

Clustering Matrix Protein of Influenza Virus Using Deep Embedded Network

BY

APURBA KUMAR ROY
ID: 191-15-12499

MD. SHAMS UDDIN MIM
ID: 191-15-12103

AND

MD. MUSHFIQUR RAHMAN
ID: 191-15-12696

This Report Presented in Partial Fulfillment of the Requirements for the Degree of
Bachelor of Science in Computer Science and Engineering

Supervised By

Md. Ferdouse Ahmed Foysal

Lecturer

Department of Computer Science and Engineering
Faculty of Science and Information Technology
Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

January 2023

©Daffodil International University


APPROVAL

This Project/internship titled: “Clustering Matrix Protein of Influenza Virus using Deep Embedded Network”, submitted by Apurba Kumar Roy ID No:191-15-12499, Md. Shams Uddin Mim ID No: 191-15-12103 and Md. Mushfiqur Rahman ID No: 191-15-12696 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on January 29, 2023.

BOARD OF EXAMINERS

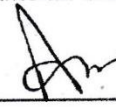
Dr. Touhid Bhuiyan
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Chairman



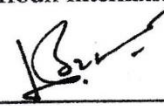
Md. Abbas Ali Khan
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Ms. Aliza Ahmed Khan
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Dr. Md. Sazzadur Rahman
Associate Professor
Institute of Information Technology
Jahangirnagar University

External Examiner

DECLARATION

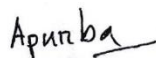
We thus declare that we completed this research project under the supervision of **Md. Ferdouse Ahmed Foysal, Lecturer, Department of CSE** Daffodil International University. Additionally, we affirm that neither this project nor any portion of it has been submitted to any institution for the award of a degree or diploma.

Supervised by:

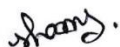


Md. Ferdouse Ahmed Foysal
Lecturer
Department of CSE
Daffodil International University

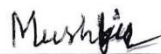
Submitted by:



Apurba Kumar Roy
ID: 191-15-12499
Department of CSE
Daffodil International University



Md. Shams Uddin Mim
ID: 191-15-12103
Department of CSE
Daffodil International University



Md. Mushfiqur Rahman
ID: 191-15-696
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

To begin with, we offer our heartfelt appreciation and gratitude to Almighty God for His divine grace, which enabled us to successfully finish the final year project/internship.

We owe a great debt of gratitude and wish to express our appreciation to **Md. Ferdouse Ahmed Foysal, Lecturer**, Department of CSE Daffodil International University, Dhaka. Our supervisor has extensive knowledge and a deep interest in the field of "Deep Learning & Bioinformatics" which helped us carry out this research. The excellent counsel, endless patience, intellectual direction, and constant encouragement with continual supervision of our supervisor have helped us in all steps of this project. Our supervisors have also been constructive critics of our work by correcting several substandard attempts of ours.

We also wish to express our heartfelt appreciation to **Dr. Touhid Bhuiyan**, Professor and Head, Department of CSE, for his generous support and assistance in completing our project. We are also grateful to other faculty members and the staff of the CSE department of Daffodil International University.

Additionally, we appreciate the encouragement and inspiration provided by all of our well-wishers, friends, family, and elders. This research is the result of a great deal of effort and the encouragement and cooperation of all those people.

And at last, we must acknowledge the unwavering support and constant encouragement of our parents. For this, we owe a great debt of gratitude to them.

ABSTRACT

Influenza A virus is a type of virus that can cause respiratory illness in humans and animals. They are classified into subtypes based on the combination of two proteins on the surface of the virus: hemagglutinin (HA) and neuraminidase (NA). There are 18 different HA subtypes and 11 different NA subtypes, and many different combinations of these subtypes are possible. One way to study these viruses is to use clustering techniques to group them based on certain features. A deep embedded network is used to learn a low-dimensional representation of the virus sequences, which is used as input to a clustering algorithm called K-means. To perform this analysis, we collected a dataset of influenza A viruses. Then The deep embedded network is used as a learning representation. K-means clustering is applied to the learned representation to cluster the virus sequences into clusters based on their similarity. The number of clusters can be determined using techniques such as the elbow method or the silhouette score. Using deep embedded networks and K-means clustering can provide insights into the relationships between different influenza A viruses and help researchers understand patterns and trends in the data. It can also be useful for tracking the evolution and spread of these viruses over time.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of Examiners.....	II
Declaration	III
Acknowledgment.....	IV
Abstract.....	V
CHAPTER	PAGE
CHAPTER 1: INTRODUCTION	1-4
1.1 Introduction.....	1-2
1.2 Motivation.....	2-3
1.3 Research Questions.....	3
1.4 Expected Outcome.....	3
1.5 Project Management and Finance.....	3-4
1.6 Layout of the Report.....	4
CHAPTER 2: BACKGROUND STUDY	5-8
2.1 Preliminaries and Terminologies.....	5
2.2 Related Works	5-6
2.3 Comparative Analysis & Summary	6
2.4 Scope of The Problem.....	7
2.5 Challenges.....	7-8
CHAPTER 3: RESEARCH METHODOLOGY	9-19
3.1 Introduction.....	9
3.2 Research Subject and Instrumentation.....	9
3.3 Workflow	9-11
3.4 Data Collection Procedure.....	11-12
3.5 Data Processing	12
3.6 Proposed Methodology	13-19

CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION	20-25
4.1 Experimental Setup	20
4.2 Performance Analysis	20-22
4.3 Result Discussion	23
4.4 Elbow Method.....	23-25
CHAPTER 5: IMPACT ON SOCIETY, ENVIRONMENT, AND SUSTAINABILITY	26-27
5.1 Impact on Society.....	26
5.2 Impact on Environment.....	26
5.3 Ethical Aspects.....	27
5.4 Sustainability Plan.....	27
CHAPTER 6: SUMMARY, CONCLUSION, RECOMMENDATION, IMPLICATION FOR FUTURE RESEARCH	28
6.1 Introduction	28
6.2 Conclusion.....	28
6.3 Future Works.....	28
REFERENCES	29
APPENDIX	30
PLAGIARISM REPORT	31-35

LIST OF FIGURES	PAGE NO
Fig 3.1: Workflow Diagram	10
Fig 3.2: Source of Data	11
Fig 3.3: Sample data from the dataset	12
Fig 3.4: Deep Embedded Network	13
Fig 3.5: Feature Vector Formation	17
Fig 3.6: Principal Components	18
Fig 3.7: Workflow of K-means clustering	19
Fig 4.1: Converted CSV data from FASTA	20
Fig 4.2: Split Properties	21
Fig 4.3: Vector Representation	21
Fig 4.4: Variance Correlation	22
Fig 4.5: Clustering with Centroid	22
Fig 4.6: Clustering Result	23
Fig 4.7: Elbow Plotting	24
Fig: 4.8 Silhouette Analysis	25

CHAPTER 1

Introduction

1.1 Introduction

The Orthomyxoviridae family of RNA viruses includes the influenza virus. Influenza A, Influenza B, Influenza C, and Influenza D, are its four primary subtypes. The most frequent of them is the influenza A-type virus. It affects many avian and mammal species, which could cause pandemics and epidemics of flu. The influenza A virus has several different serotypes and infects a variety of hosts in nature. Hemagglutinin (HA) and neuraminidase (NA), two important internal proteins, change across these serotypes. Eight single-stranded, negative-sense RNA segments try to compensate for the genomes of each influenza A virus serotype. The M1 and M2 proteins that make up the influenza A virus's membrane or matrix protein are necessary for the virus life cycle and virus replication. These influenza A virus M1 and M2 proteins have experienced evolutionary modifications over time. As a result, the influenza A virus variety has become a serious problem, making it important to find these novel genomes. One of the best clustering methods has been shown to use deep learning techniques. The analysis of relationships between protein structure and activity and genome grouping is made easier by deep learning technologies. Different types of approaches such as Clustering Recurrent Neural Networks (CRNN), convolutional variational autoencoder, novel clustering algorithms (CLUGEN), Korhonen unsupervised learning algorithms, adaptive neural network-based clustering method, hierarchical and division techniques to understand sequence-to-structure relationships, cluster protein sequences, identify similarities, predict protein genome function, etc. They solved various types of problems raised while sequencing proteins such as the prediction of protein sequences could not get a good result because of working speed, different types of protein structure, and datasets. Therefore, we thought of constructing a deep-learning approach to identify the new gene sequence by clustering. We used the nucleotide sequences of every serotype of the influenza A virus that was taken from the Influenza Virus Resource as the dataset for our unsupervised learning approach. We collected nucleotide sequences for all proteins from all H5N1 influenza A serotypes between 2000 and 2020. To cluster new sequences, we applied deep learning techniques called LSTM and CNN. By using the pre-processed and selected training dataset, we built our proposed model. After being put to the test against the training dataset, both deep learning techniques fully achieved the predetermined goals. In contrast, little research has been done on or major use of deep learning (DL)-based

representation and feature learning for clustering. Deep neural networks can be an efficient way to convert mappings from a high-dimensional data space into a lower-dimensional feature space, leading to improved clustering outcomes. This is because the quality of clustering depends not only on the distribution of data points but also on the learned representation.

1.2 Motivation

The influenza virus is a member of the RNA virus family, Orthomyxoviridae. Its four main subtypes are influenza A, influenza B, influenza C, and influenza D. The influenza A-type virus is the most prevalent of all. It affects a wide variety of bird and animal species, which could result in flu pandemics and epidemics. In nature, there is numerous influenza A virus serotypes that infect a range of hosts. The virus is continually changing, making it challenging for the human immune system to identify and effectively combat it. Analyzing the protein sequences of the viruses' many strains is one method of monitoring the virus' evolution. Protein sequence clustering is the process of putting related protein sequences together. This can help figure out connections between various virus strains and for comprehending how the virus changes over time. Multiple sequence alignment and phylogenetic tree construction techniques have traditionally been used to cluster protein sequences. These techniques, nevertheless, have drawbacks and can be time-consuming. Deep learning methods have recently been used to cluster protein sequences, with encouraging outcomes. Deep embedded networks, which employ neural networks to learn a low-dimensional representation of the protein sequences, are one such approach. The sequences can then be clustered using this representation and conventional clustering algorithms like K-means. When compared to conventional protein sequence clustering techniques, deep embedded networks offer several benefits. They can efficiently manage enormous volumes of data and understand intricate correlations between protein sequences. Furthermore, they can spot clusters that conventional approaches can miss. For the protein sequence clustering of the Influenza A virus, we suggest using deep embedded networks, PCA (principal component analysis), and K-means clustering in this study. The protein sequences are initially preprocessed to make sure they are appropriate for clustering. The sequences were then vectorized using a deep embedding network that had been trained to learn a low-dimensional representation of them. To prepare the data for k-means clustering, PCA was then used to process it. To cluster the sequences, we have finally employed K-means clustering on the learned representation. Square distance computations are used to assess the efficacy of our strategy.

Overall, the combination of deep embedded networks, PCA, and K-means clustering has the potential to be a formidable tool for the clustering of the influenza A virus protein sequence and may offer insights into the development and transmission of the virus.

1.3 Research Questions

An essential first step in starting research is formulating a precise, short, and focused study question. It provides a clear focus and purpose and outlines precisely what we want to learn. The researchers would want to present the following questions to convey their ideas and findings to reach a realistic, effective, and accurate solution to this issue.

- Can we get the influenza virus genome sequences for our deep-learning research?
- Can the raw data that was taken from the database sources be preprocessed?
- Will the field of protein gene clustering be improved by this work?
- Will this project advance bioinformatics?
- Is this study beneficial to humanity?

1.4 Expected Outcome

We have discussed the study plan for this section, which is based on the research questions. We expect to develop a well-known model to precisely cluster gene sequences. The following research expectation outcome is what the researchers would want to put up. Such as,

- Accurate and efficient protein sequence clustering of Influenza A virus strains.
- Identification of previously unknown relationships between different sequences of the virus.
- Improved ability to track the evolution of Influenza A virus over time through analysis of protein sequence data.
- Insights into the spread and transmission of the virus through analysis of the clusters produced by the method.
- A useful tool for researchers and public health officials in monitoring and responding to outbreaks of Influenza A virus.

1.5 Project Management and Finance

We first constructed a planned architecture for our research project, which we then incrementally improved. Through careful project management, we have so far succeeded in meeting our goals.

The following illustrates the project flow:

- I. Searched extensively for studies related to our project.
- II. Selected research studies that were relevant to our area of study.
- III. Analyzed all of the research papers that were chosen to find out more about protein clustering with deep learning.
- IV. Extracted specified datasets from the database of the influenza virus resource.
- V. Prepared and processed the dataset to enable deep learning algorithms to classify it.
- VI. We put the best deep learning classifiers into practice to accomplish our goals.
- VII. Trained and put to the test the classifiers to show the accuracy of our work.
- VIII. Presented the completed research project report in written form.

1.6 Layout of the Report

- The research's introduction is covered in Chapter 1, along with the study's purpose, justification, research questions, expected outcomes, project management and funding, and overall structure.
- The Background of the research is presented in Chapter 2. It discusses the Problem Scope, Related Works, Comparative Analysis & Summary, and The Challenges.
- The theoretical analysis of the research is presented in Chapter 3. The project's workflow is shown in the first section of this chapter. The process for gathering data and processing it is then described. This chapter also demonstrates the Deep Learning Classifiers' algorithmic techniques. Finally, some implementation requirements are discussed.
- The experimental results, a discussion of the findings, and an analysis of the project's effectiveness are all included in chapter four. Experimental representations are included to make the results easier to understand.
- The research's impact on society, the environment, and some ethical issues are discussed in chapter five.
- The work's conclusion and summary are found in chapter six, which serves as the book's final chapter. The chapter's conclusion identifies some of the chapter's issues and provides some suggestions for more research based on the findings.

Chapter 2

Background Study

2.1 Preliminaries and Terminologies

The reassortment, of as well as reorganization of multiple viral strains' genomes in varied hosts, as well as reorganization of multiple viral strains' genomes in varied hosts, is what leads to clustering, which produces fresh virus strains with unique features. Understanding influenza, A virus' dynamic behavior is therefore of tremendous interest and importance. To meet this need, computational biology is one way that helps. During the earlier study, deep learning algorithms were quite successful at clustering virus genome sequences. Therefore, utilizing deep learning techniques, we suggested and created protein clustering models as computer science students.

Terminologies:

- **Protein Clustering:** During viral replication, various virus strains may reassort or shuffle their genomic sequences inside an infected host cell if there are multiple serotypes [1]. Protein clustering is a process that results in the classification of virus serotypes with new and distinctive traits.
- **Clustering Sequence:** The genomic sequence of a particular virus, in particular, a protein or nucleotide sequence, must be examined to learn and cluster virus serotypes. By grouping and identifying differences in the virus's genome, protein, and nucleotide sequences, virus serotypes can be classified.
- **Deep learning clustering** is trained by newly sequenced serotypes to be able to group them into specific serotypes. This kind of clustering is known as unsupervised deep learning.

2.2 Related Works

Many studies using deep learning methods to cluster proteins have been conducted throughout the years. They employed topological similarity metrics, base clustering algorithms, and consensus approaches in detail to cluster protein-protein interaction networks in a research paper titled "An ensemble framework for clustering protein-protein interaction networks" that was published in 2007 [1]. Yet another is a study titled "Protein Classification Using Artificial Neural Networks with Different Proteins," which was published in 2007.

In "encoding approaches," proteins were categorized using genetic algorithms and artificial neural networks (ANNs) [2]. A 2005 research effort titled "Clustering Protein Sequences with a Novel Metric Transformed from Sequence Similarity Scores and Sequence Alignments Using Neural Networks" [3] led to the development of the MCL Algorithm (enhanced multileaf collimator). In a research paper titled "Protein-Protein Interaction Network-Based Knowledge Embedding with Graph Neural Network for Single-Cell RNA to Protein Prediction" that was released in November 2020, a PIKE-R2P (Protein-Protein Interaction Network-Based Knowledge Embedding with Graph Neural Network for Single-Cell RNA to Protein Prediction) was developed. [4]. Using template-based modeling (TBM) and CEThreader models, a recent research study titled "Detecting distant-homology protein structures by aligning deep neural network-based contact maps" was released in September 2019 [5]. Its goals were to annotate the biological functions of protein molecules and to develop new compounds to regulate the functions. ART-1-based clustering on Yeast Protein-Protein Interactions was employed by the authors of "Adaptive Neural Network-Based Clustering of Yeast Protein-Protein Interactions," a study that was published in 2004 [6]. In a June 2008 article titled "Ensemble non-negative matrix factorization techniques for clustering protein-protein interactions" [7], an NMF and K-means clustering algorithm model was proposed. In a study published in April 2010 titled "Clustering of protein expression data: a benchmark of statistical and neural techniques" [8], the k-means clustering methodology was applied to cluster protein expression data. Deep convolutional autoencoder was utilized in another recent work named "Unsupervised clustering of SARS-CoV-2 using deep convolutional autoencoder" that was released on August 17, 2022. [9]

2.3 Comparative Analysis and Summary

Comparative analysis will allow us to see the parallels and contrasts between our study project and earlier studies on this particular topic. We used deep learning classifiers to create a model that can cluster the protein sequence of the influenza A virus, much as the research papers mentioned in the references. If we examine the majority of the prior research, we can find that there hasn't been much progress in applying machine learning methods rather than deep learning approaches to cluster the influenza A virus protein sequence. Regarding that, the foundation of our study is the deep learning clustering of influenza A protein sequences. We used the nucleotide sequences of every influenza A virus serotype's matrix protein as a research project dataset. This method of computational biology study offers an alternative to past studies in the area.

2.4 Scope of the Problem

The size of the influenza problem Depending on the precise objectives and context of the study, viral protein sequence clustering utilizing deep embedding networks and K-means can be defined in a variety of ways. Here are some possible methods for defining the problem's size:

1. Geographical scope: The analysis may concentrate on protein sequences from a particular area or nation, or it may have a worldwide scope and include sequences from several nations.
2. Temporal scope: To follow the evolution of the virus over time, the analysis may be performed on protein sequences from a particular period, or it might be broad in scope and include sequences from various points in time.
3. Protein type: The study may concentrate on particular influenza virus proteins, such as the hemagglutinin or neuraminidase proteins, or it may take into account various protein types.
4. Data source: The analysis may rely on protein sequences from a single data source, like the Influenza Research Database, or it may draw from a variety of sources.
5. Clustering technique: The analysis might concentrate on the application of deep embedding networks and K-means in particular, or it could assess how well these techniques perform compared to other clustering techniques such as multiple sequence alignment or the creation of phylogenetic trees.

The size of the influenza problem overall Deep embedded networks and K-means can be used to cluster virus protein sequences in a way that is relevant to the analysis's objectives and situation.

2.5 Challenges

Our lack of research understanding regarding virology and the cell chemistry of the virus posed the biggest hurdle for us as computer science students. We addressed this difficulty by researching and evaluating the relevant and prior efforts in that field. Creating the ideal dataset for our model to use was our next obstacle. The raw dataset we downloaded from the database source was in the FASTA file format, which is not accessible by deep learning classifiers. So, we have to convert

the dataset file type from FASTA to CSV to make the dataset understandable by a deep learning classifier. The dataset, however, had a sizable amount of information and items that were irrelevant to our grouping approach. Consequently, we had a difficult time identifying the appropriate and crucial traits. We could overcome that difficulty by preprocessing the raw information and only using the features essential to our clustering model.

Chapter 3

Research Methodology

3.1 Introduction

In this section, we will describe the workflow for our proposed project, which clusters the influenza A protein sequence. The key elements, including data collection, processing, and the suggested model, are also explained in addition to the essential equations, graphs, tables, and descriptions. In this ground-breaking project, we employed the dataset we collected from the Influenza Resource Center together with a modified version of RNN, a deep embedding network, PCA, and K-means clustering techniques. The latter part of this chapter includes the necessary implementation requirements in addition to the statistical support for our project hypotheses.

3.2 Research Subject and Instrumentation

The research topic covers subjects that are relevant to gaining or developing a clear understanding of the issue. the application of a design model, dataset collection, dataset processing, model training, and addition of alterations based on the dataset. Instrumentation is essentially the technology and procedures that have been employed in the other section. Therefore, in the proposed work, we chose Python as the programming language and a variety of packages, including NumPy, pandas, Skit Learn, Matplotlib, Seaborn, etc. Google Colab is a web-based Python IDE that also supports the use of cloud storage for data science and machine learning applications. It was utilized for all the training and testing operations.

3.3 Workflow

Here is a general workflow for performing Influenza A virus protein sequence clustering using deep embedded networks, PCA, and K-means:

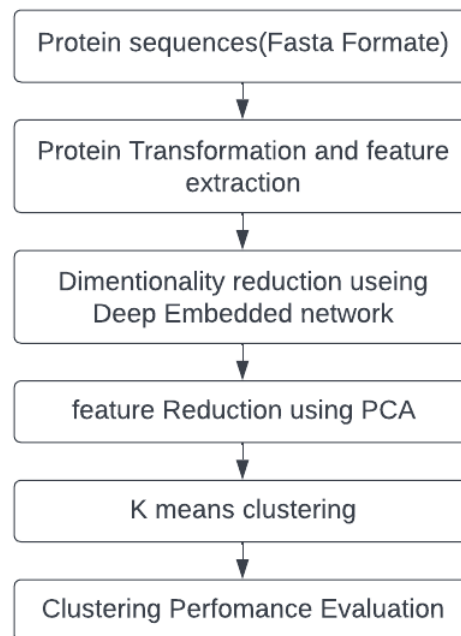


Fig 3.1: Workflow Diagram

1. Preprocessing: Before the protein sequences can be clustered, they may need to be preprocessed to ensure that they are in a suitable format. This may involve tasks such as filtering out low-quality or incomplete sequences, aligning the sequences, and encoding them for input into the deep learning model and machine learning model.
2. Training the deep embedded network: Next, a deep embedded network is applied to the preprocessed protein sequences. The goal of this step is to learn a low-dimensional representation of the sequences and convert them into vector form that captures important relationships between them.
3. Dimensionality reduction: Once the deep embedded network has been done, it is used to transform the protein sequences into the learned low-dimensional representation. Then we applied PCA to make the representation applicable to the k-means clustering.
4. Clustering: The transformed protein sequences are then fed into a traditional clustering algorithm, K-means, to group the sequences into clusters.
5. Evaluation: The clusters produced by the K-means algorithm are evaluated using various metrics such as the elbow method and adjusted Silhouette analysis, Elbow method gives us an idea on

what a good k number of clusters would be based on the sum of squared distance (SSE) between data points and their assigned clusters centroids, Silhouette analysis can be used to determine the degree of separation between clusters. The performance of the clustering can also be visualized through the use of plots such as dendrograms or scatter plots.

6. Interpretation: Finally, the clusters produced by the method are interpreted in the context of the specific goals of the analysis.

3.4 Data Collection Procedure

Accession	Length	Host	Protein	Subtype	Country	Region	Date	Virus name	Mutations	Age	Gender	Lineage	VacStr	Complete	#
BBC90936	759	Avian	PB2	H5N6	Japan	N	2017/01/23	Influenza A Virus (A/mute swan/Hyogo/2801ITM015/2017)(H5N6)						c	
BBC90940	566	Avian		H5N6	Japan	N	2017/01/23	Influenza A Virus (A/mute swan/Hyogo/2801ITM015/2017)(H5N6)						c	
BBC90946	225	Avian	NS1	H5N6	Japan	N	2017/01/23	Influenza A Virus (A/mute swan/Hyogo/2801ITM015/2017)(H5N6)						c	
AWC05827	759	Avian	PB2	H9N2	China	N	2017/06/15	Influenza A virus (A/Anhui/1/2017)(H9N2)						c	2
AWC05828	757	Avian	PB1	H9N2	China	N	2017/06/15	Influenza A virus (A/Anhui/1/2017)(H9N2)						c	2
AWC05830	716	Avian	PA	H9N2	China	N	2017/06/15	Influenza A virus (A/Anhui/1/2017)(H9N2)						c	2
AWC05832	560	Avian		H9N2	China	N	2017/06/15	Influenza A virus (A/Anhui/1/2017)(H9N2)						c	2
AWC05833	498	Avian	NP	H9N2	China	N	2017/06/15	Influenza A virus (A/Anhui/1/2017)(H9N2)						c	2
AWC05834	466	Avian	NA	H9N2	China	N	2017/06/15	Influenza A virus (A/Anhui/1/2017)(H9N2)						c	2
AWC05837	217	Avian	NS1	H9N2	China	N	2017/06/15	Influenza A virus (A/Anhui/1/2017)(H9N2)						c	2
QCF28387	566	Avian		H6N5	South Korea	N	2017/12/07	Influenza A virus (A/Alx galericulata/South Korea/K17-1638-5/2017)						c	
QCF28392	757	Avian	PB1	H6N5	South Korea	N	2017/12/07	Influenza A virus (A/Alx galericulata/South Korea/K17-1638-5/2017)						c	
QCF28393	78	Avian		H6N5	South Korea	N	2017/12/07	Influenza A virus (A/Alx galericulata/South Korea/K17-1638-5/2017)						c	
QCF28394	759	Avian	PB2	H6N5	South Korea	N	2017/12/07	Influenza A virus (A/Alx galericulata/South Korea/K17-1638-5/2017)						c	
QCF28395	473	Avian		H6N5	South Korea	N	2017/12/07	Influenza A virus (A/Alx galericulata/South Korea/K17-1638-5/2017)						c	
AQS21137	759	Human	PB2	H3N2	USA	N	2017/01/04	Influenza A virus (A/Alabama/01/2017)(H3N2)	E527K					c	
AQS21138	757	Human	PB1	H3N2	USA	N	2017/01/04	Influenza A virus (A/Alabama/01/2017)(H3N2)						c	1
AQS21139	90	Human	PB1-F2	H3N2	USA	N	2017/01/04	Influenza A virus (A/Alabama/01/2017)(H3N2)						c	1
AQS21140	716	Human	PA	H3N2	USA	N	2017/01/04	Influenza A virus (A/Alabama/01/2017)(H3N2)						c	
AQS21142	566	Human	HA	H3N2	USA	N	2017/01/04	Influenza A virus (A/Alabama/01/2017)(H3N2)						c	
AQS21144	469	Human	NA	H3N2	USA	N	2017/01/04	Influenza A virus (A/Alabama/01/2017)(H3N2)						c	
AVP08348	759	Human	PB2	H1N1	USA	N	2018/01/02	Influenza A virus (A/Alabama/01/2018)(H1N1)						c	
AVP08352	566	Human		H1N1	USA	N	2018/01/02	Influenza A virus (A/Alabama/01/2018)(H1N1)						c	
AGS21131	498	Human	NP	H3N2	USA	N	2017/01/05	Influenza A virus (A/Alabama/02/2017)(H3N2)						c	
AGS21136	121	Human	NS2	H3N2	USA	N	2017/01/05	Influenza A virus (A/Alabama/02/2017)(H3N2)						c	5
AVP08365	469	Human	NA	H1N1	USA	N	2018/01/03	Influenza A virus (A/Alabama/02/2018)(H1N1)						c	2

Fig 3.2: Source of Data

Our data collection was obtained using the NCBI (Influenza virus Resources). The database contains files for various viruses, and for our research, we collected all of the Influenza A virus's protein sequences and serotypes as well as its nucleotide sequences from all influenza serotypes from 2017 to 2019. There are some distinctive sequence types in these portions. The dataset of all protein sequences, which contains 38985 protein sequences after collapsing, was chosen after some analysis and testing with several datasets (209831total). We also obtained the accession, serotype, and segment name, Host, of each sequence from the source to provide more details.

3.4.1 Preparation of dataset

We have gathered the FASTA-formatted data files from the dataset and converted them to CSV files. The following datasets were obtained with the aid of Python and the FASTA to CSV tool in Python. An example of a dataset:

	Properties	Sequence	Accession	Host	P_Name	Sub_Type
0	AVP08348 Human H1N1 PB2	MERIKELRDLMSQSRTREILTKTTVDHMAIIKKYTSGRQEKNPALR...	AVP08348	Human	PB2	H1N1
1	AVP08352 Human H1N1	MKAILVLLYTFXTANADTLCIGYHANNSTDTVDTVLEKNVTVTHS...	AVP08352	Human		H1N1
2	AVP08365 Human H1N1 NA	MNPNQKIITIGSICMTIGMANLILQIGNIISIWVNHSIQIGNQSQI...	AVP08365	Human	NA	H1N1
3	AVP06268 Human H3N2 NA	MNPNQKIITIGSVSLTISTICFFMQIALITTVTLHFQYEFNSPT...	AVP06268	Human	NA	H3N2
4	AVP06269 Human H3N2 NS1	MDSNTVSSFQVDCFLWHIRKQVVDQKLSDAPFLDRLRRDQRSRGR...	AVP06269	Human	NS1	H3N2
...
4494	QCT08222 Human H1N1 NA	MNPNQKIITIGSICMTIGMANLILQIGNIISIWVSHSIQIGNQSQI...	QCT08222	Human	NA	H1N1
4495	QCT08357 Human H1N1 PA	MEDFVRQCFNPMIVELAEKAMKEYGEDPKIETNKFAAICTHLEVCF...	QCT08357	Human	PA	H1N1
4496	QCT08358 Human H1N1	MEDFVRQCFNPMIVELAEKAMKEYGEDPKIETNKFAAICTHLEVCF...	QCT08358	Human		H1N1
4497	QCT08118 Human H1N1	MKAILVLLYTFTTAKADTLCIGYHANNSTDTVDTVLEKNVTVTHS...	QCT08118	Human		H1N1
4498	QCT08666 Human H3N2 PB1-F2	MEQGQGLTWTQSTEHINTQRGGSGQQIQKLRPSSTQLMDHYLRIM...	QCT08666	Human	PB1-F2	H3N2

Fig 3.3: Sample data from the dataset

3.5 Data Processing

Processing of data One of the most essential aspects of the job is the pre-processing of any dataset before a model is trained on it to improve accuracy. In our research, we also performed pre-processing on the data before using a deep learning model on it. If we can break down our data preparation approach into the two processes of preparing datasets and balancing datasets, then data processing will become more significant in our thesis work.

3.5.1 Balancing Dataset

Each class's data came in an uneven amount. All serotypes' 124059 protein sequences were discovered. So, to avoid biased results, the dataset needed to be balanced. This is why we created the final dataset, which consists of 4499 sequences, using 38985 sequences from all of the chosen classes. The dataset may be applied to our suggested models once it has been balanced.

3.6 Proposed Methodology

In research work, we applied three methods which are:

1. Deep embedded network as a learning feature
2. Then we applied PCA for a dimensional reduction
3. We used k-means clustering.

3.6.1 Deep Embedded Network

The embedding technique, which converts discrete variables into continuous vectors, is one of deep learning's most effective applications. With word embeddings for machine translation and entity embeddings for categorical data, this technology has found useful uses. A discrete, categorical variable is mapped to a vector of continuous numbers using an embedding. Embeddings are low-dimensional, continuously learn vector representations of discrete variables used in neural networks. Because they may make categorical variables less dimensional and accurately reflect categories in the converted space, neural network embeddings are helpful. Neural network embeddings serve three main objectives:

1. Identifying the embedding space's closest neighbors.
2. As data for an unsupervised task in a machine learning model.
3. To display relationships between categories and concepts.

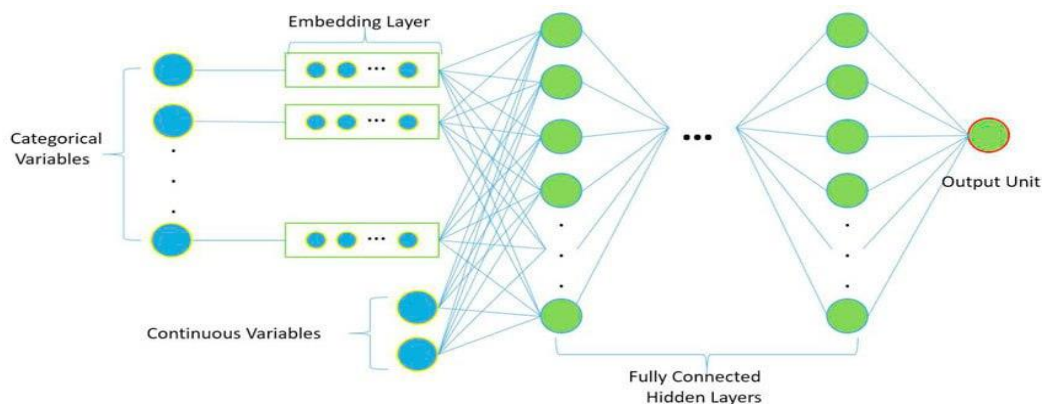


Fig 3.4: Deep Embedded Network

In fig 3.3 , As we can see in embedded network we have to take two input, consisting one categorical variable and another continuous variable to get a vector form output . So we have taken protein sub type as our categorical input and protein sequences as continuous input so that they can reduce the dimensionality of categorical variables and meaningfully represent categories in the transformed space as vector form.

3.6.1 Principal Component Analysis (PCA)

PCA is a dimensionality reduction technique that is used to condense a big collection of variables into a smaller set of variables, which also happens to contain the majority of the data in the original dataset. The entire process consists of the following five steps:

Step 1:- Standardization : This step's goal is to ensure that all of the continuous beginning variables contribute equally to the analysis in order to avoid biased outcomes. It is accomplished mathematically by deducting the mean from the value and dividing by the standard deviation.

$$z = \frac{\textit{value} - \textit{mean}}{\textit{standard deviation}}$$

Equation 3.1: Standardization

Once it is done , all the variables will be transformed to the same scale.

STEP 2: COVARIANCE MATRIX COMPUTATION: The purpose of this stage is to determine the relationship—if any—between the variables in the input data set and how they differ from the mean in relation to one another. Because variables can occasionally be highly connected to the point where they include redundant data. We compute the covariance matrix in order to find these associations. The covariance matrix, which has entries for all potential pairs of the initial variables, is a p p symmetric matrix (where p is the number of dimensions). For instance, the covariance matrix for a 2-dimensional data set with the variables x and y is a 2x2 matrix with the following from:

$$\begin{matrix} Cov(x, x) & Cov(x, y) \\ Cov(y, x) & Cov(y, y) \end{matrix}$$

Since a variable's variance is equal to its covariance with itself ($\text{Cov}(a,a)=\text{Var}(a)$), we really have the variances of each starting variable along the major diagonal (top left to bottom right). The entries of the covariance matrix are symmetric with regard to the main diagonal since the covariance is commutative ($\text{Cov}(a,b)=\text{Cov}(b,a)$), which means that the upper and lower triangular parts are equal.

STEP 3: COMPUTE THE EIGENVECTORS AND EIGENVALUES OF THE COVARIANCE MATRIX TO IDENTIFY THE PRINCIPAL COMPONENTS:

They always come in pairs, which means that every eigenvector has an eigenvalue, which is the first thing we need to understand about them. And the number of them is the same as the number of data dimensions. For instance, since there are 3 variables in a 3-dimensional data set, there are 3 eigenvectors and 3 corresponding eigenvalues. The principal components, or axis with the largest variation (or information), are what we refer to as the eigenvectors of the covariance matrix. The variance held by each Principal Component is indicated by the eigenvalues, which are simply the coefficients associated to the eigenvectors. We obtain the principal components in order of importance by ordering our eigenvectors from highest to lowest according to their eigenvalues.

Example:

Let's assume that our data set has two dimensions and the variables x and y, and that the covariance matrix's eigenvectors and eigenvalues are as follows:

$$v1 = \begin{bmatrix} 0.6778736 \\ 0.7351785 \end{bmatrix} \quad \lambda_1 = 1.284028$$

$$v2 = \begin{bmatrix} -0.7351785 \\ 0.6778736 \end{bmatrix} \quad \lambda_2 = 0.04908323$$

If we order the eigenvalues in descending order, we obtain $\lambda_1 > \lambda_2$, which denotes that the eigenvectors for the first and second principal components (PC1 and PC2, respectively) are v1 and v2, respectively.

STEP 4: FEATURE VECTOR: As we learned in the preceding phase, computing the eigenvectors and ranking them according to their eigenvalues in descending order enable us to

identify the primary components in terms of their relative importance. Choosing whether to maintain all of these components or toss out any that are less important (have low eigenvalues) allows us to combine the remaining ones into a matrix of vectors that we refer to as the "Feature vector" in this stage. The eigenvectors of the components that we choose to keep are placed in columns of the matrix that makes up the feature vector. As a result, it can be considered the initial stage in the process of dimensionality reduction since, if just p of the eigenvectors (components) out of n are retained, the resulting data set will only have p dimensions.

Example:

Continuing with the example from the previous step, we can either form a feature vector with both of the eigenvectors v_1 and v_2 :

$$\begin{bmatrix} 0.6778736 & -0.7351785 \\ 0.7351785 & 0.6778736 \end{bmatrix}$$

Or discard the eigenvector v_2 , which is the one of lesser significance, and form a feature vector with v_1 only:

$$\begin{bmatrix} 0.6778736 \\ 0.7351785 \end{bmatrix}$$

Discarding the eigenvector v_2 will reduce dimensionality by 1, and will consequently cause a loss of information in the final data set. But given that v_2 was carrying only 4% of the information, the loss will be therefore not important and we will still have 96% of the information that is carried by v_1 .

LAST STEP: RECAST THE DATA ALONG THE PRINCIPAL COMPONENTS AXES:

The input data set is always seen in terms of the original axes in the preceding processes, with the exception of standardization, where we simply choose the principal components and create the feature vector (i.e, in terms of the initial variables). The goal of this final step is to reorient the data from the original axes to those represented by the principal components using the feature vector created using the eigenvectors of the covariance matrix (hence the name Principal Components Analysis). To achieve this, multiply the feature vector's transpose by the original data set's transpose.

$$FinalDataSet = FeatureVector^T * StandardizedOriginalDataSet^T$$

Equation 3.2: Feature Vector

	x1	x2	class
0	-0.030862	-0.018930	H1N1
1	-0.029474	-0.013020	H1N1
2	-0.024641	-0.012889	H1N1
3	-0.022898	-0.012760	H3N2
4	-0.003883	-0.008616	H3N2
...
4231	-0.024878	-0.012355	H1N1
4232	-0.032135	-0.016176	H1N1
4233	-0.002429	0.000544	H1N1
4234	-0.030070	-0.013068	H1N1
4235	0.743522	-0.119411	H3N2

Fig 3.5: Feature Vector formation

This is our obtained data as a featured vector alongside two principal component and their classes.

Figure size 1000x1000 with 0 Axes>

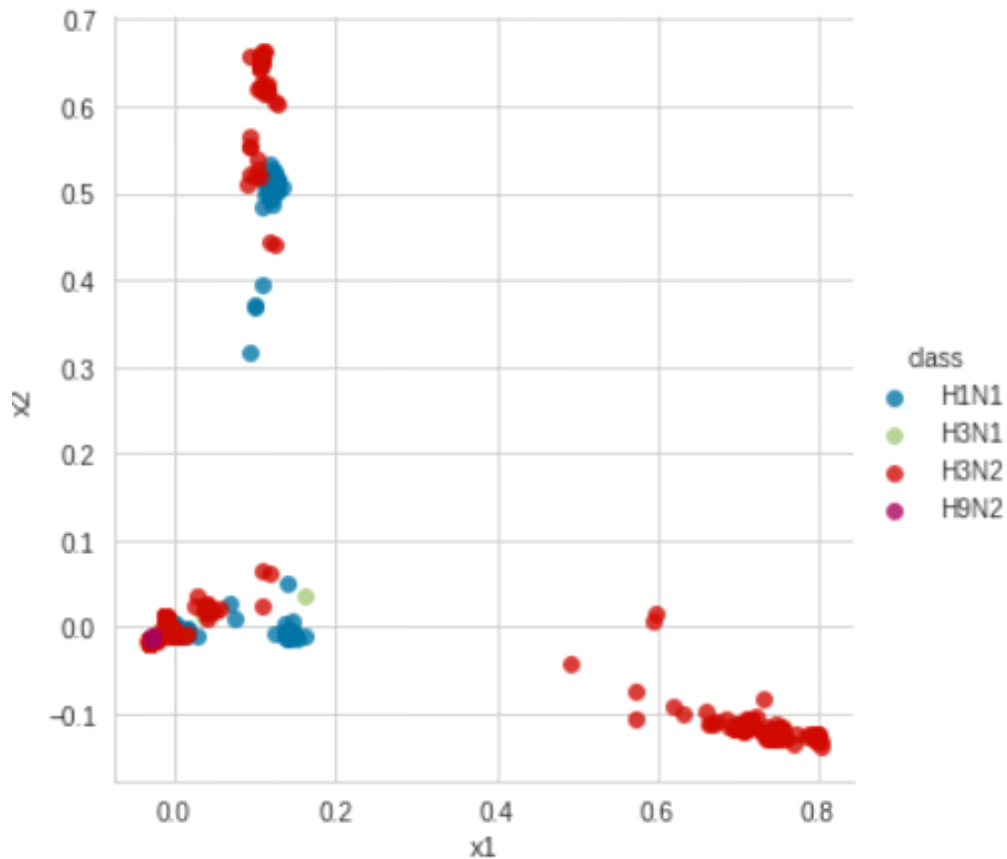


Fig 3.6: Principal Components

This figure shows the plot diagram of PCA components and their classes on our dataset.

3.6.2 K-Means Clustering

K-Means divides the dataset into k (a hyper-parameter) clusters using an iterative optimization strategy. Each cluster is represented by a center. A point belongs to a cluster whose center is closest to it. For simplicity, assume that the centers are randomly initialized. The goal of the model is to find clusters that minimize the sum of SSE over k clusters by shifting their centers. SSE or Sum of Squared Errors of a cluster is the sum of squared distances between its center and its points.

Working of K-Means Algorithm:

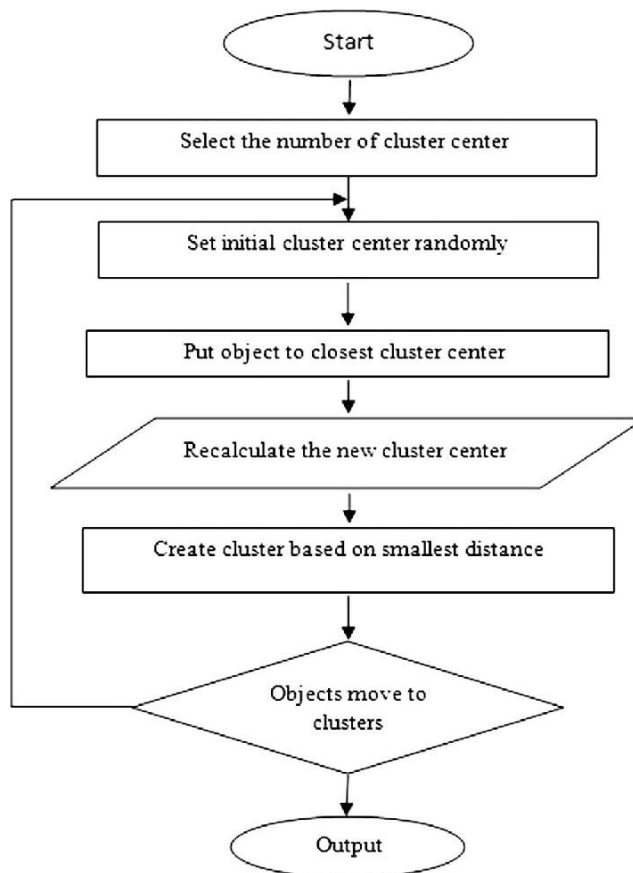


Fig 3.7: Workflow of K-means Clustering

From the above diagram, The stages listed below can help us understand how the K-Means clustering technique operates:

Step 1: The first thing we must do is state how many clusters, K , our algorithm must produce.

Step 2: Next, choose K data points at random and group each one into a cluster. Simply said, categorize the data according to the quantity of data points.

Step 3: At this point, the cluster centroids will be computed.

Step 4 - After that, keep repeating the steps below until you locate the ideal centroid, which is the assignment of data points to clusters that are stable.

4.1 The total of the squared distances between the centroids and the data points would be calculated first.

4.2 – At this stage, we must allocate each data point to the cluster that is nearest to it (centroid).

4.3 – Finally, calculate the centroids for each cluster by averaging all of its data points.

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Experimental Setup

A clustering report, the sum of square distance using elbow method and adjusted Silhouette analysis, are used in this chapter to show and discuss the study's findings. We took into account three methods while we created our model, including Deep Embedded Network, PCA, and K-means Clustering.

4.2 Performance Analysis

Before getting good clustering results we had to make sure the sequences are in a valid format and sequence analysis performance had been exact. We got our expected performance throughout the process.

	Properties	Sequence
0	AVP08348 Human H1N1 PB2	MERIKELRDLMSSQSRTEILTKTTVDHMAIIKKYTSGRQEKNPALR...
1	AVP08352 Human H1N1	MKAILVLLLYTFXTANADTLCIGYHANNSTDTVDTVLEKNVTVTHS...
2	AVP08365 Human H1N1 NA	MNPNQKIITIGSICMTIGMANLILQIGNIISIWVNHSIQIGNQSQI...
3	AVP06268 Human H3N2 NA	MNPNQKIITIGSVSLTISTICFFMQIALITTVTLHFQYEFNSPT...
4	AVP06269 Human H3N2 NS1	MDSNTVSSFQVDCFLWHIRKQVVDQKLSDAPFLDRLRRDQQRSLRGR...
...
4494	QCT08222 Human H1N1 NA	MNPNQKIITIGSICMTIGMANLILQIGNIISIWVSHSIQIGNQSQI...
4495	QCT08357 Human H1N1 PA	MEDFVRQCFNPMIVELAEKAMKEYGEDPKIETNKFAAICTHLEVCF...
4496	QCT08358 Human H1N1	MEDFVRQCFNPMIVELAEKAMKEYGEDPKIETNKFAAICTHLEVCF...
4497	QCT08118 Human H1N1	MKAILVLLLYTFTTAKADTLCIGYHANNSTDTVDTVLEKNVTVTHS...
4498	QCT08666 Human H3N2 PB1-F2	MEQQQGTLWTQSTEHINTQRGGSGQQIKLGRPSSTQLMDHYLRIM...

Figure 4.1: Converted CSV data from FASTA

Figure 4.1 shows unreadable FASTA data is converted into csv data containing to columns which are properties and sequence.

	Sequence	Accession	Host	SubType	P_Name
0	MERIKELRDLMSQSRRTREILTKTTVDHMAIIKKYTSGRQEKNPALR...	AVP08348	Human	H1N1	PB2
1	MKAILVLLYTFXTANADTLCIGYHANNSTDTVDTVLEKNVTVTHS...	AVP08352	Human	H1N1	
2	MNPNQKIITIGSICMTIGMANLILQIGNIISIWVNHSIQIGNQSQI...	AVP08365	Human	H1N1	NA
3	MNPNQKIITIGSVSLTISTICFFMQIAILITVTLHFQKQYEFNSPT...	AVP06268	Human	H3N2	NA
4	MDSNTVSSFQVDCFLWHIRKQVVDQKLSADPFLDRLRRDQSLRGR...	AVP06269	Human	H3N2	NS1
...
4494	MNPNQKIITIGSICMTIGMANLILQIGNIISIWVSHSIQIGNQSQI...	QCT08222	Human	H1N1	NA
4495	MEDFVRQCFNPMIVELA EKAMKEYGEDPKIETNKFAAICTHLEVCF...	QCT08357	Human	H1N1	PA
4496	MEDFVRQCFNPMIVELA EKAMKEYGEDPKIETNKFAAICTHLEVCF...	QCT08358	Human	H1N1	
4497	MKAILVLLYFTTAKADTLCIGYHANNSTDTVDTVLEKNVTVTHS...	QCT08118	Human	H1N1	
4498	MEQGGGLWTQSTEHINTQRGGSGQIQKLRPSSTQLMDHYLRIM...	QCT08666	Human	H3N2	PB1-F2

4499 rows × 5 columns

Figure 4.2: Split Properties

In our previous figure 4.1, some important features were included in one column “properties” which must be split into different columns as in figure 4.2 which are accession, host, protein name, and subtype.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0.002613	2.230682e-04	0.001886	0.000674	1.258297e-04	0.001363	8.908393e-06	0.002608	0.001930	0.001445	0.001476	1.352362e-03	0.002167	0.000690	0.001405	0.000724	0.002097	0.001791	0.000395	0.001345
C	0.001950	1.879529e-13	0.000003	0.000897	2.636134e-07	0.000069	2.043775e-02	0.001805	0.006388	0.000044	0.004050	4.552417e-07	0.000243	0.000592	0.000019	0.000713	0.000348	0.000042	0.013625	0.000160
D	0.001991	8.678249e-03	0.001354	0.001547	4.775425e-04	0.000719	4.117177e-03	0.001841	0.001528	0.001623	0.002204	1.225151e-03	0.002855	0.002092	0.001312	0.001549	0.001501	0.001976	0.000003	0.000843
E	0.001090	2.748404e-03	0.001434	0.002254	1.463219e-03	0.000840	7.342213e-04	0.001196	0.002017	0.001406	0.001949	6.885002e-04	0.000886	0.002711	0.001468	0.001314	0.001017	0.001825	0.000023	0.001423
F	0.001338	2.245982e-04	0.001205	0.001705	1.443948e-03	0.001760	3.807718e-04	0.000289	0.001635	0.001681	0.000635	2.531639e-03	0.002859	0.004902	0.001459	0.002232	0.001977	0.001695	0.000630	0.000927
G	0.001263	1.828869e-03	0.002004	0.001842	2.869117e-03	0.001868	3.602566e-03	0.001480	0.001305	0.001183	0.000742	1.608035e-03	0.002844	0.001237	0.001590	0.001679	0.002888	0.001728	0.000385	0.001457
H	0.001684	7.746101e-05	0.000644	0.001101	4.141815e-03	0.001513	1.699980e-04	0.000240	0.000416	0.001283	0.001754	2.085797e-04	0.000764	0.002403	0.000386	0.000959	0.001217	0.000067	0.000015	0.004173
I	0.001923	4.044009e-03	0.001113	0.001316	1.407749e-04	0.001539	1.280583e-04	0.002085	0.002528	0.002246	0.001394	1.836043e-03	0.001281	0.001075	0.001757	0.000671	0.001377	0.001756	0.000940	0.000551
K	0.002727	3.289910e-05	0.001427	0.002590	3.730971e-04	0.000653	2.287768e-03	0.001237	0.001284	0.001116	0.001909	9.459228e-04	0.000688	0.000693	0.002416	0.000178	0.002153	0.001060	0.001993	0.003564
L	0.001608	1.252040e-03	0.000597	0.001000	1.650098e-03	0.001332	4.937927e-03	0.001850	0.001855	0.001193	0.001652	1.566440e-03	0.002322	0.001645	0.002700	0.001428	0.002876	0.001765	0.001812	0.001039
M	0.001810	5.344802e-03	0.000531	0.001708	8.748049e-04	0.001182	4.214132e-03	0.001650	0.001816	0.001130	0.002189	7.874538e-04	0.000671	0.001019	0.000808	0.001065	0.000752	0.001925	0.003759	0.002537
N	0.001078	6.132081e-03	0.001599	0.002081	9.366024e-04	0.001569	1.003770e-03	0.001341	0.002048	0.001191	0.000947	7.521393e-04	0.004415	0.001953	0.001237	0.000529	0.000871	0.002111	0.005481	0.002099
P	0.002442	7.522533e-06	0.001555	0.003229	2.368088e-03	0.001300	4.356607e-03	0.000530	0.001245	0.001506	0.002045	9.894273e-04	0.001054	0.000787	0.000318	0.000961	0.002337	0.001683	0.000104	0.000460
Q	0.001032	4.027749e-03	0.002142	0.001058	8.258288e-04	0.002043	2.387480e-04	0.001901	0.001212	0.001713	0.001994	2.074550e-03	0.000763	0.000975	0.001347	0.002167	0.002026	0.000796	0.006541	0.000561
R	0.001480	3.747135e-09	0.002378	0.000921	1.465826e-03	0.001751	2.955269e-03	0.001947	0.001312	0.001411	0.001763	2.704826e-03	0.000306	0.001023	0.001752	0.000899	0.001212	0.001733	0.000711	0.001798
S	0.001498	3.343068e-04	0.001147	0.001561	1.951369e-03	0.002089	4.187908e-04	0.001648	0.000982	0.001886	0.001683	9.631902e-04	0.002003	0.002617	0.001450	0.002906	0.002132	0.001572	0.000391	0.000858
T	0.000959	3.361533e-03	0.001310	0.001853	2.004794e-03	0.001344	1.368821e-03	0.000867	0.001781	0.000925	0.001164	1.149803e-03	0.001327	0.001614	0.001430	0.001838	0.001220	0.001437	0.003451	0.003082
V	0.001272	4.626065e-04	0.002557	0.000956	1.985690e-03	0.001339	8.201427e-03	0.001063	0.001672	0.001818	0.002072	1.954124e-03	0.001032	0.000746	0.002373	0.001651	0.000730	0.001093	0.000888	0.002766
W	0.000203	3.383101e-21	0.000300	0.004123	1.476933e-08	0.001365	5.711479e-07	0.003231	0.000645	0.000003	0.003587	3.261357e-03	0.000226	0.001333	0.000714	0.002243	0.000299	0.000140	0.008345	0.000235
Y	0.000297	4.513973e-07	0.000023	0.002089	2.073649e-03	0.001356	1.184092e-04	0.001294	0.002280	0.000301	0.001850	2.442264e-03	0.004287	0.001199	0.000128	0.003145	0.002385	0.000272	0.001982	0.000575

Figure 4.3: Vector Representation

In 4.3 shows our embedded network successfully transform our data into vector representation form, which are used in PCA as input data.

	id	(A, A)	(A, B)	(A, C)	(A, D)	(A, E)	(A, F)	(A, G)	(A, H)	(A, I)	...
0	AVP08348	0.002613	0.0	2.230682e-04	0.001886	0.000674	0.000126	0.001363	0.000009	0.002608	...
1	AVP08352	0.002575	0.0	1.914635e-03	0.004121	0.001469	0.002068	0.003050	0.000537	0.003818	...
2	AVP08365	0.005229	0.0	3.760094e-03	0.000304	0.001469	0.000132	0.001056	0.002740	0.001132	...
3	AVP06268	0.001689	0.0	7.188875e-05	0.004292	0.001883	0.006529	0.002131	0.001945	0.002176	...
4	AVP06269	0.000951	0.0	3.119208e-14	0.000624	0.002392	0.018316	0.002793	0.015037	0.016618	...
...
4494	QCT08222	0.005926	0.0	3.760094e-03	0.000320	0.000015	0.000048	0.001056	0.002740	0.001128	...
4495	QCT08357	0.001258	0.0	2.457181e-03	0.001680	0.001951	0.000186	0.000405	0.000359	0.002509	...
4496	QCT08358	0.005137	0.0	7.150674e-03	0.006612	0.002952	0.000059	0.000209	0.000953	0.006634	...
4497	QCT08118	0.002481	0.0	1.967951e-03	0.004011	0.000893	0.002068	0.002877	0.001141	0.003614	...
4498	QCT08666	0.000000	0.0	0.000000e+00	0.000000	0.000000	0.000912	0.000017	0.000000	0.000000	...

4499 rows x 577 columns

Figure 4.4: Variance Correlation.

Figure 4.4 shows relationship between components are successfully evaluated using PC.

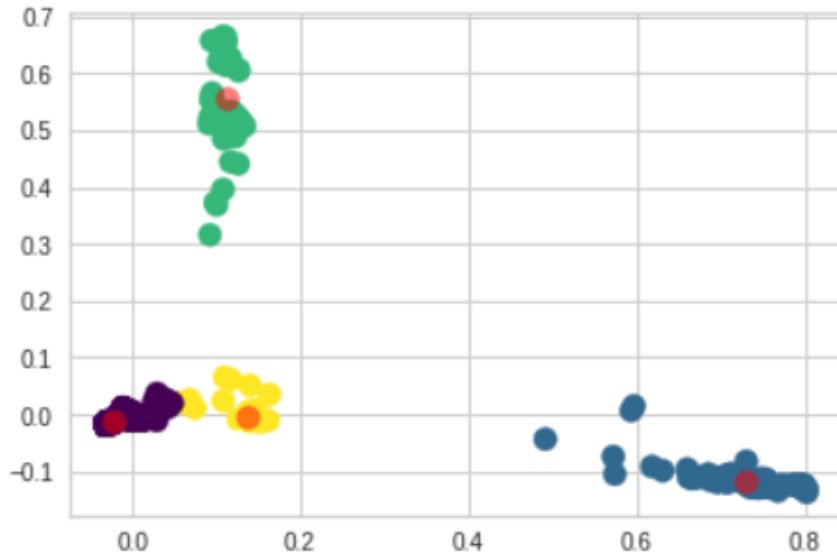


Figure 4.5: Clustering with Centroid

Figure 4.5 shows the similarity among sequences after applying k-means clustering based on centroid with different colors to show the same group.

4.3 Result Discussion

From 4499 sequence vector representation primarily, we worked with 3 components, and successfully we were able to cluster them as we expected.

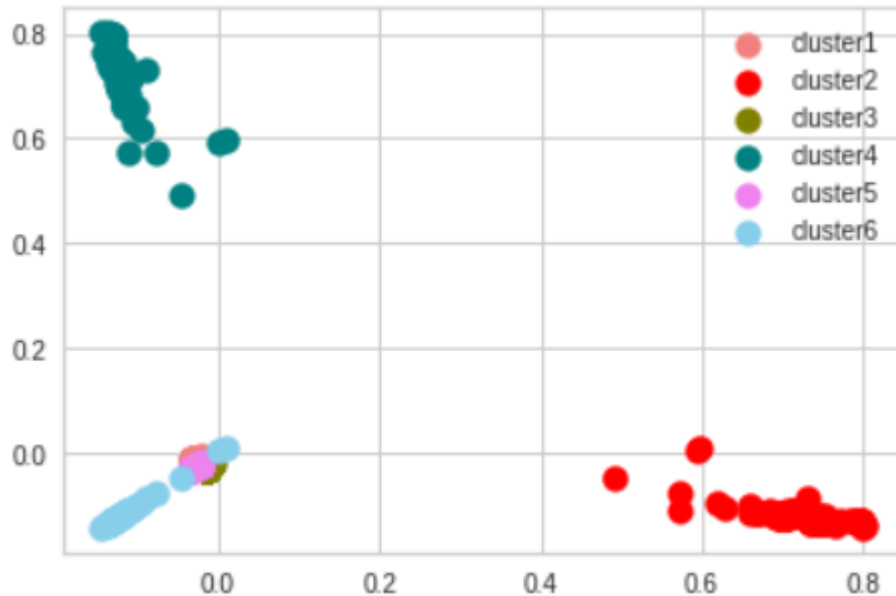


Figure 4.6: Clustering Result

After preliminary work then we worked with 6 components to check whether our model can cluster multiple components and we got our desired result as we can see in figure 4.6.

4.4 Elbow Method

Based on the sum of squared distance (SSE) between data points and the centroids of their assigned clusters, the elbow technique offers us an estimate of what a suitable k number of clusters might be. At the point where SSE begins to flatten out and form an elbow, we pick k.

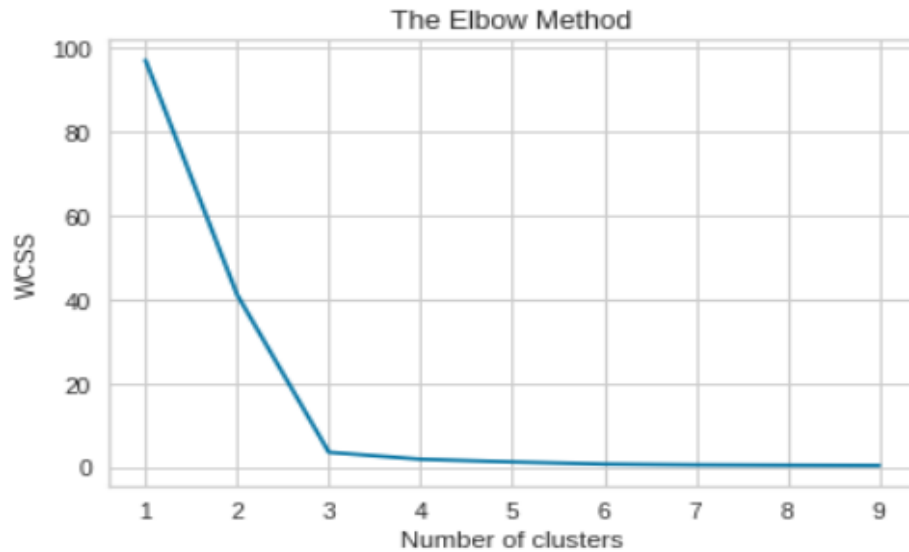


Figure 4.7: Elbow Plotting

The graph demonstrates that $k=3$ is a reasonable choice. Because the curve is monotonically declining, it can be challenging to determine how many clusters to utilize because there may be no elbow or a clear point where the curve begins to flatten out.

Silhouette Analysis:

The degree of separation between clusters can be found via silhouette analysis.

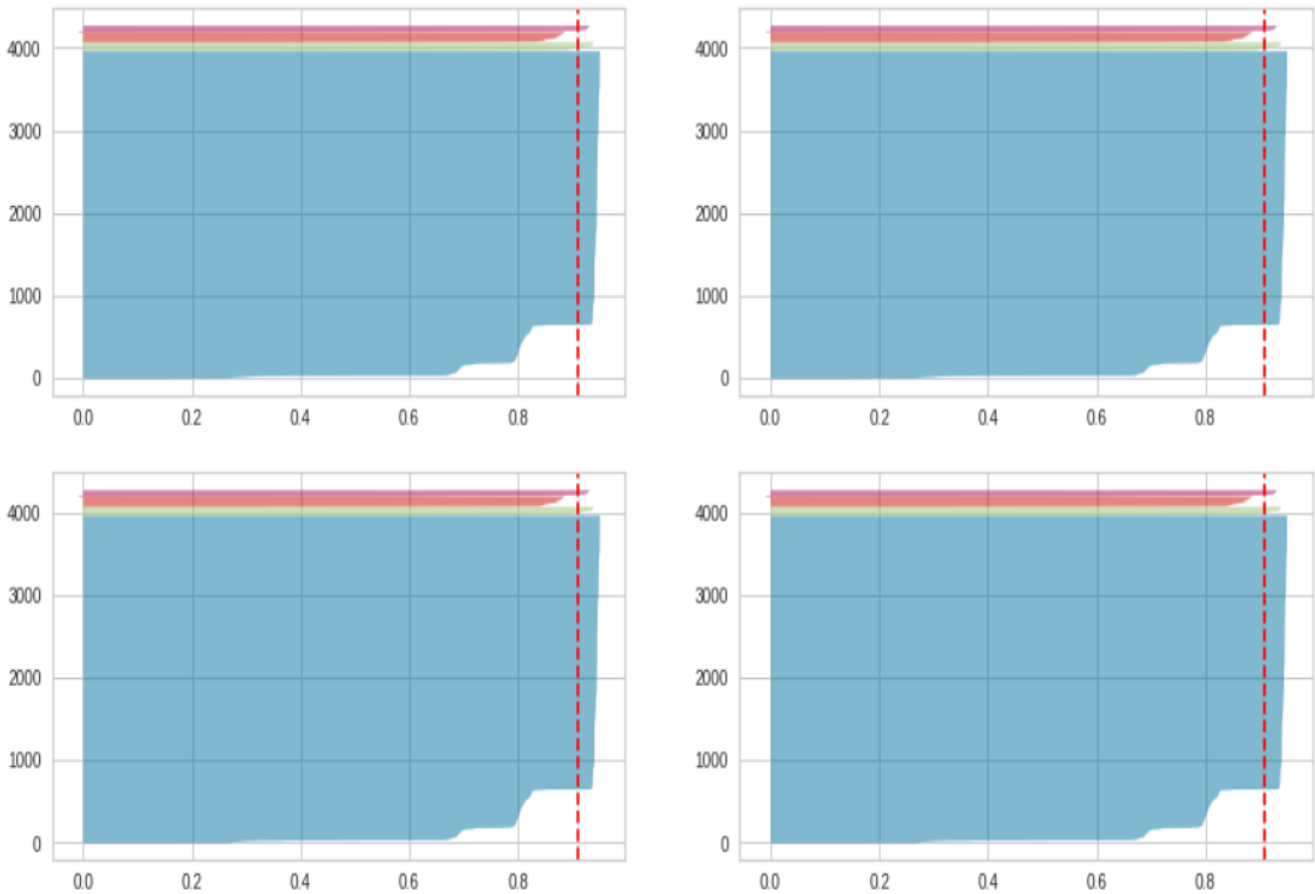
- Calculate the average distance between all the data points in the same cluster for each sample (a_i).
- Calculate the median distance between each data point in the nearest cluster (b_i). Compute the coefficient:

$$\frac{b^i - a^i}{\max(a^i, b^i)}$$

Equation 4.1: Coefficient

The coefficient can take values in the interval $[-1, 1]$.

- If it is 0, the sample is in close proximity to the nearby clusters.
- If the answer is 1, the sample is far from the nearby clusters.



- If it is -1, the sample has been placed in the incorrect clusters.

Figure 4.8: Silhouette Analysis

As we can see above in figure 4.8, we achieved a 0.9 coefficient which is very close to the neighboring cluster to evaluate the performance of our model analysis.

CHAPTER 5

Impact on Society, Environment, and Sustainability

5.1 Impact on Society

The development of potent vaccinations is one way that influenza A virus protein sequence clustering can have an impact on civilization. The main proteins that are necessary for the influenza A virus to multiply and infect host cells can be found by studying the protein sequences of several influenza A virus strains. This knowledge can be utilized to create vaccinations that specifically target these proteins and offer protection from the virus. Additionally, protein sequence clustering can aid in enhancing the precision of diagnostic testing for the influenza A virus. Researchers can create more precise tests that can distinguish between several strains of the influenza A virus and aid in a more precise infection diagnosis by identifying and studying the protein sequences of the various viral strains. Overall, the ability to manage influenza A virus epidemics, enhance the precision of diagnostic tests, and provide information for the creation of efficient vaccinations can be achieved by using protein sequence clustering techniques such as deep embedding networks, PCA, and K means. Last but not least, protein sequence clustering can assist in informing public health policies and methods for limiting influenza A virus outbreaks. Researchers can gain a better understanding of how the virus spreads and changes by examining the protein sequences of various virus strains. This knowledge can be used to develop strategies to stop the transmission of the virus and manage outbreaks.

5.2 Impact on the Environment

Since our research is computer-based, there are no negative environmental effects. In actuality, the research examines viruses, a crucial component of the ecosystem. Because of their variations, viruses are constantly evolving and changing in nature. According to our findings, influenza proteins are clustered. a virus model that allows us to understand how viral strains have changed over time. A perfect clustering can help us to make a relationship among other viruses beside making us understand how the environment caused in making a new sequence. Then it will be easier for the scientist to prevent another pandemic caused by different types of viruses. The way our model would put an important impact on the Environment is assisting scientists to invent vaccines against most dangerous viruses since a new strain of virus can cause a pandemic and a pandemic means loss of lives at a particular period of time that can affect our environment.

5.3 Ethical Aspects

When employing deep embedded networks, PCA, and K methods to cluster the influenza A virus protein sequence, there are a lot of ethical issues to take into account. Potential misuse of the knowledge produced by protein sequence clustering is one ethical issue that can come up. The creation of bioweapons or the discrimination of particular groups could both be made possible by this information. It is crucial to make sure that the outcomes of protein sequence clustering are put to good use for society and that the right controls are in place to prevent this knowledge from being misused. The possible effects of protein sequence clustering on vulnerable populations should be taken into account as another ethical problem. For instance, not all populations may have equal access to the creation of vaccinations or other therapies based on protein sequence clustering. It is crucial to make sure that everyone who requires these therapies can get them and that the advantages of protein sequence clustering are fairly distributed. The possibility of using research subjects for personal gain is another ethical concern. Protein sequence clustering occasionally necessitates the acquisition of human biological samples, which poses questions of informed permission and privacy protection. The privacy of research participants must be maintained, and it is crucial to ensure that they are properly informed about the study. Overall, it is crucial to give careful thought to the moral implications of protein sequence clustering and to take action to make sure that the findings of this study are applied morally and for the good of society. To control the use of this technology and safeguard the rights of research subjects and the general public, this may include developing rules and regulations.

5.4 Sustainability Plan

Our study is a strategic undertaking that can support itself over the long run. Any sustainability measure, including financial, environmental, and community sustainability, can be met by the project with success. The resources used in the project cannot be simply lost; therefore, the developed model can continue to work throughout time. As a result of the project's modest maintenance requirements, sustainability upkeep won't be overly difficult in the future. Our research shows promise and has the potential to develop. Our project has a growth component, which we are aware of and may work on in the future to keep it continuing.

CHAPTER 6

Summary, Conclusion, Recommendation, Implication for Future Research

6.1 Summary of the study

Influenza A virus protein sequence clustering using deep embedded networks, PCA, and K means is a technique that involves analyzing the protein sequences of different strains of influenza A virus in order to better understand the spread and evolution of the virus. This type of protein sequence clustering can have a number of impacts on society, including aiding in the development of effective vaccines, improving the accuracy of diagnostic tests, and informing public health policies and strategies for controlling outbreaks of the virus. There are also ethical considerations to be aware of in the use of this technology, such as the potential for the misuse of the information generated and the potential impact on vulnerable populations. Overall, protein sequence clustering is a valuable tool for understanding and controlling the influenza A virus and its impact on society.

6.2 Conclusions

A method for grouping protein sequences into classes and making distinctions between them is presented in this study. For grouping protein sequences based on unsupervised data, we used the deep embedded network, PCA, and K-means Clustering models. The performance of the clustering model was then examined and tested. Both deep learning methods successfully met the targets. After completing the entire work process, we conclude that the research work has achieved its goals and complied with all standards.

6.3 Implication for Future Study

Our research is an ambitious idea with lots of potential for improvement. We are aware of the room for expansion in our project and can try to take advantage of it in the future. We intend to use this established model to cluster the sequences of the other varieties of the influenza virus genome to make further progress. Future vaccine development for the influenza virus may benefit from the work done in this area.

- [1] Asur, Sitaram, Duygu Ucar, and Srinivasan Parthasarathy. "An ensemble framework for clustering protein–protein interaction networks." *Bioinformatics* 23.13 (2007): i29-i40.
- [2] Schatz, Karsten, et al. "Analyzing the similarity of protein domains by clustering Molecular Surface Maps." *Computers & Graphics* 99 (2021): 114-127.
- [3] Rossi, Andre Luis Debiase, and Maria Angelica de Oliveira Camargo-Brunetto. "Protein classification using artificial neural networks with different protein encoding methods." *Seventh International Conference on Intelligent Systems Design and Applications (ISDA 2007)*. IEEE, 2007.
- [4] Ma, Qicheng, et al. "Clustering protein sequences with a novel metric transformed from sequence similarity scores and sequence alignments with neural networks." *BMC Bioinformatics* 6.1 (2005): 1-13.
- [5] Dai, Xinnan, et al. "PIKE-R2P: Protein–protein interaction network-based knowledge embedding with graph neural network for single-cell RNA to protein prediction." *BMC bioinformatics* 22.6 (2021): 1-17.
- [6] Gutteridge, Alex, Gail J. Bartlett, and Janet M. Thornton. "Using a neural network and spatial clustering to predict the location of active sites in enzymes." *Journal of molecular biology* 330.4 (2003): 719-734.
- [7] Eom, Jae-Hong, and Byoung-Tak Zhang. "Adaptive Neural Network-Based Clustering of Yeast Protein–Protein Interactions." *International Conference on Intelligent Information Technology*. Springer, Berlin, Heidelberg, 2004.
- [8] Khanmohammadi, Mohammadreza, et al. "Artificial neural network for quantitative determination of total protein in yogurt by infrared spectrometry." *Microchemical Journal* 91.1 (2009): 47-52.
- [9] Zheng, Wei, et al. "Detecting distant-homology protein structures by aligning deep neural-network based contact maps." *PLoS computational biology* 15.10 (2019): e1007411.
- [10] Datta, Ayan, et al. "A neural network-based approach for protein structural class prediction." *Journal of Intelligent & Fuzzy Systems* 20.1-2 (2009): 61-71.
- [11] Greene, Derek, et al. "Ensemble non-negative matrix factorization methods for clustering protein–protein interactions." *Bioinformatics* 24.15 (2008): 1722-1728.
- [12] Jarman, Ian H., et al. "Clustering of protein expression data: a benchmark of statistical and neural approaches." *Soft Computing* 15.8 (2011): 1459-1469.
- [13] Sherif, Fayroz F., and Khaled S. Ahmed. "Unsupervised clustering of SARS-CoV-2 using deep convolutional autoencoder." *Journal of Engineering and Applied Science* 69.1 (2022): 1-22.
- [14] Qi, Ren, et al. "Clustering and classification methods for single-cell RNA-sequencing data." *Briefings in bioinformatics* 21.4 (2020): 1196-1208.
- [15] <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8412085>
- [16] <https://naserian-elahe.medium.com/deep-embedding-and-clustering-an-step-by-step-python-implementation-bd2c9d51c80f>
- [17] <https://www.ncbi.nlm.nih.gov/genomes/FLU/Database/nph-select.cgi#mainform>.
- [18] <https://arxiv.org/pdf/2109.15149v1.pdf>
- [19] <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8412085>
- [20] <https://jeas.springeropen.com/articles/10.1186/s44147-022-00125-0>

APPENDICES

We had a lot of challenges finishing the research study. The first step was choosing a methodological strategy for the project. However, the largest problem we ran into during the research was that, as computer science majors, we lacked a solid understanding of virology and the cell chemistry of the virus. By researching and examining relevant and prior efforts in that field, we were able to overcome this challenge. The best dataset for our model to work on needed to be created as our next assignment. The raw dataset we obtained from the database source was a FASTA file, which deep learning classifiers cannot read. So, to make the dataset readable by a deep learning classifier, we had to change the dataset's file format from FASTA to CSV.

Clustering Matrix Protein of Influenza Virus Using Deep Embedded Network

ORIGINALITY REPORT

24%
SIMILARITY INDEX

17%
INTERNET SOURCES

8%
PUBLICATIONS

20%
STUDENT PAPER

PRIMARY SOURCES

1	Submitted to Liverpool John Moores University Student Paper	3%
2	Submitted to Daffodil International University Student Paper	2%
3	Submitted to Carnegie Mellon University Student Paper	2%
4	Submitted to Birla Institute of Technology and Science Pilani Student Paper	2%
5	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
6	towardsdatascience.com Internet Source	1%
7	bmcbioinformatics.biomedcentral.com	1%
8	Submitted to Southampton Solent University Student Paper	1%

