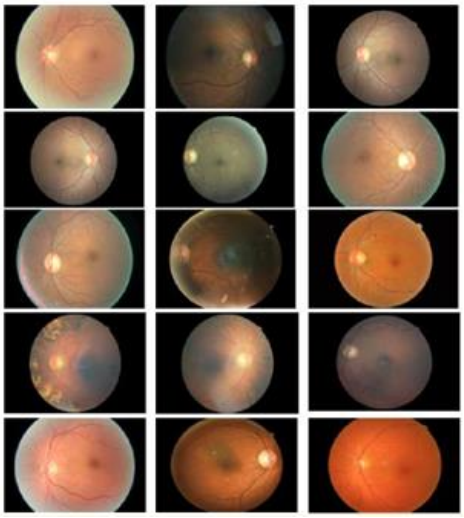

Fig. 3: Diabetes Dataset

3.3. Test Data

The process of testing the data to make comparisons on the KNN and CNN methods uses two different data for each method used. Classification using the KNN method uses diagnostic-based data while classification uses the CNN method using photo-based data. The data tested for the two methods used are as follows:

KNN Method Data

Table 2: Diagnostic Data

X1	X2	X3	X4	X5	X6
46	54	155	1	0	2

CNN Method Data

Fig. 4: Diabetes Disease's photo

3.4. Test Result

Based on the data tested using the two methods, the following results were obtained:

KNN Method Results

The results obtained from the system using the KNN method can detect diseases based on diagnostic data obtained from patients. The calculation of diagnostic data uses the previous dataset that has been made in the previous table.

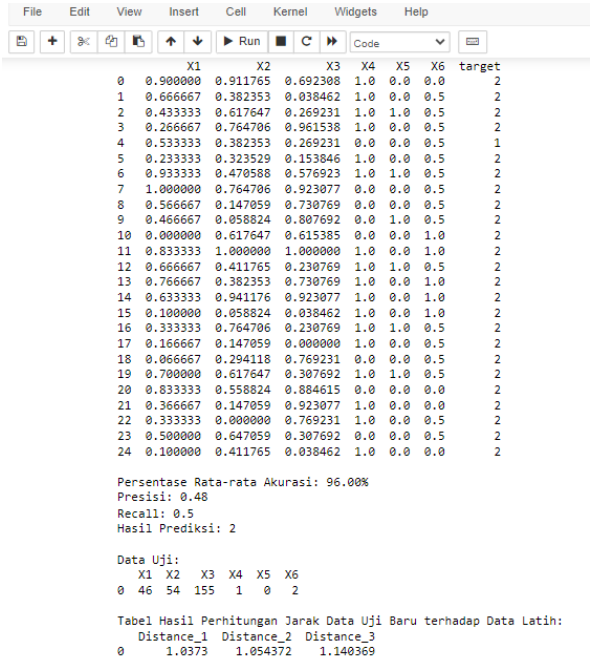

Fig 5: Results Using the KNN Method

Table 3: Test data

X1	X2	X3	X4	X5	X6
46	54	155	1	0	2

Table 4: Dataset Accuracy

Precision	Recall	Average accuracy
0.48	0.5	96.00%

Table 5: Table of Distance K=3

Distance_1	Distance_2	Distance_3	Target
1.0373	1.054372	1.140369	2

Based on the results obtained, the detection results lead to target 2, which means the detection is correct that the patient has diabetes.

CNN Method Results

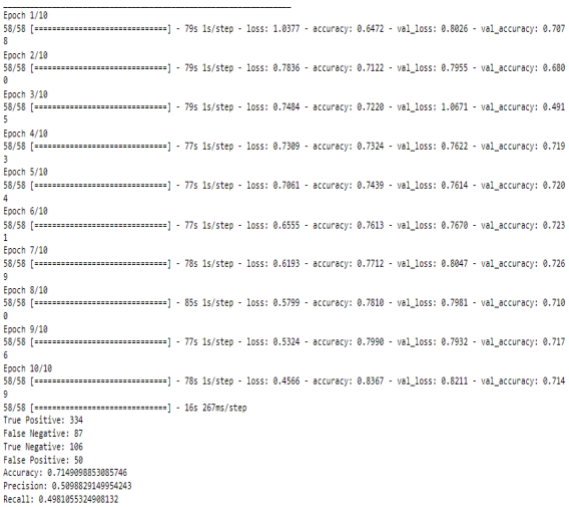

Fig. 6: Results Using the CNN Method

Fig. 7: Test Image

Table 6: Table Training

Epoch	*Iterations per epoch*	*Learning Rate*	*Validation frequency*
10	32	0.001	50

Table 7: Table of Accuracy and Loss

Epoch	*Processing Time*	*Accuracy*	*Loss*
1	70 sec	0.6423	1.1058
2	71 sec	0.7073	0.7843
3	69 seconds	0.7176	0.7562
4	68 seconds	0.7351	0.7249
5	69 seconds	0.7395	0.7013
6	66 seconds	0.7471	0.6764
7	69 seconds	0.7608	0.6383
8	71 sec	0.7712	0.5914
9	70 sec	0.8165	0.5326
10	69 seconds	0.8280	0.5039

```
1/1 [==============================] - 0s 121ms/step
label_dict: {'Mild': 0, 'Moderate': 1, 'No_DR': 2, 'Proliferate_DR': 3, 'Severe': 4}
prediction: [[0.01435558 0.26334026 0.01963569 0.66026384 0.04240463]]
predicted_class_idx: 3
predicted_class_label: Proliferate_DR
Predicted Class: Proliferate_DR
Prediction Score: 0.66026384
```

Fig. 8: Results Using the CNN Method

Table 8: Result

Mark	**Percentage**	**Target**
0.9499877	94%	2

Based on the results obtained, the detection results using image test data lead to target 2, which means that the detection is correct that the patient has diabetes.

4. Conclusion

Based on the results of the comparative design of the K-Nearest Neighbor and CNN classification methods in the data set of people with diabetes, in making the training dataset for testing, the percentage results were 96% in the KNN method and 94% in the CNN method. To detect diabetes, both the KNN and CNN methods are capable of detecting the disease, it's just that by using the CNN method there is a percentage given so that the accuracy of the data is more convincing, whereas by using the KNN method only the results of the target disease are given without any percentage given.

5. SUGGESTION

The suggestions that the writer can convey are as follows:

1. The system is designed only for comparison of detection results between the KNN and CNN methods in knowing the percentage of detection using both methods by using datasets for the KNN method and image images for the CNN method.

2. The system created to make a comparison of the KNN and CNN methods uses a console-based system or application using the python programming language. This system is not made available to the public but only to do a comparison of the two methods.

Reference

[1] Argina, A.M. (2020). Application of the K-Nearest Neigbor Classification Method to Diabetic Datasets. Indonesian Journal of Data and Science, 1(2), 29-33.
[2] Muslih, M., & Rachmawanto, E. H. (2022). Convolutional Neural Network (CNN) for Image Classification of Diabetic Retinopathy. SKANIKA, 5(2), 167-176.
[3] Sholeh, M., Andayati, D., & Rachmawati, R. Y. (2022). DATA MINING CLASSIFICATION MODEL USING K-NEAREST NEIGHBOR ALGORITHM WITH NORMALIZATION FOR DIABETES PREDICTION. TeIKA, 12(02), 77-87.
[4] Syahrul, F. H., & Sasongko, P. S. Application of the Convolutional Neural Network for Classifying the Severity of Diabetic Retinopathy in Patients with Diabetes Mellitus. Journal of the Informatics Society, 13(1), 1-14.
[5] Marcella, Dewi, Yohannes Yohannes, and Siska Devella. "Classification of eye disease using Convolutional Neural Network with VGG-19 architecture." Journal of Algorithms 3.1 (2022): 60-70.
[6] Agustin, Tinuk. ANALYSIS OF THE COMBINATION OF CONVOLUTIONAL NEURAL NETWORK (CNN) AND SUPPORT VECTOR MACHINE (SVM) IN THE AUTOMATIC DETECTION OF NON-PROLIFERATIVE DIABETIC RETINOPATHY. Diss. AMIKOM Yogyakarta University, 2021.
[7] Zakiya, Putri Nada, and Ledya Novamizanti. "Classification of Retinal Macular Pathology Through Oct Image Using Convolutional Neural Network With Mobilenet Architecture." eProceedings of Engineering 8.5 (2021).
[8] Putry, N.M. (2022). COMPARISON OF KNN AND NAÏVE BAYES ALGORITHMS FOR DIABETES MELLITUS DIAGNOSIS CLASSIFICATION. Evolution: Journal of Science and Management, 10(1).