Prediction of neurological diseases using data mining

Gholam Hassan Shirdel[1]*, Mojtaba Ghanbari[2], Shabnam Zarghami[3]

[1]* [3]Department of Mathematics, University of Qom, Qom, Iran.
[2]Department of Mathematics, Farahan Azad University, Farahan, Iran.

*Corresponding Author: Gholam Hassan Shirdel
*Department of Mathematics, University of Qom, Qom, Iran, Email: shirdel81math@gmail.com.

Abstract

Data mining is one of the stages of acquiring knowledge in a database to collect useful information. Data mining is a new field that has various applications and it is known as one of the top ten sciences affecting technology. Data mining analyzes databases and massive data sets in order to discover and extract knowledge, and machine (and semi-machine) mines. Such studies and explorations can actually be considered the extension and continuation of the ancient and ubiquitous knowledge of statistics. The major difference is the scale, breadth and variety of fields and applications, as well as the dimensions and sizes of today's data, which require machine learning, modeling, and training methods. In the 1960s, statisticians used the term "Data Fishing" or "Data Dredging" to discover any relationship in a very large volume of data without considering any assumptions. After thirty years and with the accumulation of data in databases, the term "Data Mining" became more popular around 1990. The purpose of this research is to predict brain and nerve diseases using data mining algorithms. The purpose of this research is to help medical professionals to predict disease. In this research, after data preparation, disease prediction has been attempted using large matrix methods and data mining techniques. By examining the new vector, we can find out which of the diseases in the matrix will be closer to this new disease with new symptoms using the rows of the matrix. The conducted research is one of descriptive-analytical and applied studies. In this research, we used different meters such as Jacquard distance, cosine similarity L_1-norm and cosine similarity L_2-norm implemented a program using Python software to predict brain and neurological diseases. The algorithm implemented by Python software, the doctor enters the symptoms of the patient and the output of the program shows three diseases close to the input symptoms for each meter, and finally all the meters are compared and the meter that has a weaker result is determined each time it is run. The advantages of each of these meters are explained below.

***Keywords:** prediction of neurological, treatment, treatment methods, diseases, data mining.*

1- Introduction

The ever-growing healthcare industry is generating large amounts of useful data on patient demographics, treatment plans, payment, and insurance coverage that are attracting the attention of clinicians and scientists alike (1, 2). A number of peer-reviewed articles have focused on various dimensions of data mining

applications in healthcare in recent years (3). Some data is collected and processed for different purposes due to the ever-increasing advancements of technology in all different fields.

Data mining refers to the process of determining and analyzing hidden information from different perspectives to obtain useful knowledge. Data mining has various applications, one of which is the diagnosis of various diseases in the field of medicine. Nowadays, many ailments and diseases are dangerous and deadly. (4)

We hope that this study makes it possible to obtain useful and effective information for future research. The article is organized as follows: Section 2 deals with the methods and techniques of knowledge discovery in databases and the concepts of data mining and the research strategy used in this research. Part 3 focuses on part of the data related to symptoms and diseases extracted from Aminov's book and expert consultations during many meetings with a neurologist. Section 4 deals with the diagnosis and prediction algorithm of the disease. Section 5 focuses on the implementation of code implemented by Python software to predict diseases. Section 6 presents the conclusion.

After this matrix was created using different methods in data mining, we will discuss this issue that if the disease has its own characteristic symptoms, it enters the software as an input, and three outputs are the result of the algorithm that we implemented using Python software and the Python program implemented by us, which is the closest diseases to the input symptoms, or to be more precise, using large matrix methods and data mining techniques, after preparing the matrix, if we have a disease with symptoms, we can check the new vector It was found that this new disease with new symptoms by using the rows of the matrix that this disease will be closer to which of the matrix diseases.

The next step compares the different data mining methods used for this matrix and it is observed which one gives a more optimal answer or its error is less.

The important result of this research is to choose the best method that has the least time complexity to achieve the result. The noteworthy point is that this method can be extended to many other situations.

The ever-increasing advances in health science have led to an increase in the lifespan of human societies, a decrease in deaths, and an increase in the elderly population.

2. Material and method

This study is considered as a descriptive-analytical and applied study. Data mining is one of the best research methods that have been used. In this research, various data mining methods and techniques have been used to predict and diagnose the disease early (5, 6, 7).

2-1- Knowledge of the system

It is very vital and necessary to know the field in which data mining should be done and also to have relevant knowledge to conduct this research. Therefore, in the first step, an attempt has been made to get a proper understanding of the field under study through consultation with a neurologist, a detailed study of Aminoff's book , as well as a study of diseases of the brain and nerves, determination of factors affecting the disease, treatment methods, diagnosis and methods of preventing this disease.

2-2- Data preparation (disease and symptoms)

The data used in this study are taken from Aminoff's book clinical neurology book, consultation with neurologist and clinical data. A matrix with about 150 rows and 500 columns is formed after consultation with the relevant doctor and using clinical data in the archive, and the rows of this matrix represent the jth symptom for the ith disease. Studying the texts and consulting with a specialist doctor was designed.

The set of diseases and data in the matrix is defined as a table in Excel, Table (2), (1) is considered as an example whose rows show the diseases and its columns show the symptoms, the method of data collection is presented according to the description. Table (3) is considered as an example showing symptom codes and Table (4) as an example showing disease codes. Disease diagnosis code was modeled using Python programming and data mining techniques such as Jacquard distance, L1-soft cosine similarity and L2-soft cosine similarity. For the types of symptoms (matrix column), data (matrix row) 1 to 150 have been obtained, which represent diseases.

Charts **(Figures 1-3)** help us find the correct understanding of the data. By drawing graphs, we try to make numerical information in a state that is possible for humans to understand, because mere numerical information does not help us, and it is by modeling and analyzing the structure of these data that we can have a correct understanding of the reality behind these numbers. One of the most important diagrams is the heatmap diagram. In fact, the purpose of this chart is to create an initial clustering and display numerical information in the form of color. In the figure below, you can see a heat map, which is displayed in the column and row of numbers in

the form of colors. Each cell of this spectrum chart represents a numerical value. This spectrum is shown in the figure with different colors. Numbers with values below zero are displayed in red and above in blue, and zero values are in black. By looking at this graph, we can see that numbers with values below zero are displayed in yellow and above in blue. By looking at this chart, we can see which part has what value.

There are several varieties for the mean in mathematics and especially in statistics. In the study of the distribution of a statistical population, the representative value around which the measurements are distributed is called the central value, and any numerical measure that represents the center of the data set is called the measure of central tendency. Mean and median are the most common measures of center tendency.

Median in statistics and probability theory is one of the measurements of the tendency to the center. The median is a number that divides a statistical population or a probability distribution into two equal parts. One of the important advantages of the median over the mean is that the mean is not affected by very large and very small numbers in the set of measurements. One of the most important properties of the median is that the sum of the absolute value of the differences of different values of the random variable is the minimum.

Standard deviation **(figure1)** (symbol σ) is one of the dispersion indices that shows how far the data is from the average value. If the standard deviation of a set of data is close to zero, it is a sign that the data are close to the mean and have little dispersion; While the large standard deviation indicates the significant dispersion of the data. The standard deviation is equal to the square root of the variance. Its advantage over variance is that it is also dimensional with data. Standard deviation is also used to determine the reliability coefficient in statistical analysis. In scientific studies, usually data with a difference of more than two standard deviations from the mean value are considered as outliers and are excluded from the analysis.

Table 1- Symptoms (headache) (7) The set of diseases and data in the matrix is defined as a table in Excel.

Figure 1 whose line shows the diseases and the numbers opposite the lines are actually the same columns that show the symptoms. It has been numbered by a specialized study based on the severity of symptoms in the range of 0 to 10.

Table 1. Symptoms A (headache) [7]

Symptoms (headaches)	Disease
10	**Subarachnoid hemorrhage**
9	**Meningitis or cephalitis**
9	**Hypertensive encephalopathy**
9	**Giant cell arthritis**
8	**Intracranial mass**
9	**Pseudotumor cerebri (false brain tumor)**
0	**Trigerminal veralgia**
0	**Neuralgia throat sufaz artery**
0	**Neuralgia following herpes**
9	**Hypertension**

Table 2. Symptoms of B (decreased level of consciousness, confusion)(7)

Symptoms (decreased level of consciousness, confusion)	Disease
9	**Subarachnoid hemorrhage**
9	**Meningitis or cephalitis**
0	**Hypertensive encephalopathy**
0	**Giant cell arthritis**
7	**Intracranial mass**
0	**Pseudotumor cerebri (false brain tumor)**
0	**Trigerminal veralgia**
0	**Neuralgia throat sufaz artery**
0	**Neuralgia following herpes**
0	**Hypertension**

Table 3. Some symptoms extracted from Aminoff's book and consultation with a specialist doctor (7)

Symptoms	Code	Symptoms	
Plantar reflex (bilateral extensor or plantar reflex)	G	Headache	A
Hemiparesis (paralysis of one limb or one half of the body)	H	Level reduction	B
Aphasia (language disorder)	I	Vomit	C
Visual Field Defects	J	neck stiffness	D
tentacle hernia	K	high blood pressure	E
Progressive drowsiness	L	Fever	F

Table 4- Some diseases extracted from Aminoff's book and consultation with a specialist doctor(7)

1) Subarachnoid hemorrhage
2) Meningitis or cephalitis
3) Hypertensive encephalopathy
4) Giant cell arthritis
5) Mass inside the skull
6) Cerebral pseudotumor (false brain tumor)
7) Trigerminal neuralgia
8) Neuralgia of the throat, sufase artery
9) neuralgia following herpes
10) Hypertension
11) Atypical facial pain
12) Migraine
13) Cluster headache
14) Tension-type headache
15) Ice-breaking pain

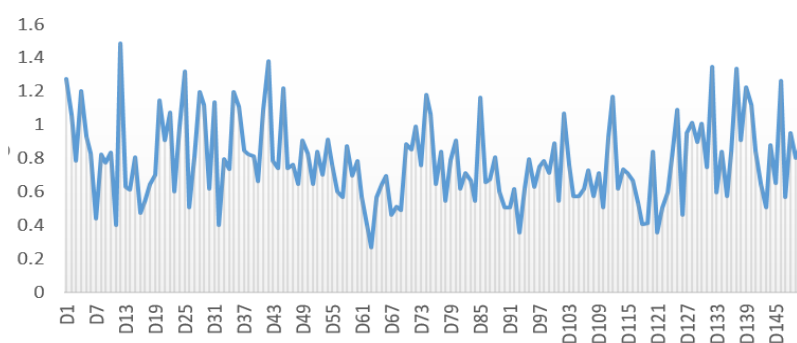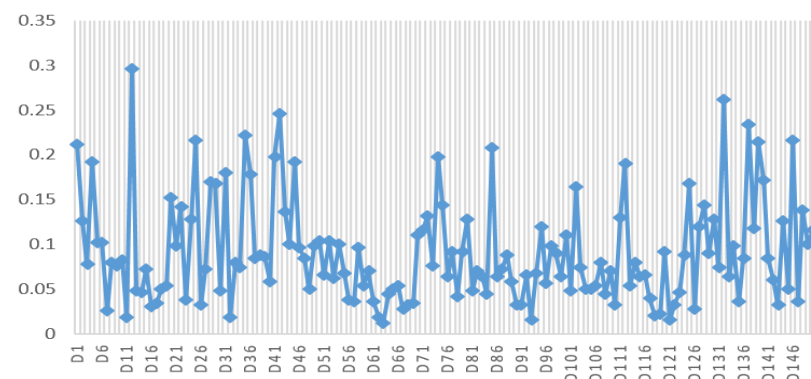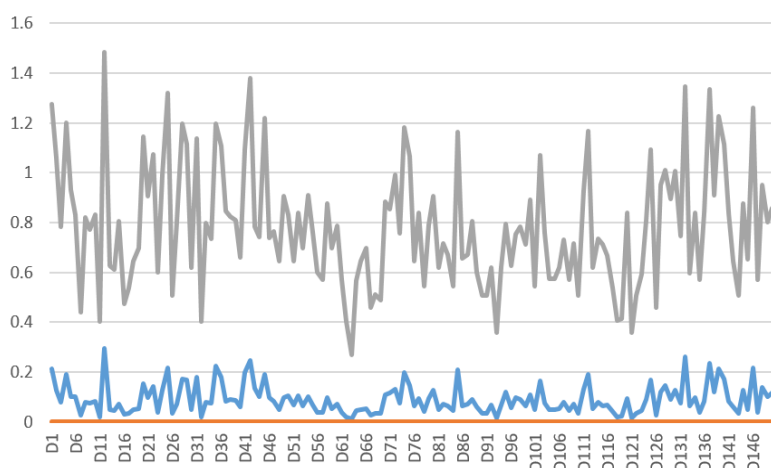**Figure 1.** Standard deviation data

Figure 2. Mean data

Figure 3. Check the median data and the standard deviation data and the mean data

2- 3-Modeling
Thus, there are various data mining methods for modeling; Therefore, in this study, modeling is done in Python software and predictive models are presented using data mining techniques.

2-4- Jaccard distance
The Jaccard similarity coefficient[1] was originally invented by Paul Jaccard, which is used to obtain a distance to show how close two sets are to each other, which was called the Jaccard index or the Jaccard similarity coefficient to find the similarity between Two items, its more formal form is written as follows:

$$similarity_{Jaccard}(i,j) = \frac{\#users\ that\ bought\ both\ items}{\#users\ who\ bought\ either\ i\ or\ j} \tag{1}$$

Where, i represents item 1 and j represents item 2.

2-5-Measuring the distance with Lp - norms
Lp-norms are a general way to measure distances, so in this section, we will take a look at two different L1-norms, L2- norms, and Lp- norms measures.

2-6- L1-norm (Manhattan distance)
Manhattan distance is the simplest measure of distance calculation, which is also famous for Taxicab distance, the advantage of this method is its high speed.
Manhattan distance calculation:
It is obtained from the sum of the absolute value of the x difference and the y difference (equation (2)).

$$|x_1 - x_2| + |y_1 - y_2| \tag{2}$$

2-7- L-2 Norm
L2-norm is also known as Euclidean norm and is defined as follows:

1) Distance (Sara,Pietro)

$$= \left\| r_{sara} - r_{pietro} \right\|_2 = \sqrt{\sum_{i=1}^{n} \left| r_{sara,i} - r_{pietro,i} \right|^2} \tag{3}$$

---

[1] Coefficient de communaté

2- 8-Cosine similarity

Cosine similarity is very well known in the field of text processing, this formula is also used in collaborative filtering. Cosine similarity ignores the 1-1 criteria. This index is introduced as equation (4):

1) $$\cos(x, y) = \frac{x, y}{\|x\| \times \|y\|}$$

2) $$sim\ (i, j)$$

3) $$= \frac{ri.rj}{|r_1\|_2 |r_j|_2} = \frac{\Sigma u^r i, u^r j, u}{\sqrt{\Sigma u^{r^2} i, u}\sqrt{\Sigma u^{r^2} j, u}}$$

(4)

1) $$\cos(x, y) = \frac{x, y}{\|x\| \times \|y\|}$$

2) $$sim\ (i, j)$$

3) $$= \frac{ri.rj}{|r_1\|_2 |r_j|_2} = \frac{\Sigma u^r i, u^r j, u}{\sqrt{\Sigma u^{r^2} i, u}\sqrt{\Sigma u^{r^2} j, u}}$$

(5)

3. Findings

1-3- Python code implemented on the data (5, 6)

The algorithm in the figure was implemented after collecting data using Python software. This algorithm was implemented with different Jaccard distance meters, L1-Norm cosine similarity and L2-Norm cosine similarity by Python software for additional explanations, which meters are fully explained in the above sections. These meters show the closest similarities to our input signals. When a patient with certain symptoms goes to the doctor, the doctor enters the symptoms into the software using the software implemented by us, and finally the output will show the doctor three diseases close to the symptoms in order to predict the disease. Next, the meters were compared, then, the worst meter program is identified, which has poorer results than the rest of the meters. In the next section, which includes the implementation of the program, the visible results are presented for more information.

2-3- Code implementation and disease prediction

The code implementation for disease prediction is shown in **Figure 1** after data preparation and algorithm implementation in Table 5. Data mining is able to discover and extract new knowledge from retrospective data. The way of pre-processing the data as well as the selected variables has a significant effect on the discovery of knowledge. There are various data techniques that are used to predict diseases. Three data mining algorithms were used in this article, which will be explained later. According to the experimental results, the efficiency and effectiveness of all three methods were compared based on sensitivity, specificity and accuracy. Pruning and strengthening methods were used to find the desired structure and increase the accuracy of the results. This study has examined existing database, the predictive approaches of data mining in neurological diseases and their diagnosis. Therefore, new research can be developed in this field according to the notes mentioned about research gaps and the use of predictive data mining approaches in the early diagnosis of various (medical) diseases. Example-5-1- **in figure 4,** For example, we have examined a sample, and entered the following disease symptoms into the software as input, reduced level of consciousness, confusion (B=9), neck stiffness (D=8) and plantar reaction (Two-way extensor), plantar reflex (G=6), respectively, which we have implemented with Python to show the diseases close to this patient, and the output is shown with different meters respectively. Jacquard meter output shows three diseases, which are migraine, cerebellar degeneration caused by phenytoin and vasovagal syncope. The output of the Minkowski meter is muscular dystrophy, Duchenne dystrophy, and fasciosculo-brachial dystrophy. The output of the Minkowski meter is muscular dystrophy, Duchenne dystrophy, and Facioscapulohumeral Muscular Dystrophy. All the output diseases are close to the input symptoms in the field of movement disorders, Somesthetic Senses disorders, movement defects and vision defects. Our software has identified the weakest meter that shows weaker results than other

meters, which is the Jacquard meter. We reviewed the results with a neurologist, and the result of the review showed that the results are completely acceptable and accurate.

Table 5. Implemented Python code (5, 6): Algorithm implemented by Python software to predict brain and neurological diseases.

```python
from math import sqrt import pandas as pd import numpy from numpy import dot from collections
import Counter from itertools import chain def jaccard_similarity(arr1, arr2):
    intersection = len(list(set(arr1).intersection(arr2)))
    union = (len(set(arr1)) + len(set(arr2))) - intersection
    return float(intersection) / union
def l1_norm(arr):
    return sum([abs(i) for i in arr])
def l2_norm(arr):
    return sqrt(sum([pow(i, 2) for i in arr]))
def cosine_similarity_l1(arr1, arr2):
    return dot(arr1, arr2) / (l1_norm(arr1) * l1_norm(arr2))
def cosine_similarity_l2(arr1, arr2):
    return dot(arr1, arr2) / (l2_norm(arr1) * l2_norm(arr2))
def recommend(df, disease_name):
    distance_type = input(
        'Distance Type ( 1: Jaccard similarity, 2: Cosine similarity l1-norm, 3: Cosine similarity l2-
norm): ')
    do_compare = input('Do you want to compare distances? (y (yes), n (no)): ')
    distance = []
    distances = []
    for i in range(df.shape[0]):
        if i != disease_name:
            distance.append(jaccard_similarity(df.iloc[disease_name].values, df.iloc[i].values))
    distances.append(distance)
    distance = []
    for i in range(df.shape[0]):
        if i != disease_name:
            distance.append(cosine_similarity_l1(df.iloc[disease_name].values, df.iloc[i].values))
    distances.append(distance)
    distance = []
    for i in range(df.shape[0]):
        if i != disease_name:
            distance.append(cosine_similarity_l2(df.iloc[disease_name].values, df.iloc[i].values))
    distances.append(distance)
    if distance_type == '1':
        print('Jaccard similarity (top 3): ', numpy.argsort(distances[0])[0:3], sorted(distances[0])[0:3])
    if distance_type == '2':
        print('Cosine    similarity    l1-norm    (top    3):    ',    numpy.argsort(distances[1])[0:3],
sorted(distances[1])[0:3])
    if distance_type == '3':
        print('Cosine    similarity    l2-norm    (top    3):    ',    numpy.argsort(distances[2])[0:3],
sorted(distances[2])[0:3])
    if do_compare == 'y':
        print('Jaccard similarity (top 3): ', numpy.argsort(distances[0])[0:3], sorted(distances[0])[0:3])
        print('Cosine    similarity    l1-norm    (top    3):    ',    numpy.argsort(distances[1])[0:3],
sorted(distances[1])[0:3])
        print('Cosine    similarity    l2-norm    (top    3):    ',    numpy.argsort(distances[2])[0:3],
sorted(distances[2])[0:3])
        counts = Counter(chain(*map(set,
                        [sorted(distances[0])[0:3],                        sorted(distances[1])[0:3],
sorted(distances[2])[0:3]])))
```

```python
    common_remove = [[i for i in sublist if counts[i] == 1] for sublist in
                [sorted(distances[0])[0:3], sorted(distances[1])[0:3], sorted(distances[2])[0:3]]]
    list_size = []
    for i in range(len(common_remove)):
        list_size.append(len(common_remove[i]))
    max_size_list = list_size.index(max(list_size))

    if max_size_list == 0:
        print('\n' + 'Worst distance calculation is Jaccard similarity')
    if max_size_list == 1:
        print('\n' + 'Worst distance calculation is Cosine similarity l1-norm')
    if max_size_list == 2:
        print('\n' + 'Worst distance calculation is Cosine similarity l2-norm')
disease_name = 100
dfs = pd.read_excel('Future_Signs_final.xlsx', header=None)
first_row = dfs.iloc[0, 1:]
first_column = dfs.iloc[1:, 0]
first_row.to_excel("features.xlsx", sheet_name='Sheet1')
first_column.to_excel("disease_names.xlsx", sheet_name='Sheet1')
df = dfs.iloc[1:, 1:]
show_type = input(
    'Choose: 1: Input NoN-Zero Features From User, 2: Input Features From File, 3: Input all Features From user ')
if show_type == '1':
    how_many_number = int(input("How many non-zero Features do you have? "))
    print("Enter two values for %d times: First is the Feature Index (from 1 to %d) and Second is the Feature Value" % (
        how_many_number, (len(df.columns))))
    df.iloc[disease_name][1:len(df)] = 0
    for i in range(how_many_number):
        index, df.iloc[disease_name][index] = [int(x) for x in input().split()]
    recommend(df, int(disease_name))

if show_type == '2':
    disease_name = input('Enter the Number of the disease: ')
    recommend(df, int(disease_name))

if show_type == '3':
    print("Enter %d Feature Values" % (len(df.columns)))
    for i in range(len(df.columns)):
        df.iloc[disease_name][i] = input()
    recommend(df, int(disease_name))
```

```
how_many_number, (len(df.columns))))
    df.iloc[disease_name][1:len(df)] = 0
    for i in range(how_many_number):
        index, df.iloc[disease_name][index] = [int(x) for x in input().split()]
    recommend(df, int(disease_name))

if show_type == '2':
    disease_name = input('Enter the Number of the disease: ')
    recommend(df, int(disease_name))

if show_type == '3':
    print("Enter %d Feature Values" % (len(df.columns)))
    for i in range(len(df.columns)):
        df.iloc[disease_name][i] = input()
    recommend(df, int(disease_name))
```

```
Choose: 1: Input NoN-Zero Features From User, 2: Input Features From File, 3: Input all Features From user 1
How many non-zero Features do you have? 3
Enter two values for 3 times: First is the Feature Index (from 1 to 500) and Second is the Feature Value
2 9
4 8
7 6
Distance Type ( 1: Jaccard similarity, 2: Cosine similarity l1-norm, 3: Cosine similarity l2-norm): 1
Do you want to compare distances? (y (yes), n (no)): y
Jaccard similarity (top 3):  [ 12  37 146] [0.125, 0.125, 0.14285714285714285]
Jaccard similarity (top 3):  [ 12  37 146] [0.125, 0.125, 0.14285714285714285]
Cosine similarity l1-norm (top 3):  [  0 104 103] [0.0, 0.0, 0.0]
Cosine similarity l2-norm (top 3):  [  0 104 103] [0.0, 0.0, 0.0]

Worst distance calculation is Jaccard similarity
```

Figure 4. Disease and interpretation of results in example5-1

6- Conclusion

This study has been carried out with the aim of designing an efficient model to discover the knowledge of predicting neurological diseases based on the latest data set of indicators in this field that are related to the community health, and has presented the analysis of accurate data mining techniques for predicting neurological diseases. In other words, this study has presented efforts made by research to apply data mining techniques based on the use of disease and symptom data sets through the business intelligence program to provide important results in making the right decisions at the right time. Different metrics have been used to provide complete explanations and conclusions, and the algorithm has been implemented using Python software to predict neurological diseases. The advantages of each of these meters are explained below. Distance measures such as Euclidean or Manhattan are used for dense data (most features have non-zero numbers), and the coefficient of feature values is of great importance.
The cosine similarity measure is used for sparse data. Therefore, it is possible to analyze very big medical data and discover very useful knowledge and decision-making using data mining methods. Data mining can be used as a guide for doctors in predicting neurological diseases and the accuracy of the results obtained from the models is also quite close to reality.these results are shown in **Figure 5**.
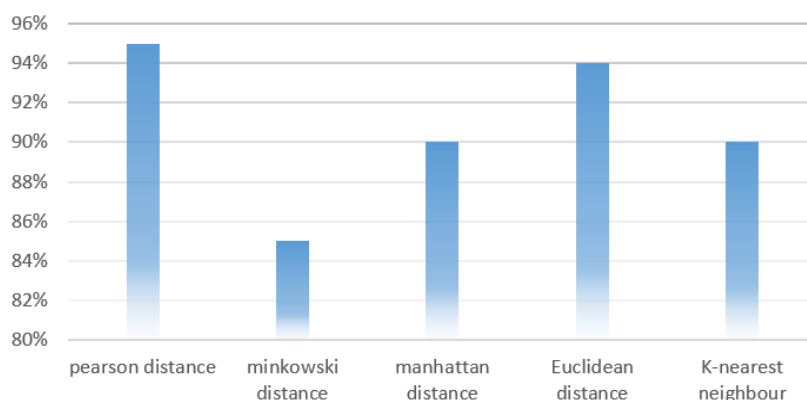

Figure 5. the result of comparison of algorithm and meters in predicting diseases using data mining

7-Acknowledgments

8-Declarations

Regarding the statements given below, it is not applicable to the content of our submitted article, financial resources, relational benefits, ethical approval because the financial aid in our article was not taken from anywhere, and also the data collection was collected from the main reference and the main book of clinical neurology of Aminoff with the help of a neurologist so animal and human cases have not been used. The complete data table has been uploaded in related files(optional) . Regarding the participation of the authors, the consent of the authors has been taken to include their names in the relevant article. The names of the authors along with their information are listed aboved and the authors have consented to use their names in this research.

8-1-Ethical Approval
According to the above explanation, the data collection was collected from the main reference and the main book of clinical neurology of Aminoff with the help of a neurologist so animal and human cases have not been used therefore, this instruction is not applicable.

8-2-Authors'contributions
Study concept and design:Z.SH., SH.GH.,GH.M.,;Acquisition of data Z.SH.;Analysis and interpretation of data:GH.M.,Z.SH.,; Drafting of the manuscript:Z.SH.,;Critical revision of the manuscript for important intellectual content:SH.GH.;Statistical analysis:Z.SH.,;Administrative, technical, and material support:SH.GH., GH.M.;Study supervision:SH.GH

8-3-Funding
the financial aid in our article was not taken from anywhere, so this instruction is not applicable.

8-4- Availability of data and materials
Data sets generated or analyzed during the current study are available from the corresponding author upon request.

8-5- Competing interests
This research has no conflict of interest.

References

1. Delcea C, Bululoi AS, Gyorgy M, Rad D. Psychological Distress Prediction Based on Maladaptive Cognitive Schemas and Anxiety with Random Forest Regression Algorithm. Pharmacophore. 2023;14(5):62-9.
2. Delcea C, Rad D, Gyorgy M, Runcan R, Breaz A, Gavrilă-Ardelean M, et al. A Network Analysis Approach to Romanian Resilience - Coping Mechanisms in the Covid-19 Era. Pharmacophore. 2023;14(4):57-63.
3. Delcea C, Bululoi AS, Gyorgy M, Siserman CV. Medico-Legal Approach to Incestuous Sexual Orientation in Men. Arch Pharm Pract. 2023;14(4):69-74.
4. Ghorbani R, Ghousi R. Predictive data mining approaches in medical diagnosis: A review of some diseases prediction. International Journal of Data and Network Science. 2019,vol.3(2), pp.47-70.
5. zacharski.ron. A prorammers Gui to Data Mining: the ancient Art of the Numeriati. 2021,pp.395.
6. Falk K. Practical Recommender Systems. 2019,pp.432.
7. Michael, J., Aminoff Md DSc, F., & Andrew Josephson, S. Aminoff's Neurology and General Medicine (6 ed.), Academic Press.March 8, 2021.