Chapter Four

Research Methodology

General Stochastic Proximity Embedding

(GSPE)

Chapter Four

General Stochastic Proximity Embedding (GSPE)

4.1 Introduction:

The goal of utilizing data reduction is to move from large-dimensional data sets to small-dimensional space while retaining the important original information in the process. The most straightforward way is to use linear algorithms; however the complexity of current datasets renders conventional approaches ineffective.

Using DR in the display of data sets may result in the occurrence of two types of mistakes: continuity errors and false neighbourhood errors, both of which are undesirable. The data points in the immediate vicinity of the original area may be projected farther out within projected area if there is a problem with continuity. False neighbourhood errors, on the other, cause the further distant data points to be projected in the original area near proximity to the projected area.

We will develop a new approach to investigate the greatest visualization and how it can be used to overcome the problems associated with (DR). Our proposed method, named General Stochastic Proximity Embedding (GSPE), will be used to investigate the best visualization of high big data[1].

4.2 General Stochastic Proximity Embedding (GSPE):

Because of the large dimensionality of their points, high big data sets have a complex structure. Although high big data sets are large and have a lot of dimensions, visualization is the best approach to display and understand them quickly.

In this thesis a new dimension reduction method is suggested which is known as “General Stochastic Proximity Embedding” (GSPE), it tries to visualize any big data sets as accurately as possible. Our suggested method (GSPE) is a novel non-linear dimensionality reduction (DR) technique for visualizing various data sets, such as hyperspectral pictures and reduces the high dimension of big data sets which preserves the neighbourhood relation between points in the original space. When compared to other well-known dimension reduction approaches, the flexibility of the GSPS process increases the likelihood of successful visualization. The amount of data lost is kept to a minimum, and the original data sets' structure is preserved. As a result, if we want a decent result, we should utilize this method first. GSPE addresses the DR's major flaws by removing the bogus neighbors points while maintaining the neighbourhood connection to the real neighbour points that are located within the locality. The visual representation of our proposed strategy shows the true, useful, and relevant points, allowing the image's objects to be readily identified.

The GSPE technique is a generic a method that is being used to show datasets in a better quality. Our method's main idea is to use projected area as a primary factor in directing projecting by identifying local neighbourhoods rather than a world society.

In addition, rather than existing metrics that provide a single measurement value, to assess the reliability of each pixel in the screen, a brand-new point-by-point statistic has been added.. Our measurement will assist the user in finding the untrustworthy places in the representation, thus identifying the pixels whose points are not reflective of the connections between the original dimension space's matching points.

Furthermore, the tests showed that the suggested approach outperforms conventional trustworthy methods since the GSPE avoids false neighbourhood errors from occurring in the findings. Furthermore, the findings demonstrated that the GSPE is superior to the SPE and FSPE in that the error projection is reduced by concentrating rather than the original space, on a low-dimensional space It may be impossible to do so without losing information. In general, DR tries to minimize as shown in the equation below:

$$\text{Stress}=\sqrt[2]{\sum_{i,j}^{n}(rij-dij)2} \quad \text{----------------------------} \; (3.1)$$

In our proposed method, both the SPE algorithm and the FSPE algorithm will be combined, beginning with a random starting configuration, the GSPE The embed is refined repeatedly by the method by randomly choosing two points i and j and changing their coordinates in the same way that stochastic gradient works, as seen below:

$$\boldsymbol{y_i \leftarrow y_i + \lambda(t)T\left(d_{ij}\right)\frac{r_{ij} - d_{ij}}{d_{ij} + \epsilon}\left(y_{i_{y_j}}\right) - - - - - - - - - - (3.2)}$$

$$\boldsymbol{y_i \leftarrow y_j + \lambda(t)T\left(d_{ij}\right)\frac{r_{ij} - d_{ij}}{d_{ij} + \epsilon}\left(y_j - y_i\right) - - - - - - - - -(3.3)}$$

$$\boldsymbol{T\left(d_{ij}\right) = \begin{cases} 1 \;\; if\; rij \neq dij \\ 0 \;\; otherwise \end{cases} \qquad - - - - - - - - - - - -(3.4)}$$

R$_{ij}$ represent the Euclidean distance in original space of SPE algorithm, and d$_{ij}$ represent the Euclidean distance in projection space of FSPE.

ʎ(t) is the rate of learning that slows down with time t,Є is a small number that is used to prevent dividing by 0.T (d$_{ij}$) is a function of distance d$_{ij}$ in the projection space.

And through the diagram 3.1 below, the work of the GSPE algorithm can be visualized.

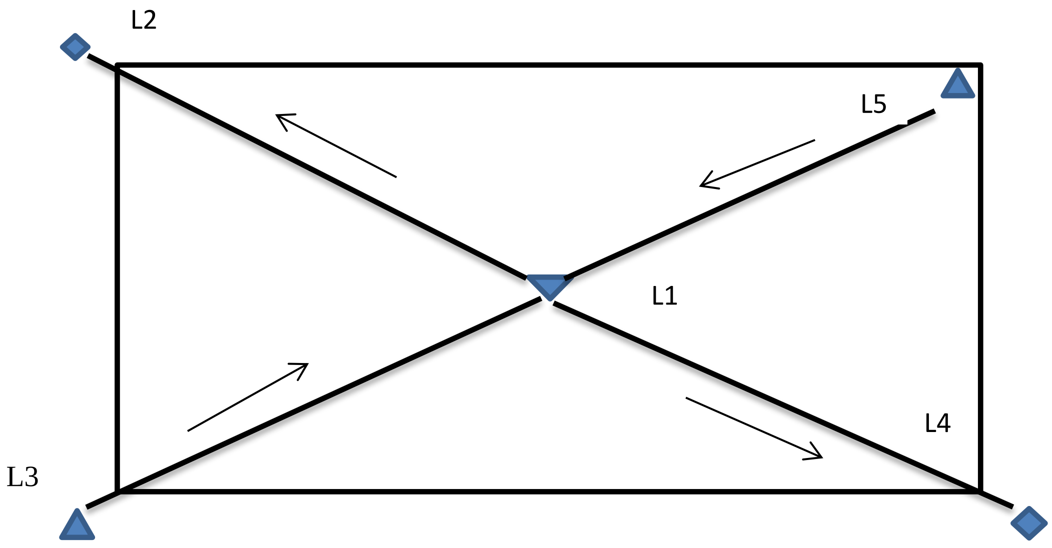


FIG.3.1: This diagram shows how the GSPE algorithm works.

Through the diagram, we notice that the point L1 maintains the distance with the true neighbours, as is the case with L3 and L5, and tries to push the pseudo-neighbours away, as in L2 and L4.

4.3 Applying GSPE to visualize Covid-19:-

The database that was worked on was stored in an excel file in the form of a table, this table consists of 16 rows and 63 columns, the rows represent the viruses that infect the respiratory system (16 rows) (the first row is influenza and the last row represents the Covid-19 virus), and the columns represented all the symptoms of these viruses.

In order to visualize the covid-19, we will apply the GSPE on the Covid-19 datasets. The best way to find the relation between all viruses, the dataset will project gradually. The step-by-step visualization is necessary to discover the unseen relationship. Thus, the method is start by the following algorithm:-

1. D=3.
2. From the dataset table, the first three columns are selected.
3. Applying GSPE on the selected dimensionality to reduce to 2D.
4. The result of step 2 illustrates the relation among viruses.
5. D=d+1.
6. If d < 63, go to step 2.

Figure 4.2 shows the algorithm to visualize Covid-19 datasets. The result of each projection, by applying the algorithm, is projected in 2D space. The results explain that the nearest viruses to Covid-19 are closed viruses to it.

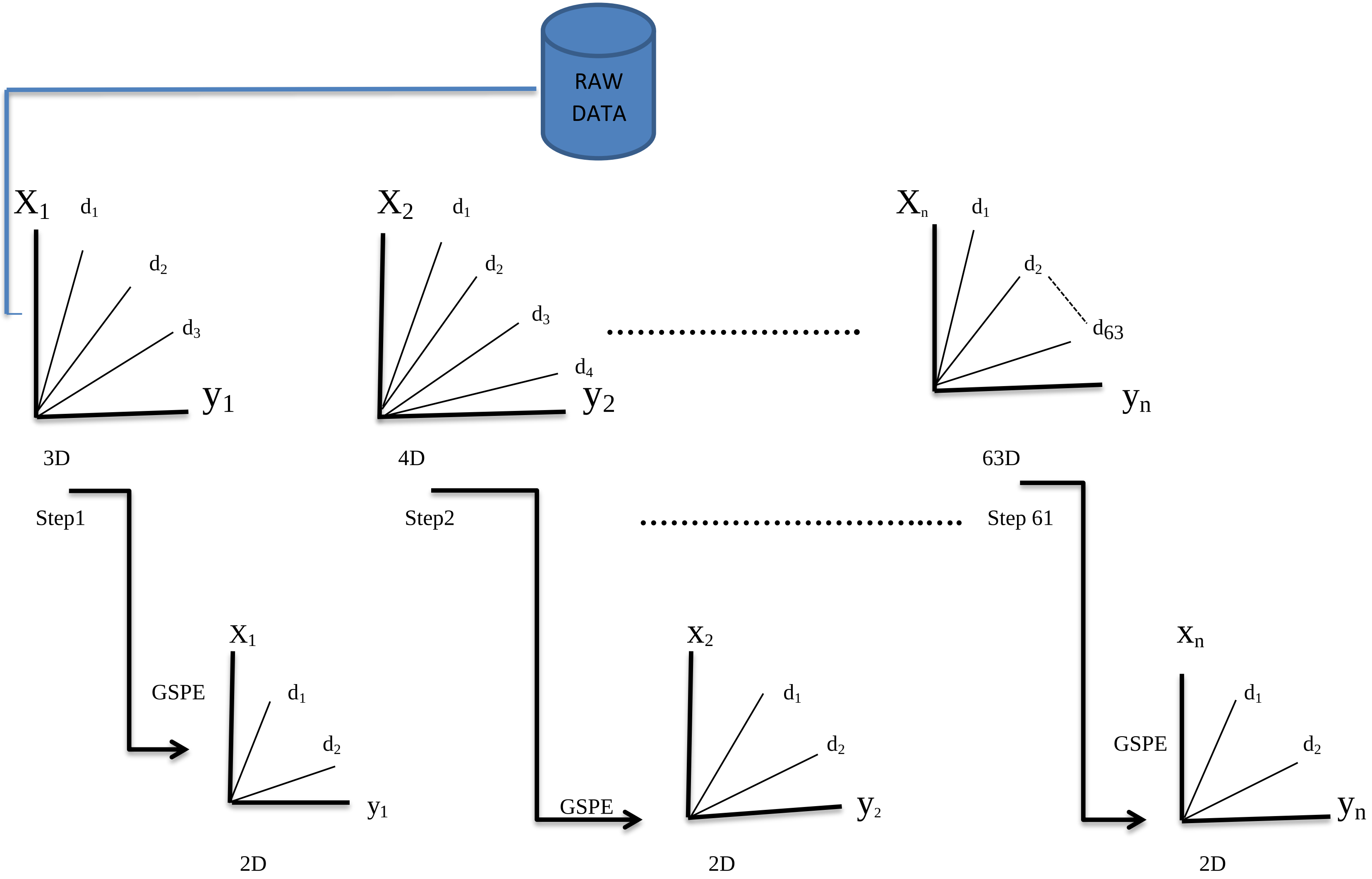
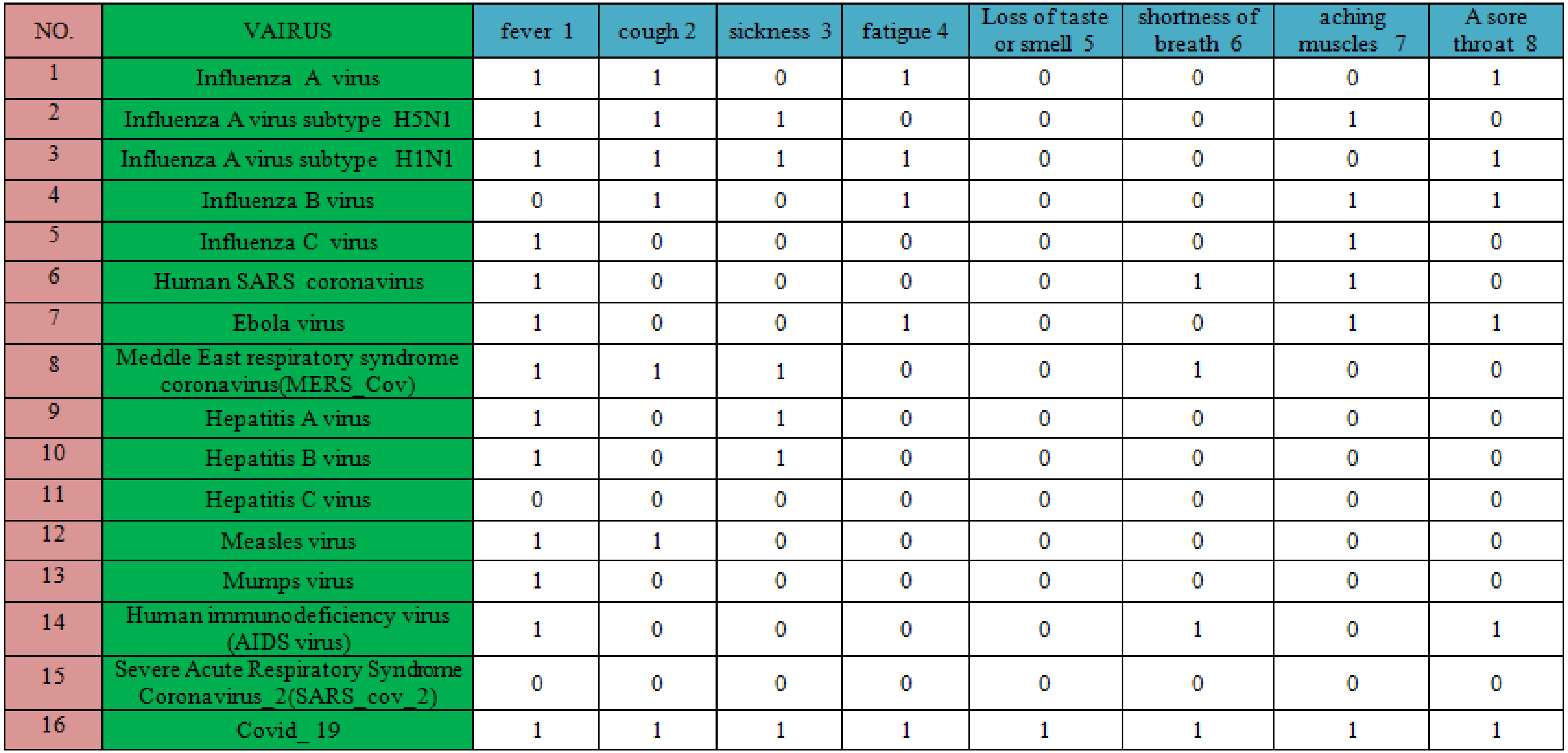


FIG.3.2: Diagram showing how the GSPE algorithm works.

The process, through the above scheme, is that we take three dimensions, then four, then five until we reach the 63 dimension, and the application of the GPSE algorithm will reduce the dimensions into two dimensions in order to reveal the relationship and amount of similarity between the data.

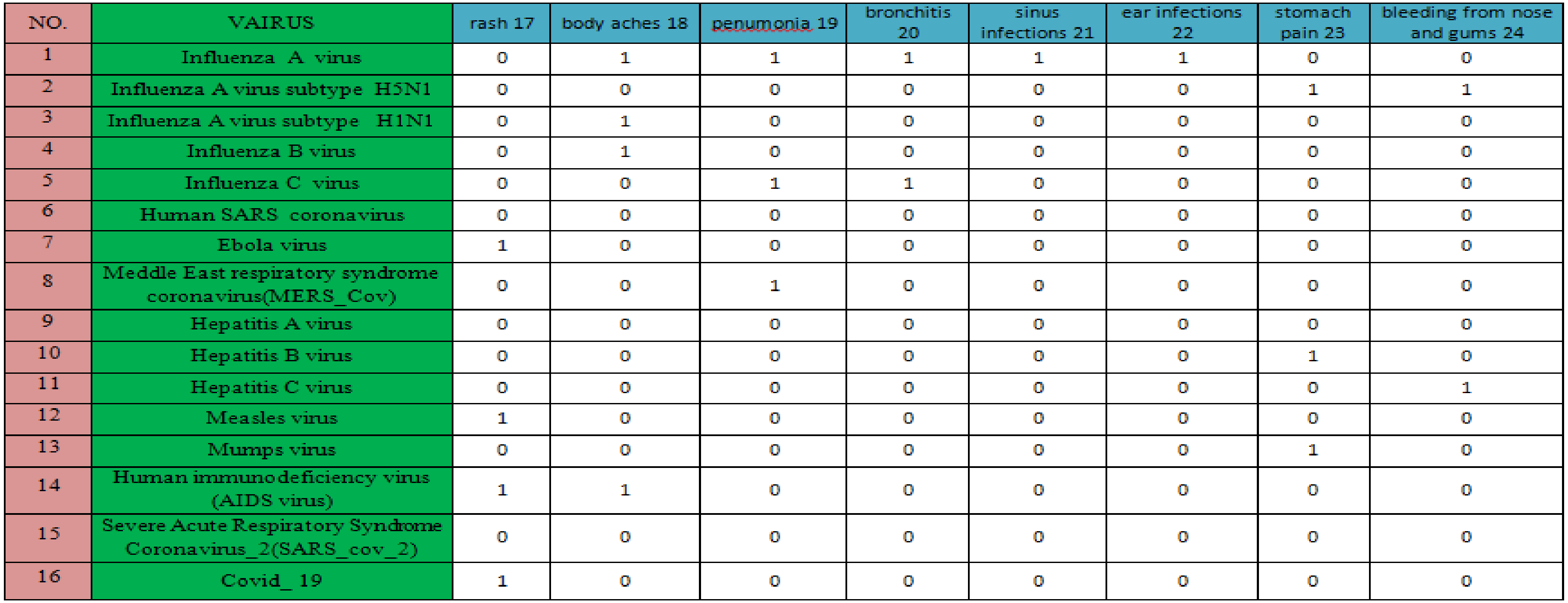
The datasets of all viruses are explained in table 4.1. In this table, we can say, there are 16 viruses, and each virus has 63 states. The Covid-19 virus in the bottom of table 4.1, where the other viruses in the above of it.

Table 3.1 COVID 63-symptoms data

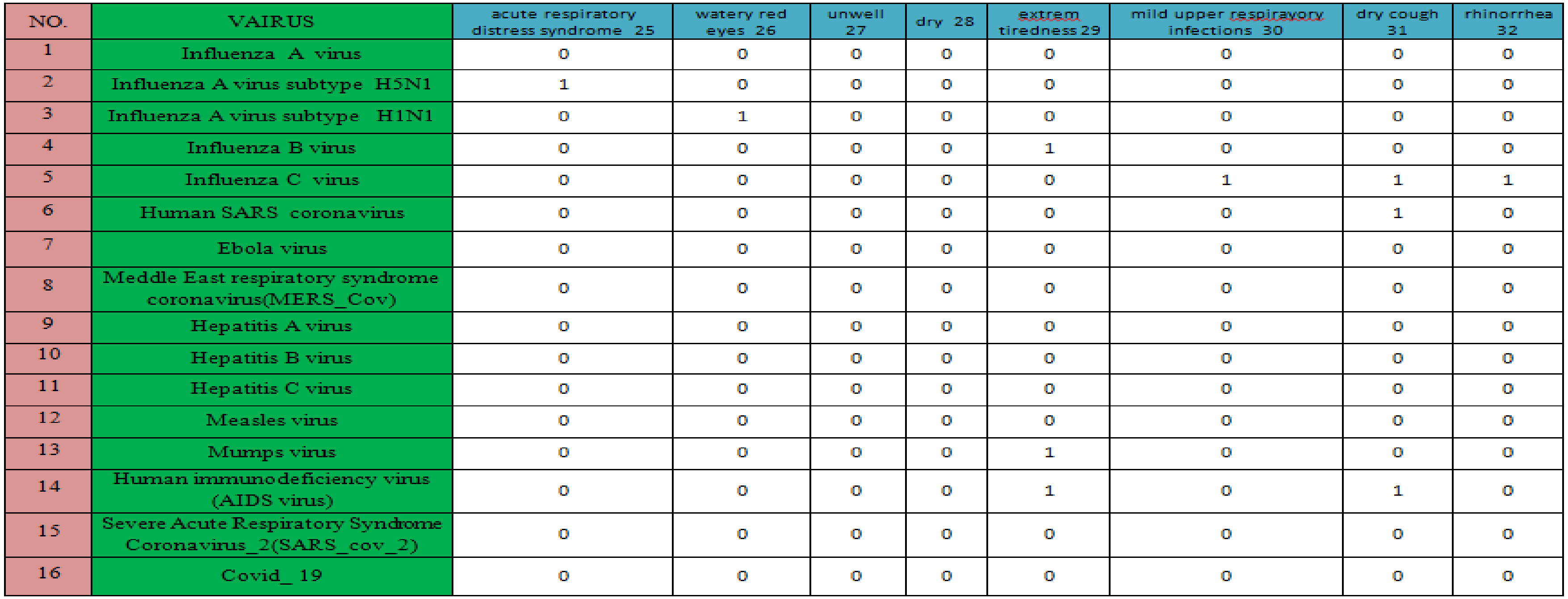


NO.	VAIRUS	fever 1	cough 2	sickness 3	fatigue 4	Loss of taste or smell 5	shortness of breath 6	aching muscles 7	A sore throat 8
1	Influenza A virus	1	1	0	1	0	0	0	1
2	Influenza A virus subtype H5N1	1	1	1	0	0	0	1	0
3	Influenza A virus subtype H1N1	1	1	1	1	0	0	0	1
4	Influenza B virus	0	1	0	1	0	0	1	1
5	Influenza C virus	1	0	0	0	0	0	1	0
6	Human SARS coronavirus	1	0	0	0	0	1	1	0
7	Ebola virus	1	0	0	1	0	0	1	1
8	Meddle East respiratory syndrome coronavirus(MERS_Cov)	1	1	1	0	0	1	0	0
9	Hepatitis A virus	1	0	1	0	0	0	0	0
10	Hepatitis B virus	1	0	1	0	0	0	0	0
11	Hepatitis C virus	0	0	0	0	0	0	0	0
12	Measles virus	1	1	0	0	0	0	0	0
13	Mumps virus	1	0	0	0	0	0	0	0
14	Human immunodeficiency virus (AIDS virus)	1	0	0	0	0	1	0	1
15	Severe Acute Respiratory Syndrome Coronavirus_2(SARS_cov_2)	0	0	0	0	0	0	0	0
16	Covid_ 19	1	1	1	1	1	1	1	1

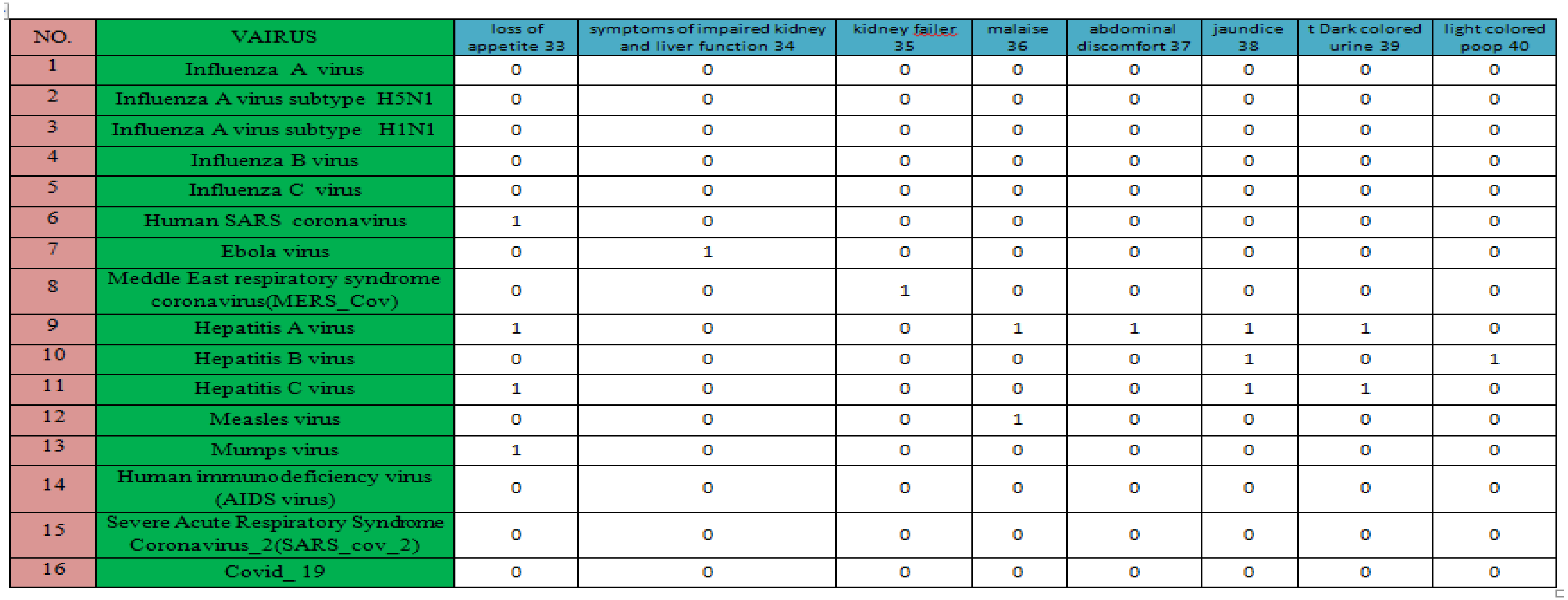
NO.	VAIRUS	chills 9	runny or stuffy nose 10	headache 11	chest pain 12	conjunctivitis 13	nasal congestion 14	diarrhea 15	vomiting 16
1	Influenza A virus	1	0	1	0	0	1	1	1
2	Influenza A virus subtype H5N1	0	0	1	1	1	0	1	0
3	Influenza A virus subtype H1N1	1	1	1	0	0	0	1	1
4	Influenza B virus	0	1	1	0	0	0	0	0
5	Influenza C virus	0	0	1	0	0	0	0	0
6	Human SARS coronavirus	1	0	1	0	0	0	1	0
7	Ebola virus	0	0	1	0	0	0	0	1
8	Meddle East respiratory syndrome coronavirus(MERS_Cov)	0	0	0	0	0	0	0	1
9	Hepatitis A virus	0	0	0	0	0	0	1	0
10	Hepatitis B virus	0	0	0	0	0	0	0	1
11	Hepatitis C virus	0	0	0	0	0	0	0	0
12	Measles virus	0	0	0	0	1	0	0	0
13	Mumps virus	0	0	1	0	0	0	0	0
14	Human immunodeficiency virus (AIDS virus)	0	0	1	0	0	0	1	0
15	Severe Acute Respiratory Syndrome Coronavirus_2(SARS_cov_2)	0	0	0	0	0	0	0	0
16	Covid_ 19	1	1	1	1	1	1	1	1



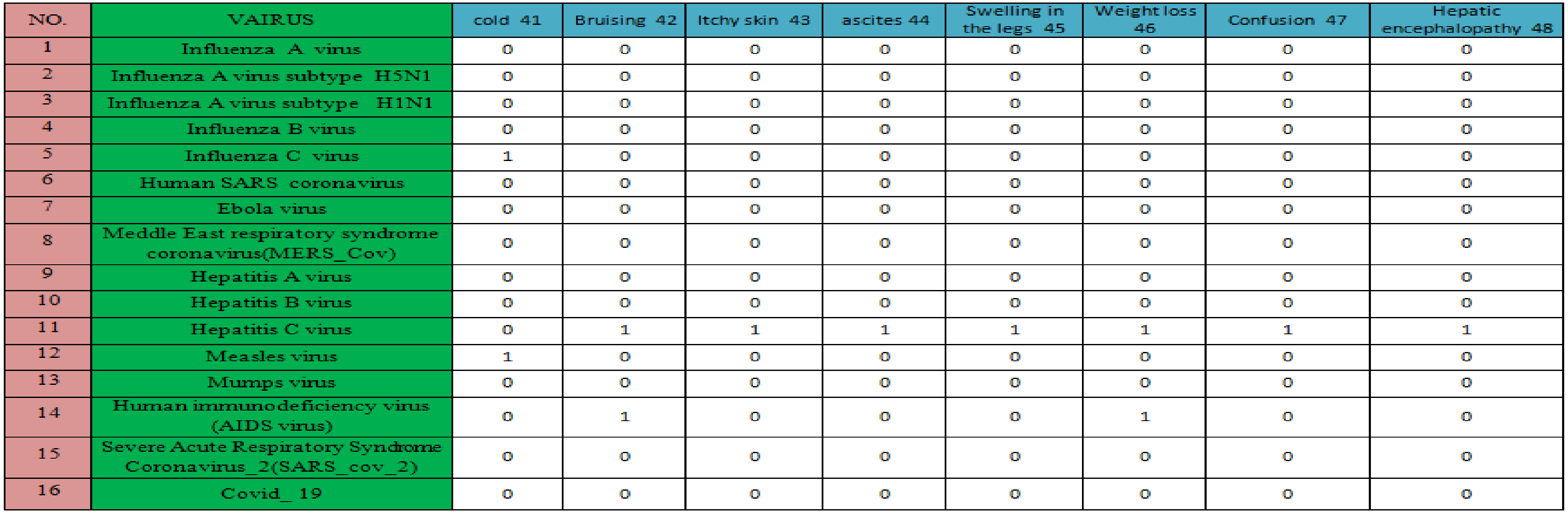
NO.	VAIRUS	rash 17	body aches 18	penumonia 19	bronchitis 20	sinus infections 21	ear infections 22	stomach pain 23	bleeding from nose and gums 24
1	Influenza A virus	0	1	1	1	1	1	0	0
2	Influenza A virus subtype H5N1	0	0	0	0	0	0	1	1
3	Influenza A virus subtype H1N1	0	1	0	0	0	0	0	0
4	Influenza B virus	0	1	0	0	0	0	0	0
5	Influenza C virus	0	0	1	1	0	0	0	0
6	Human SARS coronavirus	0	0	0	0	0	0	0	0
7	Ebola virus	1	0	0	0	0	0	0	0
8	Meddle East respiratory syndrome coronavirus(MERS_Cov)	0	0	1	0	0	0	0	0
9	Hepatitis A virus	0	0	0	0	0	0	0	0
10	Hepatitis B virus	0	0	0	0	0	0	1	0
11	Hepatitis C virus	0	0	0	0	0	0	0	1
12	Measles virus	1	0	0	0	0	0	0	0
13	Mumps virus	0	0	0	0	0	0	1	0
14	Human immunodeficiency virus (AIDS virus)	1	1	0	0	0	0	0	0
15	Severe Acute Respiratory Syndrome Coronavirus_2(SARS_cov_2)	0	0	0	0	0	0	0	0
16	Covid_ 19	1	0	0	0	0	0	0	0



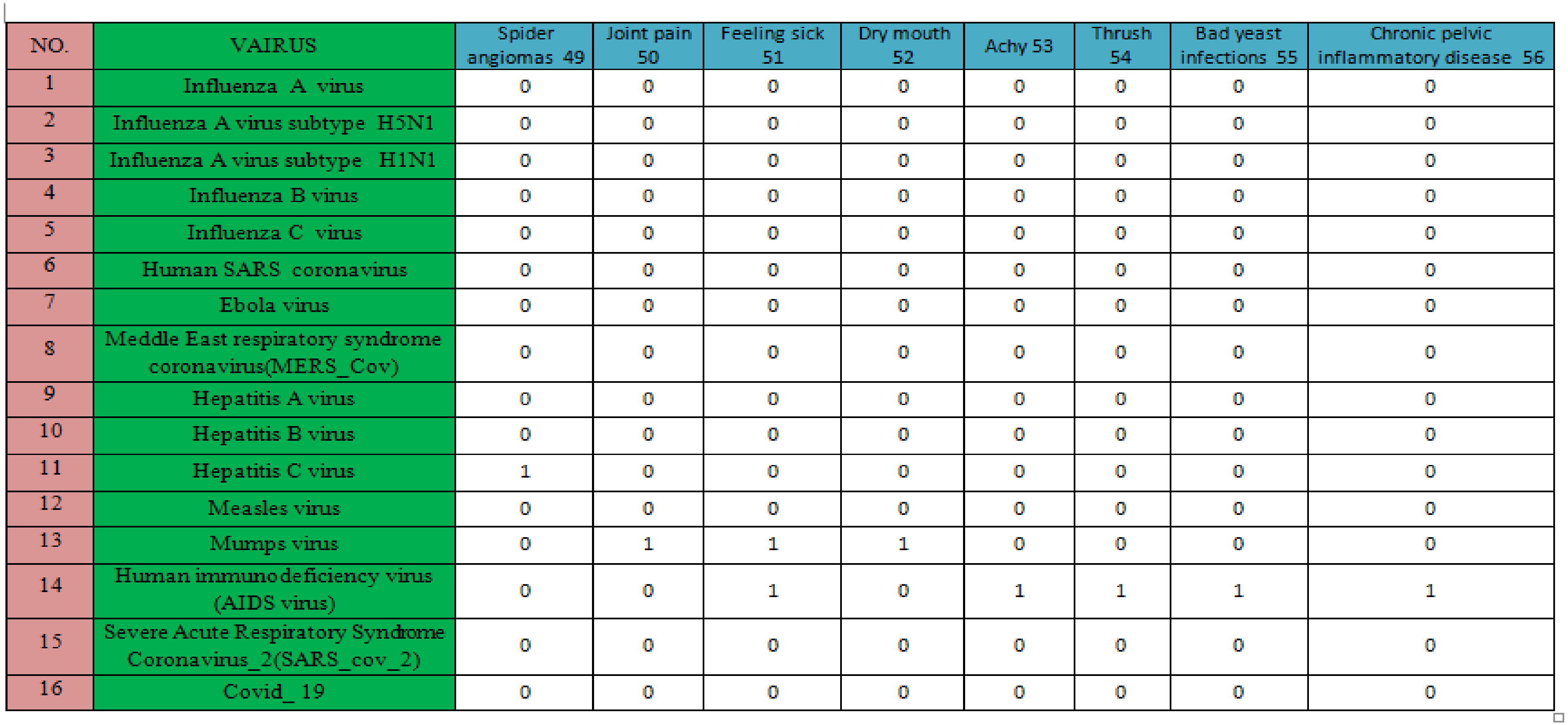
NO.	VAIRUS	acute respiratory distress syndrome 25	watery red eyes 26	unwell 27	dry 28	extrem tiredness 29	mild upper respiravory infections 30	dry cough 31	rhinorrhea 32
1	Influenza A virus	0	0	0	0	0	0	0	0
2	Influenza A virus subtype H5N1	1	0	0	0	0	0	0	0
3	Influenza A virus subtype H1N1	0	1	0	0	0	0	0	0
4	Influenza B virus	0	0	0	0	1	0	0	0
5	Influenza C virus	0	0	0	0	0	1	1	1
6	Human SARS coronavirus	0	0	0	0	0	0	1	0
7	Ebola virus	0	0	0	0	0	0	0	0
8	Meddle East respiratory syndrome coronavirus(MERS_Cov)	0	0	0	0	0	0	0	0
9	Hepatitis A virus	0	0	0	0	0	0	0	0
10	Hepatitis B virus	0	0	0	0	0	0	0	0
11	Hepatitis C virus	0	0	0	0	0	0	0	0
12	Measles virus	0	0	0	0	0	0	0	0
13	Mumps virus	0	0	0	0	1	0	0	0
14	Human immunodeficiency virus (AIDS virus)	0	0	0	0	1	0	1	0
15	Severe Acute Respiratory Syndrome Coronavirus_2(SARS_cov_2)	0	0	0	0	0	0	0	0
16	Covid_ 19	0	0	0	0	0	0	0	0



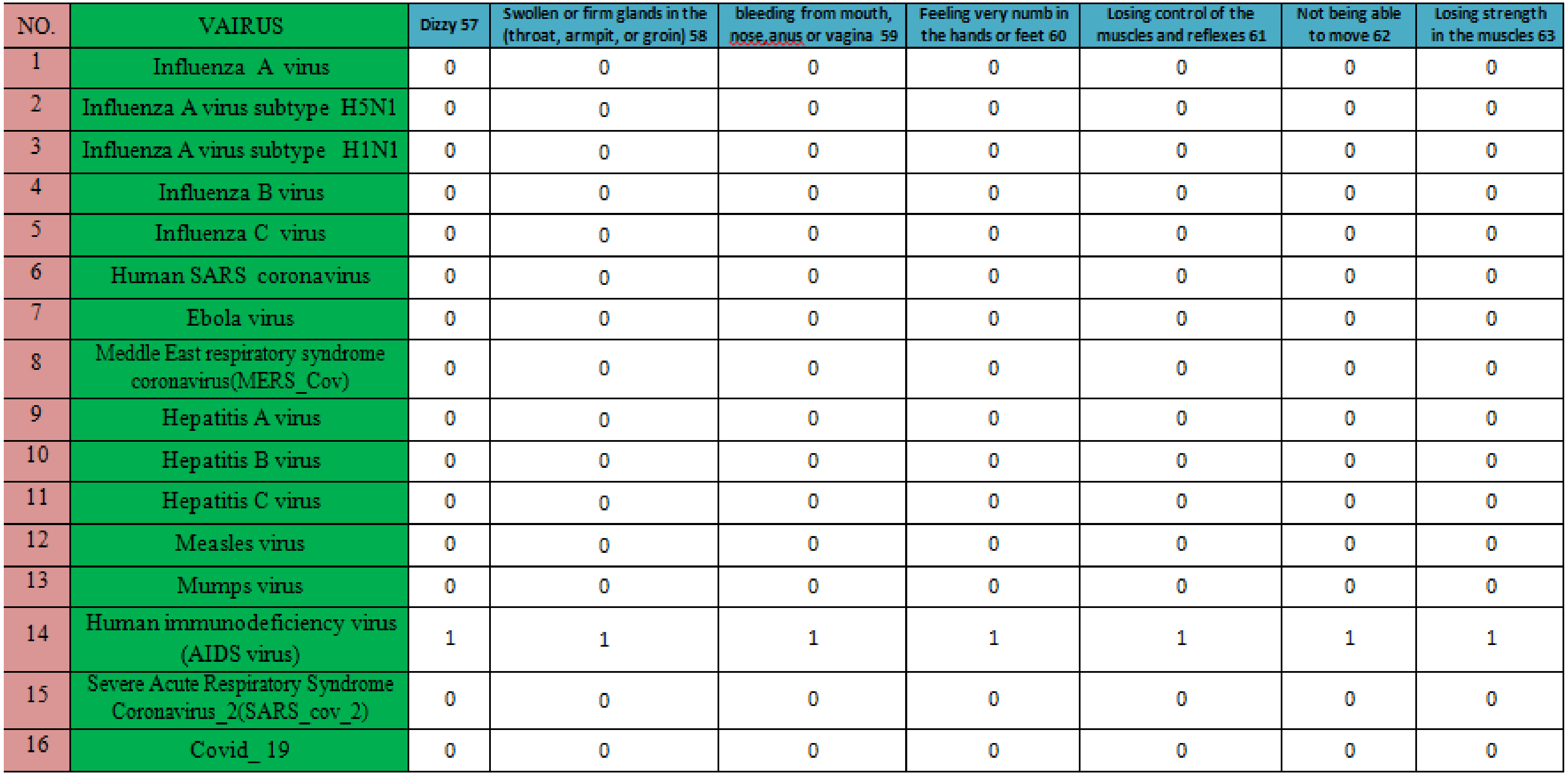
NO.	VAIRUS	loss of appetite 33	symptoms of impaired kidney and liver function 34	kidney failer 35	malaise 36	abdominal discomfort 37	jaundice 38	t Dark colored urine 39	light colored poop 40
1	Influenza A virus	0	0	0	0	0	0	0	0
2	Influenza A virus subtype H5N1	0	0	0	0	0	0	0	0
3	Influenza A virus subtype H1N1	0	0	0	0	0	0	0	0
4	Influenza B virus	0	0	0	0	0	0	0	0
5	Influenza C virus	0	0	0	0	0	0	0	0
6	Human SARS coronavirus	1	0	0	0	0	0	0	0
7	Ebola virus	0	1	0	0	0	0	0	0
8	Meddle East respiratory syndrome coronavirus(MERS_Cov)	0	0	1	0	0	0	0	0
9	Hepatitis A virus	1	0	0	1	1	1	1	0
10	Hepatitis B virus	0	0	0	0	0	1	0	1
11	Hepatitis C virus	1	0	0	0	0	1	1	0
12	Measles virus	0	0	0	1	0	0	0	0
13	Mumps virus	1	0	0	0	0	0	0	0
14	Human immunodeficiency virus (AIDS virus)	0	0	0	0	0	0	0	0
15	Severe Acute Respiratory Syndrome Coronavirus_2(SARS_cov_2)	0	0	0	0	0	0	0	0
16	Covid_ 19	0	0	0	0	0	0	0	0



NO.	VAIRUS	cold 41	Bruising 42	Itchy skin 43	ascites 44	Swelling in the legs 45	Weight loss 46	Confusion 47	Hepatic encephalopathy 48
1	Influenza A virus	0	0	0	0	0	0	0	0
2	Influenza A virus subtype H5N1	0	0	0	0	0	0	0	0
3	Influenza A virus subtype H1N1	0	0	0	0	0	0	0	0
4	Influenza B virus	0	0	0	0	0	0	0	0
5	Influenza C virus	1	0	0	0	0	0	0	0
6	Human SARS coronavirus	0	0	0	0	0	0	0	0
7	Ebola virus	0	0	0	0	0	0	0	0
8	Meddle East respiratory syndrome coronavirus(MERS_Cov)	0	0	0	0	0	0	0	0
9	Hepatitis A virus	0	0	0	0	0	0	0	0
10	Hepatitis B virus	0	0	0	0	0	0	0	0
11	Hepatitis C virus	0	1	1	1	1	1	1	1
12	Measles virus	1	0	0	0	0	0	0	0
13	Mumps virus	0	0	0	0	0	0	0	0
14	Human immunodeficiency virus (AIDS virus)	0	1	0	0	0	1	0	0
15	Severe Acute Respiratory Syndrome Coronavirus_2(SARS_cov_2)	0	0	0	0	0	0	0	0
16	Covid_ 19	0	0	0	0	0	0	0	0



NO.	VAIRUS	Spider angiomas 49	Joint pain 50	Feeling sick 51	Dry mouth 52	Achy 53	Thrush 54	Bad yeast infections 55	Chronic pelvic inflammatory disease 56
1	Influenza A virus	0	0	0	0	0	0	0	0
2	Influenza A virus subtype H5N1	0	0	0	0	0	0	0	0
3	Influenza A virus subtype H1N1	0	0	0	0	0	0	0	0
4	Influenza B virus	0	0	0	0	0	0	0	0
5	Influenza C virus	0	0	0	0	0	0	0	0
6	Human SARS coronavirus	0	0	0	0	0	0	0	0
7	Ebola virus	0	0	0	0	0	0	0	0
8	Meddle East respiratory syndrome coronavirus(MERS_Cov)	0	0	0	0	0	0	0	0
9	Hepatitis A virus	0	0	0	0	0	0	0	0
10	Hepatitis B virus	0	0	0	0	0	0	0	0
11	Hepatitis C virus	1	0	0	0	0	0	0	0
12	Measles virus	0	0	0	0	0	0	0	0
13	Mumps virus	0	1	1	1	0	0	0	0
14	Human immunodeficiency virus (AIDS virus)	0	0	1	0	1	1	1	1
15	Severe Acute Respiratory Syndrome Coronavirus_2(SARS_cov_2)	0	0	0	0	0	0	0	0
16	Covid_ 19	0	0	0	0	0	0	0	0



NO.	VAIRUS	Dizzy 57	Swollen or firm glands in the (throat, armpit, or groin) 58	bleeding from mouth, nose,anus or vagina 59	Feeling very numb in the hands or feet 60	Losing control of the muscles and reflexes 61	Not being able to move 62	Losing strength in the muscles 63
1	Influenza A virus	0	0	0	0	0	0	0
2	Influenza A virus subtype H5N1	0	0	0	0	0	0	0
3	Influenza A virus subtype H1N1	0	0	0	0	0	0	0
4	Influenza B virus	0	0	0	0	0	0	0
5	Influenza C virus	0	0	0	0	0	0	0
6	Human SARS coronavirus	0	0	0	0	0	0	0
7	Ebola virus	0	0	0	0	0	0	0
8	Meddle East respiratory syndrome coronavirus(MERS_Cov)	0	0	0	0	0	0	0
9	Hepatitis A virus	0	0	0	0	0	0	0
10	Hepatitis B virus	0	0	0	0	0	0	0
11	Hepatitis C virus	0	0	0	0	0	0	0
12	Measles virus	0	0	0	0	0	0	0
13	Mumps virus	0	0	0	0	0	0	0
14	Human immunodeficiency virus (AIDS virus)	1	1	1	1	1	1	1
15	Severe Acute Respiratory Syndrome Coronavirus_2(SARS_cov_2)	0	0	0	0	0	0	0
16	Covid_ 19	0	0	0	0	0	0	0

4.4-Conclusion:

This chapter explains our suggested GSPE. This method has ability to reduce the dimensionality of the dataset into lower. GSPE tries to project the datasets into another lower space which is more clear to the user . the ability of the our suggested method (GSPE) led us to use it in visualizing the Covid-19 viruses with other viruses. In the third chapter, we prove the accuracy of GSPE in visualizing the bioinformatics datasets.