**Malware Categorisation Using Machine Learning**

A dissertation submitted in partial fulfilment of

the requirements for the degree of

BACHELOR OF ENGINEERING in Computer Science

in

The Queen's University of Belfast

By

Jon Kyle

09/05/2017

Table of Contents

[Acknowledgements 4](#_Toc480994591)

[Abstract 4](#_Toc480994592)

[Introduction 4](#_Toc480994593)

[Literature Review 5](#_Toc480994594)

[Dynamic Analysis 6](#_Toc480994595)

[Static Analysis 6](#_Toc480994596)

[PE File Structure 7](#_Toc480994597)

[Proposed Machine Learning Features 8](#_Toc480994598)

[File Entropy 8](#_Toc480994599)

[Ratio of Code to Initialised Data 8](#_Toc480994600)

[Major Image Version 8](#_Toc480994601)

[Number Of Sections 8](#_Toc480994602)

[Common DLL Imports 9](#_Toc480994603)

[Dimensionality Reduction Techniques 10](#_Toc480994604)

[Principle Component Analysis (PCA) 10](#_Toc480994605)

[Non-Negative Matrix Factorisation (NMF) 10](#_Toc480994606)

[Clustering Algorithms 11](#_Toc480994607)

[KMeans 11](#_Toc480994608)

[Mean Shift 12](#_Toc480994609)

[System Requirements Specification 13](#_Toc480994610)

[System constraints 13](#_Toc480994611)

[Functional Requirements 13](#_Toc480994612)

[Non-functional requirements 13](#_Toc480994613)

[User Interface requirements 14](#_Toc480994614)

[User Characteristics 14](#_Toc480994615)

[Design 15](#_Toc480994616)

[System Architecture 15](#_Toc480994617)

[User Interface Design 15](#_Toc480994618)

[Software Design 18](#_Toc480994619)

[Implementation 20](#_Toc480994620)

[Choice of Languages and Libraries 20](#_Toc480994621)

[Development Environment 21](#_Toc480994622)

[Development Hardware 21](#_Toc480994623)

[Key Implementation Decisions 21](#_Toc480994624)

[Implementation of Important Functions/Algorithms 24](#_Toc480994625)

[Calculation of Entropy 24](#_Toc480994626)

[Conversion of Byte Data to Decimal 24](#_Toc480994627)

[Reading Values from PE Files 25](#_Toc480994628)

[Getting Names of .dll Imports from a PE File 26](#_Toc480994629)

[Automation of Optimizing KMeans Parameters 27](#_Toc480994630)

[Accuracy Check 27](#_Toc480994631)

[Decomposition and Clustering Plots 28](#_Toc480994632)

[Color Plots 29](#_Toc480994633)

[Silhouette Plots 30](#_Toc480994634)

[Implementation of individual components 32](#_Toc480994635)

[PE File Parser 32](#_Toc480994636)

[Feature Extraction 32](#_Toc480994637)

[User Interface 32](#_Toc480994638)

[Clustering and Categorisation 33](#_Toc480994639)

[System Evaluation and Experimental Results 34](#_Toc480994640)

[Evaluation of Malware Detection 34](#_Toc480994641)

[Effect of Decomposition Algorithm Choice on Detection Rate 35](#_Toc480994642)

[Effect of Decomposition Number of Components on Detection Rate 35](#_Toc480994643)

[Effect of Clustering Algorithm Choice on Detection Rate 36](#_Toc480994644)

[Effect of Clustering Number of Clusters on Detection Rate 36](#_Toc480994645)

[Evaluation of Malware Family Categorisation 37](#_Toc480994646)

[Effect of Decomposition Algorithm Choice on Clustering Accuracy 37](#_Toc480994647)

[Effect of Number of Components on Clustering Accuracy 38](#_Toc480994648)

[Effect of Clustering Algorithm Choice on Clustering Accuracy 38](#_Toc480994649)

[Effect of Number of Clusters on Clustering Accuracy 39](#_Toc480994650)

[Conclusion 40](#_Toc480994651)

[References 41](#_Toc480994652)

[Appendices 43](#_Toc480994653)

[Appendix 1. Meeting Minutes 43](#_Toc480994654)

[Appendix 2. PE32 Class Full UML 50](#_Toc480994655)

[Appendix 3. peparse.py Sample Command Line Output 51](#_Toc480994656)

[Appendix 4 Full Table of results for Malware Detection Accuracy 54](#_Toc480994657)

[Appendix 5 Full Table of Results for Clustering Accuracy 63](#_Toc480994658)

# **Acknowledgements**

# **Abstract**

Malware categorization techniques play a critical role in safeguarding existing technical infrastructure. To defend against cyber-attacks, companies need to quickly categorise malware into low-risk, medium-risk or high-risk families. Malware comes in many forms and uses many different attack vectors to gain access to computers, in order to cause undesired and potentially harmful behaviors to occur.

This project presents a heuristic-based static analysis tool which extracts features from the binary representation of WIN32 Portable Executable files. These features are then analysed using a number of dimensionality reduction techniques including Principal Component Analysis and Non-Negative Matrix Factorisation, to maximize the variance within the data. This processed data is then used as an input to the KMeans clustering algorithm to categorise the files into clusters.

This project has demonstrated that machine learning is highly effective at categorising malware for the purpose of malware detection within datasets of benign and malicious files. Using the Non-Negative Matrix Factorisation and KMeans algorithms a detection rate of 100% was achieved. This project has also demonstrated that machine learning is not effective when trying to categorise malware into families, achieving only 46.76% accuracy.

# **Introduction**

Malware can often cause harm to companies and in some cases individuals through the interception of network data, encryption or deletion of files, and in extreme cases can cause catastrophic failure on a hardware level through infecting or replacing existing firmware.

The symptoms of malware differ, depending on the malware type or family; As the symptoms differ, so does the potential risk to systems infected by the malware.

Categorization of malware allows us to make an approximation of the potential risk a file poses. Though there are many existing anti-virus scanners available, the results of these scanners can disagree when it comes to the categorization of a piece of malware, and in some rare cases an anti-virus may classify a file as safe, when in fact it is not.

Malware comes in many forms and employs many different propagation techniques to deliver payloads to vulnerable machines. Common propagation techniques include Virus, and Worm which self-replicate locally or across networks. Payloads can range from relatively benign adware, to ransomware and rootkits. These payloads determine the relative risk of a malware sample, for example a piece of adware presents little risk when compared to a rootkit.

Since 2012 the number of network-capable devices has increased from 8.7 billion to 28.4 billion and is expected to increase to 50.1 billion by 2020 [1], this poses a problem for Security Operation Centres as this increasing number of devices gives attackers an increasing number of targets and opportunities to exploit the networked nature of these devices for malicious purposes [2]. Regardless of platform, there is likely no perfect way of detecting or categorizing malicious files. Currently, anti-virus vendors and Security Operations Centers classify malware according to its behaviors on its target machine or Operating System, this is done through analyzing the malware signature and comparing to known signatures of other malware samples. As the evasion and propagation techniques used by attackers increase in complexity, we need a similar evolution in detection techniques in order for Security Operations Centres to quickly and accurately assess the potential risk associated with a file and put the necessary mitigations and remediation in place.

For this project, I will be focusing on heuristic based static analysis of the malware file structure to determine its relative risk. The goal is to investigate this method of malware categorization, which will involve data-mining a set of Windows Portable Executable files and make use of existing machine learning libraries and clustering algorithms to categorise these files.

# **Literature Review**

**Exploration of Existing Malware Detection Methods**

Most anti-virus software looks at syntactic signatures of a file to determine whether it is malicious, these signatures can be made up of combinations, or patterns, of instructions which have been previously flagged as malicious. At a higher level, anti-virus software can look at the hash of a file and compare to a database of hash values from known malware samples [3]. The disadvantage of the signature based approach most anti-virus software takes is that small changes to malware source code can result in a signature which may not have been seen before and therefore not in the most up-to-date database of signature definitions. Similarly, these signature based methods provide little protection from zero-day attacks. Metamorphic and polymorphic malware also causes issues for signature based detection systems as one piece of malware can have multiple signatures.[4]

There has been some exploration into semantic-based detection systems as a way to combat the previously mentioned methods of evading syntactic signature-based systems. These systems seek to identify *what* an executable is doing rather than *how* it is doing it through use of abstractions and templates. [4]

## **Dynamic Analysis**

Dynamic analysis involves executing the malware in a sandboxed environment, often on a virtual machine. The run-time behaviors of the malware are then observed through the tracking of system events and network traffic. This approach to malware analysis excels as it removes a large amount of the guess work which would be present using static analysis because the malware is being run in a relatively realistic environment. [5]

Dynamic Analysis has its drawbacks, there is a non-zero chance of infecting other machines on the same network. Researchers have also observed malware which behaves differently when run in a virtualized environment[6], [7], which prevents accurate categorization.

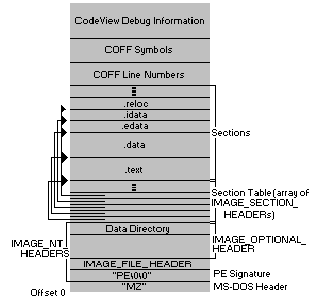
## **Static Analysis**

Static analysis is a technique used to inspect properties of a piece of malware without loading the executable into memory. It involves looking at the heuristics of the malware file for features which reveal behaviors the malware may exhibit if it were executed. The obvious advantage of this approach is that, since the malware is not being executed on the machine, then there is no chance of the malware spreading to other machines on the network and this also prevents damage to the host operating system. Static analysis also has limitations, such as its inability to fully analyze the behavior of a binary file which uses self-modifying code techniques or relies on data which is not static e.g. The current date/time. [8]. Accurate results can be computationally expensive to obtain. The processing time poses a problem for systems which rely on real-time detection of threats, although this is not the case for all static analysis systems as near real-time detection was observed with the PE Miner tool created by Shafiq et al. [9]

Static analysis has also been shown to be susceptible to reduction in effectiveness when used with obfuscated binaries [10],[11].

## **PE File Structure**

The Windows Portable Executable File Format is an extension of the Common Object File Format. There is a large amount of information available from an executable file including its target operating system, the date and time it was compiled and the DLLs it imports. A typical PE32 file consists of several sub-structures shown below.



**Fig 1.** The PE File Format [12]

Each of the components of a PE File have a clearly designed specification which can be found at [12].

This specification forms the basis for our static analysis system, allowing us to create a parser which is able to extract information and features of the file quickly and conveniently. There are structural differences between 32 bit binaries and 64 bit binaries. For this solution, only 32 bit binaries are being targeted.

## **Proposed Machine Learning Features**

### File Entropy

This is a number between 0 and 8 and represents the randomness of the data within the executable. According to the results proposed by Lyda and Hamrock [13], packed files and encrypted executables usually have a higher value for entropy when compared to non-packed and non-encrypted executables. Since a large proportion of malware use these techniques while benign files rarely use them. Entropy values can be used as an indicator of whether a file has been packed or encrypted and therefore if it is benign or malicious. This indicator alone would obviously not be good enough to distinguish between a benign or malicious with 100% accuracy but it is a useful metric nonetheless.

### Ratio of Code to Initialised Data

This is the ratio of executable code to initialised data within the PE File. During my initial research, mining data from my dataset of malicious and benign files, I observed that benign files often have a large amount of data but a small amount of executable code in comparison. Conversely I observed that malicious files tend to have little to no data but a larger proportion of executable code. To normalise these observations and make them more meaningful I have chosen to express this ratio as a single machine learning feature as opposed to two distinct features consisting of the size of code and size of initialised data.

### Major Image Version

This number is the version of the executable. According to Raman [14] malicious files often have a value of zero and benign files often have a higher value. I observed the same pattern during my initial research and as such this feature has been the best indicator of whether a file is benign or malicious, that I have come across.

### Number Of Sections

This is the number of section headers in the section header table. As Yonts[15] observed, the number of sections in an executable can be a very good indicator of whether the file is benign or malicious. In general, benign files range from 0-10 sections while malicious files almost always have 3 or 4 sections. I was also able to observe this pattern when data mining my dataset.

### Common DLL Imports

The DLLs imported by an executable paint a very good picture of the functionality and behaviour of the executable. For example, if an executable is importing Wsock32.dll then you can surmise that the executable makes use of networking. Although there will be overlap between the imports of benign files and malicious files, in theory malware should be able to be classified by family using this information, as behaviours differ between families. We will be using the same technique Shafiq et. Al. [9] proposed where we look for a particular import and set a flag to true or false depending on the whether it was present or not, in order to have a numeric representation of the data.

|  |  |
| --- | --- |
| **Import Name** | **Function** |
| KERNEL32.DLL | Memory Management & I/O operations |
| USER32.DLL | Window Management |
| ADVAPI32.dll | Security & registry Management |
| msvcrt.dll | C library for the Visual C++ Compiler |
| GDI32.dll | Graphics Device Interface |
| SHELL32.dll | Opening webpages and files |
| ole32.dll | Object Linking and Embedding |
| WS2\_32.dll | Networking |
| SHLWAPI.dll | Registry Entry, URL and Colour Management |
| COMCTL32.dll | UI Components |

**Fig 2.** List of 10 most common imports from dataset

After analyzing the data set, the 10 most common imports were found their presence or absence from a file are being used as features in the solution. The functions for the above dll files were sourced from [16]

## Dimensionality Reduction Techniques

Dimensionality reduction techniques seek to find the subset of data which best represents the whole collection of data for a particular object. That is, dimensionality reduction involves mapping a high-dimensionality dataset to a lower-dimension sub-space by filtering out data which is deemed non-essential. This has the benefit of providing an easier visualization of the whole data-set as it is not practical to represent anything with more than 3 dimensions on a graph. Low-dimensional datasets also improve performance when processing data with machine learning algorithms.[17]

### Principle Component Analysis (PCA)

Principle component Analysis is the most common form of dimensionality reduction and it achieves this though factorization of a matrix (our dataset) into two sub matrices, one of which is the Principle Component Model which best represents the data in the original matrix, this is the data which will be plotted on a graph, and the other is considered noise. [18]

### Non-Negative Matrix Factorisation (NMF)

Much like Principle Component Analysis, Non-Negative Matrix Factorisation involves factorizing a high-dimensionality matrix into two smaller matrices. Where this approach differs is that it operates under the constraint that the numerical data which makes up the data-set or matrix is non-negative. As a result of this constraint NMF can result in a different representation of the original matrix.[19]

## **Clustering Algorithms**

### KMeans

The K-Means algorithm divides provided data-points into K clusters such that the sum of squares within each cluster is minimized.[20][21] A crucial component of the KMeans algorithm is the number of clusters which the algorithm is instructed to create. Unlike other clustering algorithms such as MeanShift the KMeans algorithm does not decide the number of clusters internally based on the data.

#### Silhouette Metrics

There are several techniques which can be used to determine the optimal number of clusters, one of which is silhouette metric. This number is essentially a score for how well each data-point fits into its cluster. This metric is expressed as a score between -1 and 1. A score tending towards -1 would indicate that there may be to many or too few clusters and that the data does not belong in this cluster. A score tending towards 1 would indicate that the data has been clustered correctly. A score around 0 indicates that the data may fit more than one cluster [20].

|  |  |  |  |
| --- | --- | --- | --- |
| **Number Of Clusters** | **Decompositon** | **Clustering** | **Silhouette Score** |
| 2 | NMF | KMeans | 0.810315175 |
| 2 | PCA | KMeans | 0.704351353 |
| 3 | NMF | MeanShift | 0.660036554 |
| 3 | NMF | KMeans | 0.647568178 |
| 3 | PCA | KMeans | 0.63639525 |
| 4 | PCA | KMeans | 0.628463924 |
| 4 | NMF | KMeans | 0.626236348 |
| 4 | PCA | MeanShift | 0.617196567 |
| 5 | NMF | KMeans | 0.595266334 |
| 6 | PCA | KMeans | 0.569866616 |
| 6 | NMF | KMeans | 0.540762171 |
| 5 | PCA | KMeans | 0.510923939 |

**Fig 3.** Examples of silhouette metric scores

#### Elbow Method

The Elbow Method is an alternative technique for determining the optimal number of clusters for a given dataset. This involves calculating the cost function for a given value of K starting with k = 2 and increasing K by 1 until the cost function of K drops drastically [20]. The cost function can be defined as the sum of squared errors, that is, the sum of the squared linear distance of each point to its nearest centroid. The reason the we see the cost of K decrease as K increases is that, as the number of centroids increases, the distances from each point to its nearest centroid decreases.



**Fig 4.** Example elbow plot comparison for PCA and NMF

In both instances shown above it is obvious that the optimal number of clusters for the dataset analyzed is two, this is because the largest drop observed in the Cost Function of K (y-axis) occurs between 1 cluster and 2 clusters.

### Mean Shift

The Mean Shift algorithm is a form of hierarchical clustering algorithm. It works by iteratively shifting each data-point towards its nearest peak. These peaks are identified using a probability density function calculated from the given dataset. To start, each point is identified as a peak. We need to provide a kernel bandwidth value which will indirectly determine how many clusters will end up with. The kernel bandwidth is a radial distance from each point. If other points fall within this radial distance then the mean value for all of these points is used as a peak. This process is then repeated iteratively until all points shift towards their nearest peak. [22] This forms clusters of data-points.

# **System Requirements Specification**

## **System constraints**

As previously mentioned, this project is only concerned with WIN32 binary files and not with their 64 bit equivalents. The system also does not make any attempt to decrypt or unpack malware files. This is due to the large number of implementations of packing algorithms and encryption algorithms used by attackers. It is unrealistic to have perfect conditions for each of these implementations. The extracted feature of file entropy will help combat this as this value gives a good indication of whether a file is malicious or not. Another possible approach to this problem is to attempt to unpack or decrypt the malware using existing tools before analysis with our solution.

## **Functional Requirements**

1. Provide methods for the extraction of all features outlined in the COFF specification [12]
2. Extract chosen features from a directory of WIN32 binary files
3. Provide a percentage measure for how accurate the clustering was.
4. Display the output of clustering algorithms in 2d and 3d graphs

## **Non-functional requirements**

If this project is a success it should:

1. Be able to distinguish between benign and malicious files with greater than 95% accuracy.
2. Be able to distinguish between different malware families with greater than 80% accuracy.
3. Provide an easy to use tool for static analysis of a PE32 file. i.e small setup time and cost with minimal configuration required to run.
4. Be platform independent for increased usability.
5. Handle errors gracefully and not crash or hang
6. Be scalable to large datasets (1000+ files)
7. Data-mine a file in no more than 5 seconds on average across the dataset

## **User Interface requirements**

The user interface for this tool should:

1. Allow the user to browse to a chosen directory
2. Display this directory for reference
3. Have a button for performing analysis on the specified directory.
4. Allow the user to specify which clustering algorithms and dimensionality reduction algorithms to use
5. Show output from each of the selected algorithms
6. Show a percentage accuracy value for each of the selected algorithms

## **User Characteristics**

This tool is aimed at Security Operation Centres as well as InfoSec researchers in a professional environment.

Provides ability to compare an unknown file against a known dataset to determine its relative risk

# **Design**

## **System Architecture**

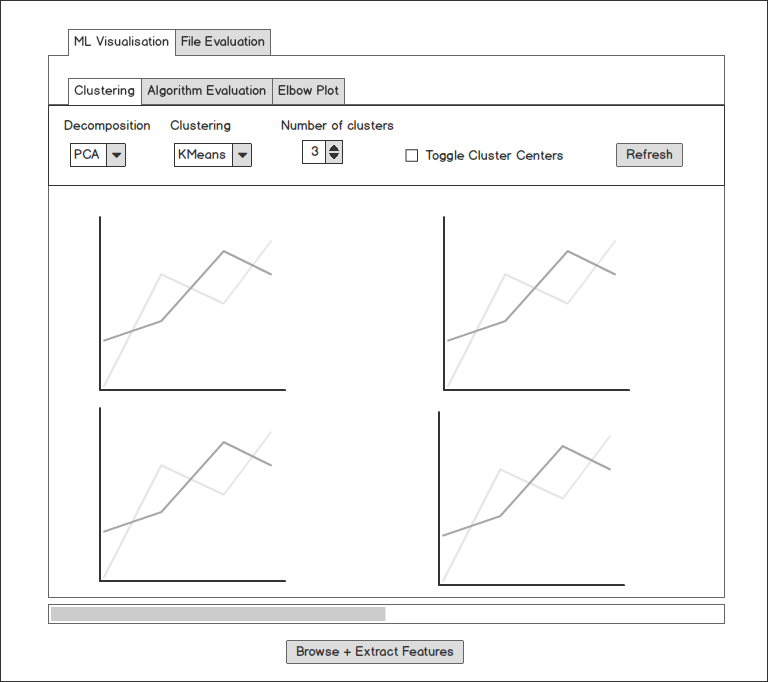


**Fig 5.** Information flow in the system

The system consists of 3 classes. The PE32 class performs IO operations to read hexadecimal byte strings from Win32 PE Files. The PEMachineLearning class takes the data exposed by the PE32 class and performs dimensionality reduction, clustering and visualization operations. This class also handles interaction logic for the UI, attaching event listeners to UI components present in the UI\_MainWindow. The outputs from the operations carried out by the PEMachineLearning class are displayed through the UI\_MainWindow class.

## **User Interface Design**

The user interface is split up into 2 main tabs and 3 sub-tabs within the first main tab. Outside of the main area of the application there is a progress bar and a button which allows the user to browse to a directory.



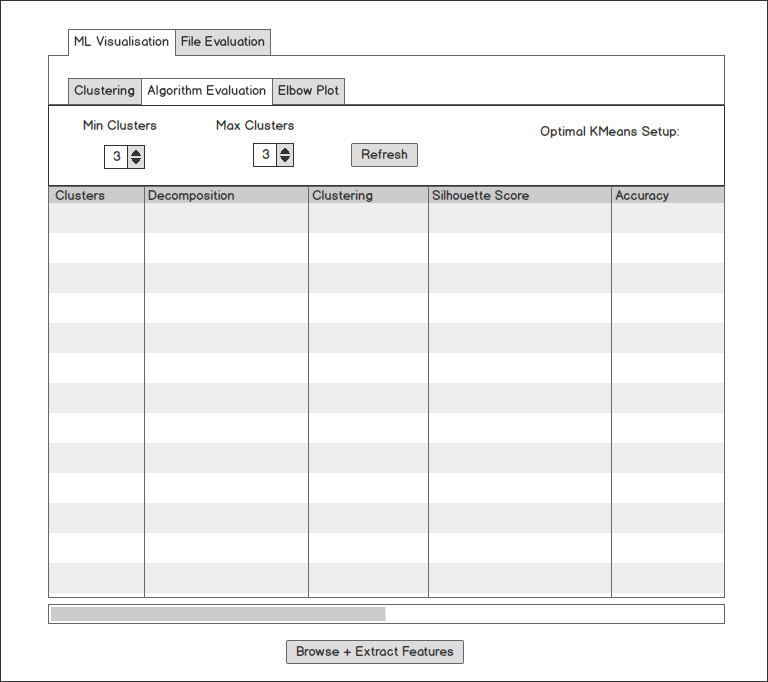
**Fig 6.** Design for the clustering tab

The first main tab is the visualization tab, this contains three sub-tabs for generating a number of different graphs and evaluating different combinations of decomposition and clustering algorithms on a dataset. The first sub-tab allows the user to choose a decomposition and clustering algorithm, as well as a number of clusters, which is only used for the KMeans algorithm. The user can toggle the display of cluster centers or centroids on the graphs plotted in this tab.

The four graphs are, starting from top-left clockwise :

1. Decomposition Plot
2. Clustering Plot
3. Color Map Plot showing linear discriminants between clusters
4. Silhouette Plot depicting how well clustered the data is

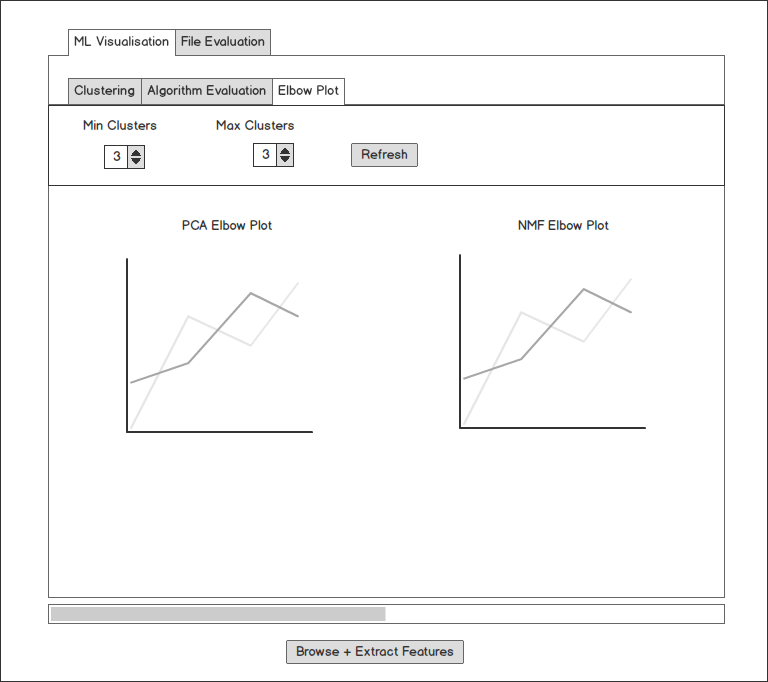
When the plotting is complete a silhouette score will be made visible below the plots.



**Fig7.** Design for the Algorithm Evaluation tab

The second sub-tab consists of controls for specifying maximum and minimum numbers of clusters to evaluate algorithms between. This tab allows the user to quickly specify constraints for automatic evaluation of all possible decomposition and clustering algorithm combinations, and species the optimal KMeans parameters which can be used on the visualization page. The evaluation is presented as a table consisting of 5 columns:

1. Number of clusters
2. Decomposition Algorithm
3. Clustering Algorithm
4. Silhouette Score
5. Detection Accuracy

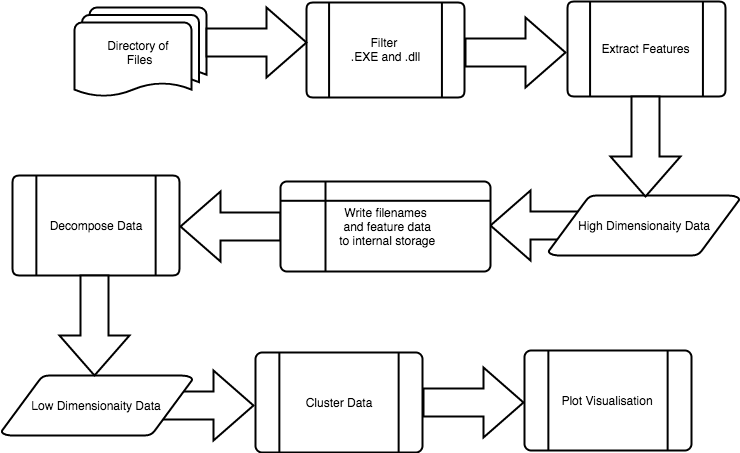


**Fig 8.** Design for elbow plot tab

The third sub-tab provides the user with elbow plots for the KMeans algorithm for both supported decomposition algorithms. This can be used as an alternative method for determining the optimum KMeans parameters for the first tab.

**FILE EVAL**

## **Software Design**



**Fig 9.** Proposed pipeline for the system

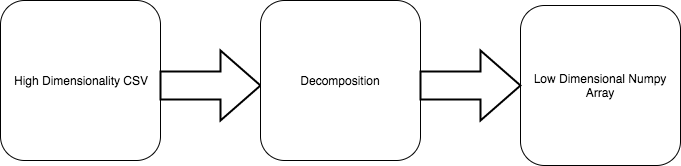
Figure 9 shows a high level version of the data-flow within the completed project. Each stage of this pipeline is explained in more detail below.

1. User browses to a directory of files. For each .EXE or .dll file present in the directory, the tool creates an instance of the PE32 class.
2. For each PE32 object, extract our chosen machine learning features to a row in a CSV file.

| **Feature** | **Source** |
| --- | --- |
| File Entropy | PE32.get\_file\_entropy() |
| Code/Data Ratio | PE32.get\_code\_data\_ratio() |
| Major Image Version | PE32.get\_major\_image\_version() |
| Number Of Sections | PE32.number\_of\_sections() |
| CommonDll Imports | PE32.get\_imports() - map this list of imports to an array of true or false values for the presence of our chosen dlls. |

**Fig 10.** Proposed features and source within the PE32 class

1. Perform decomposition on the high dimensionality data in the CSV file and project to a lower dimensional subspace suitable for plotting. Store this in a Numpy array.



**Fig 12.** Decomposition stage of the pipeline

1. Cluster data from array using chosen clustering algorithm and plot data points on a 2d Matplotlib scatter graph.



**Fig 12.** Example K=2 KMeans clustering for malicious (red) and benign(green) files.

1. Use labels assigned by clustering algorithm in conjunction with filenames to determine the accuracy of the clustering.

# **Implementation**

## **Choice of Languages and Libraries**

This project was implemented in Python 3.5. The reasoning for this is that Python is ideal for rapid prototyping; often requiring significantly less code than other object-orientated languages such as Java and C++ due to its dynamic typing. There are of course trade-offs with this choice, namely performance in comparison to C++ and other low level languages.

Due to the specialised requirements of this project, I made use of several libraries to aide in the rapid development of a working piece of software.

For the machine learning component of the project, I opted to use scikit-learn. This is an open source machine learning library for python which has implementations of common clustering algorithms such as KMeans and MeanShift. There are also several implementations for dimensionality reduction/decomposition available, namely Principle Component Analysis (PCA), Non-Negative Matrix Factorisation (NMF) and Fast Independent Component Analysis (FastICA). In my solution I make use of both PCA and NMF. To cluster the decomposed data, I made use of the KMeans and MeanShift algorithms.

I have chosen to use Matplotlib for data visualisation as it provides the ability to plot graphs with relatively little code. The data-mined features are initially be too high-dimensional to plot on a graph, so I used PCA and NMF to project the data to a lower dimensional subspace which retains most of the variance from the initial dataset [17]. I then use the output of this as the input for the KMeans algorithm and then subsequently plot the clustered data using Matplotlib.

To implement a rudimentary GUI which displays the plotted graphs and evaluation of the clustering labels, I used PyQt which is a cross platform python binding used for rapid GUI development.

## **Development Environment**

Due to the malicious nature of the Windows PE files being analysed, my host operating system for my development environment was Mac OS X. This ensured that the malware files did not execute, as OS X uses the ELF format for executables, similar to a Linux environment.

I used Spyder as my Python IDE which is packaged with the Anaconda Python data-science platform. Anaconda comes with the Pylab package which includes Numpy and Matplotlib. My reasons for choosing the Anaconda Python platform were its ease of set-up and its ability to create multiple python environments with different versions of referenced libraries. It also comes bundled with scikit-learn and PyQt which minimises the time required to set-up a full development environment.

To create my GUI, I used the tools contained within QtCreator IDE. QtCreator is a cross platform C++ IDE which comes bundled with QtDesigner. QtDesigner is a WYSIWYG form design tool, which outputs a .ui file that can be used to generate a PyQt GUI class, using the pyuic5 command line tool. This solution to designing and creating a GUI minimised time spent creating complex layouts which was be better spent elsewhere in the development of the solution. It also allows for significantly more complex layouts than those possible with built in Python GUI libraries such as TKinter.

## **Development Hardware**

**Machine:** Apple Macbook Pro

**Operation System:** OS X El Capitan Version 10.11.6

**CPU:** 2.5 GHz Intel i5

**Memory:** 8GB DDR3 1600 MHz

## **Key Implementation Decisions**

At the outset of the project it was decided that the implementation of machine learning algorithms and dimensionality reduction algorithms was out of scope of the project, subsequently, existing machine learning libraries were examined to determine their utility for what we wanted to accomplish in this project. The development efforts were largely focused in establishing a pipeline of operations which would give a useful visual representation of clusters within a given a dataset of benign and malicious files.

When the project was underway it was apparent that evaluation and comparison of multiple algorithms proved to be time consuming and that creating a tool based on a single algorithm was too rigid to be useful for any other experimentation with the dataset. As a result, elements were added to the UI to enable the tweaking of input parameters to the algorithms and also selection of which decomposition or clustering algorithms to use in the pipeline. Screenshots of these UI elements can be seen in Figure 13.

****

**Fig 13.** Output on the Clustering Tab.

To tackle the issue of algorithm evaluation I implemented a way of automatically evaluating each permutation of decomposition and clustering algorithms along with their inputs, within certain constraints set by the user similar to the tab in the previous screenshot. This tab finds the optimal setup for the KMeans algorithm based on the silhouette score, expressing how well the data fits the cluster labels generated for that particular permutation. This recommendation can be seen in the upper right-hand corner of the below screenshot. This tab builds a table of results automatically which is also written to a csv file for easy parsing in external programs such as Excel.



**Fig 14.** Output on the algorithm evaluation tab

This window allows the user to quickly evaluate which algorithm combinations produce the best clustering and, optionally, can also be used to determine how well the data has been clustered based on the detection rate of malware samples. The detection rate is the percentage of malware files it correctly identified in the dataset when compared with benign files. It should be noted that because of how the filenames and extracted features are stored separately, the naming of a file has no effect on how the clustering algorithms determine which cluster is most appropriate for a particular file. In the case of the MeanShift algorithm, fewer permutations are possible as the algorithm internally decides how many clusters there will be.

## **Implementation of Important Functions/Algorithms**

### Calculation of Entropy



**Fig 15.** Entropy Calculation Code

Calculation of entropy is done by first initializing an empty integer array of length 256, with zeroes. The CalculateEntropy method allows for the entropy of sub-sections within the file to be calculated as well as the entropy value for the entire file. This array will hold a count of how many occurrences a particular byte value has between the startOffset and startOffset+length, the index of the count in the array is also the decimal value for that byte, allowing for an efficient iteration of this array later. The byte\_occurences array is then looped over and any byte values which have not occurred are ignored. For all byte values which have occurred, we calculate the frequency and then set the entropy value equal to: previous entropy value - frequency \* the base 256 log of the frequency \* 8. This results in an entropy value between 0 and 8. A higher value here represents more random data in the file within the constraints set by the caller.

### Conversion of Byte Data to Decimal



**Fig 16.** Byte to Decimal Conversion Code

As the data required for decomposition and clustering needed to be in a numerical format, it was important to be able to convert between the hexadecimal byte strings being parsed from the PE file and their equivalent decimal representations. The simple method above was created to prevent the duplication of this conversion operation, utilizing pythons built in ability to convert between numbers in different base representations.

### Reading Values from PE Files



**Fig 17.** Code Used for Reading PE32 Binary Files

Due to the large number of times the tool needs to read values from a file, it was very important to write these functions in the most reusable way possible. The ReadByte method acts as the smallest component within the PE32 class, every call to the class to get a value will result in one or more calls to the ReadByte method. The ReadBytes method shows this in practice, where it loops until it has read the specified number of bytes, appending each to a local variable to be returned to the caller. The ReadBigEndianBytes and ReadLittleEndianBytes methods add another abstraction layer over the ReadBytes method, with the only difference in the returned result being ReadBigEndianBytes’ use of pythons slicing functionality to reverse to order of the characters in the returned byte string. It should also be noted that these two methods accept parent offsets as well as offsets which are relative to the parent offset, this was a necessary step as the Windows PE File Specification[12] does not guarantee a value can be found at an offset from the start of the file, only that it can be found in relation to a previously calculated offset, usually the COFF header, Optional header or PE header.



**Fig 18.** Example calls to Code shown in Figure 17

Figure 18 shows just how flexible the previously discussed methods for reading bytes from a PE executable are. The only differences between calls to these methods were the input parameters, which were sourced from the PE File Specification[12].

### Getting Names of Imports from a PE File



**Fig 19.** Code Used to Extract Imports from Binary Files

Reading the string representation of the import names from the PE files proved to be more complicated that initially thought. The original implementation involved using the import table to fetch the relative virtual address of a specific import and then reading from this address until a null termination byte was encountered. While this approach is the correct and recommended way to fetch import names according to the PE Specification[12], it proved to be unreliable for some of the files within the dataset. For some files, the relative virtual address was outside the range of the file length causing a null value to be returned and the system to hang. The solution to this problem was to make use of the built in ‘strings’ command present on Unix based systems. This command returns all readable strings in a given binary file. A similar command line tool exists on windows called *strings.exe*. The subprocess module was used to call the strings command from within the system and return the output as a list of strings separated by line break characters. The next issue faced was filtering out all the strings which were not imports, this was relatively easy to accomplish using python’s regex library, *re.* Using a simple find all in the returned list of strings with a regular expression which would look for one or more word characters, that is, any non-special characters, followed by .dll, .exe, .DLL, or .EXE. This proved to be very effective and reliable method of fetching the import names. Figure 19 shows the complete method, found in the PE32 class.

### Automation of Optimizing KMeans Parameters



**Fig 20.** Partial Code: Automatic Algorithm Evaluaton

As previously shown in Figure !!!!, the finished tool is able to provide a recommendation of the best KMeans parameters. The implementation of this functionality is relatively simple. A small part of the *populateTable()* method from the peMachineLearning class can be seen in Figure !!!. This code simply stores the best silhouette score and the decomposition algorithm, used to pre-process the data before performing the KMeans clustering, as two variables which are then used to updated the label shown in the top right corner of Figure !!!. This logic is carried out as the different permutations algorithms used are iteratively evaluated for the purposes of populating the table in the Algorithm Evaluation tab shown in Figure !!!

### Calculation of Detection Accuracy

****

**Fig 21.** Code Which Calculates Detection Accuracy

The calculate\_detection\_accuracy method takes a list of labels and a list of filenames as parameters in order to determine if the files were labeled correctly. The first step in calculating the accuracy is establishing which label should be used as a baseline for the comparison. For this reason, the method checks the first 3 letters of the first filename that is benign and sets the variables used for comparison accordingly. Once it has established which label belongs to the benign cluster, it loops through all of the filenames and keeps a count of how many match the expected label and how many which don’t match and were not expected to match based off the first three letters in the filename. This count is then returned as a percentage of the total number of files.

### Calculation of Clustering Accuracy

### Decomposition and Clustering Plots



**Fig 22.** Code Used to Perform PCA and Plot Result on Scatter Graph

Figure !!! shows the implementation for the PCA decomposition plot. The first step is specifying how many dimensions we want to reduce the data to, this can be seen in the call to the constructor with n\_components =2. The next step is performing the dimensionality reduction on the dataset with the call to the PCA.fit() and PCA.transform() methods. This data is then plotted on a simple scatter graph with the x and y values being the two-dimensional data returned by the PCA.transform() method call. It should be noted that the plot for NMF is created the same way. The plots for the clustering algorithms used follow a similar pattern, but use the data returned by the decomposition algorithms as an input rather than the initial dataset. Figure !!! shows how colours are assigned to clusters on the plot based on the value of their label. This list of colours is then specified when plotting the points on a scatter graph.



**Fig 23.** Partial Code: Plotting Clusters and Centroid Labels

In Figure !!! you can see the logic for plotting cluster centers or centroids, this boolean flag is set by the user. When this is set to true, white circles are plotted over the points on the grid which are cluster centers, then the corresponding label values are plotted over the top of those white circles to give the output shown in Figure !!!.



**Fig 24.** Cluster Plot with Centroid labels

### Colour Plots



**Fig 25.** PartialCode: Plotting a Colour Map based on Cluster Labels

Color plots proved more complex to plot in comparison to the relatively simple scatter graphs discussed in the previous section. The first step when plotting the graph was to establish upper and lower bounds for both dimensions of the dataset. These were set to +- 1 of the actual value to prevent points being plotted on the absolute edge of the graph. A step size is then assigned which represents the granularity or precision of the linear discriminant lines plotted later. We then use this step size to generate lists of values for both dimensions between the upper and lower bounds with sequential values being incremented by the step size. These two lists of values are then used to create a grid of points. At this point we treat each point on the grid as if it was a piece of data we want to apply the clustering algorithm to. This allows us to obtain labels for each of the points on the graph and therefore know where the boundaries for each cluster are on the grid. We use this information in the call to the imshow() method and specify a color map to draw solid colours on the graph. As a last step we then plot the original two-dimensional data on top of the solid colours. This plot can be seen in Figure !!!, below.



**Fig 26.** Example: Colour Map with 2 clusters

### Silhouette Plots



**Fig 27.** PartialCode: Plotting Silhouette Scores for the Clustered Dataset

Silhouette plots were implemented using scikit learns silhouette metric methods. We first calculate silhouette scores for each data-point in the dataset. We then set a value for the gap at the bottom of the graph, y\_lower. The next step is to iterate from 0 to the number of clusters specified by the user. Next, silhouette scores for the data-points in the current cluster are sorted. We then calculate how big the cluster is and set an upper boundary on the y-axis for plotting. We assign a color based on the current cluster number and fill the area between 0 on the x-axis and the silhouette score for each data-point. On the y-axis the fill is limited by y\_lower and y\_upper. Once this is done, we place a label beside the filled in area with the cluster number and then set the lower boundary for the next filled in area to be 5 more than the upper boundary of the current area, creating a gap between the filled areas. This gives the output shown in Figure !!!, below.



**Fig 28.** Example: Silhouette Plot for 2 Clusters

### Elbow Plot



**Fig 29.** PartialCode: Elbow Plot

## Implementation of Individual Components

### PE File Parser

The PE file parser component of this project has been implemented as a single wrapper class, PE32, which exposes properties of the PE File it is instantiated with. A full UML diagram of this class can be found in Appendix 2. For outputting these properties of a file I have also created a command line tool which, when given a path to an executable file, will print all fields which the wrapper class exposes for examination. The following command will run this tool ***python peparse.py {path-to-file}/{file-name}.exe****.* Appendix 3 contains a sample output from this tool.

### Feature Extraction

Feature extraction is accomplished by instantiating the PE32 class with all acceptable files in a given directory. The tool will only ever attempt to extract features from .EXE and .dll files as these are the only files which it was designed for. Once the PE32 class has been instantiated, calls are made to the methods on the PE32 class to get the values from the file. These values are then stored in a csv file, features.csv; The associated filenames are stored in a separate csv file, files.csv. The justification for storing the filenames separately, is that in later stages of the pipeline the finished tool uses, it is imperative that the data being input into the algorithms is numerical. When the filenames are required we only read from the files.csv and when we need data for our machine learning algorithms we only read from features.csv, it should be noted that the indexes are consistent between the files so the first piece of data in one file will correlate with the first line of data in the other etc.

### User Interface

The user interface for the finished tool is implemented in two classes, UI\_MainWindow and PeMachineLearning. The UI\_MainWindow acts as a superclass for PeMachineLearning, with PeMachineLearning inheriting all properties of the superclass. UI\_MainWindow acts as the setup for the UI and cannot be instantiated without being inherited, it handles the layout, naming and most of the display details of the UI components used. The PeMachineLearning class handles more specialized operations such as plotting graphs and performing IO operations, it uses its relationship with the superclass to display the results of these operations on the UI.

### Clustering and Categorisation

Clustering and categorisation are handled by the PeMachineLearning class. The methods have been designed so that they are flexible and reusable, accepting parameters for tweaking their behaviors and outputs, where appropriate.

**Expand**

**File Evaluation**

# **System Evaluation and Experimental Results**

The results discussed in the following sections were taken from the ‘Algorithm Evaluation’ tab of the finished system.

## Evaluation of Malware Detection

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Decomposition** | **No. Features** | **Clustering** | **No. Clusters** | **Detection Accuracy** |
| NMF | 3 | KMEANS | 11 | 100.000 |
| NMF | 10 | KMEANS | 10 | 99.281 |
| PCA | 10 | KMEANS | 12 | 97.842 |
| NMF | 14 | KMEANS | 5 | 96.403 |
| PCA | 5 | KMEANS | 8 | 93.525 |
| NMF | 11 | KMEANS | 7 | 93.525 |
| NMF | 16 | KMEANS | 6 | 93.525 |
| PCA | 16 | KMEANS | 4 | 93.525 |
| PCA | 6 | KMEANS | 3 | 93.525 |
| PCA | 9 | KMEANS | 8 | 92.806 |
| PCA | 8 | KMEANS | 7 | 92.806 |
| PCA | 16 | KMEANS | 7 | 92.806 |
| PCA | 14 | KMEANS | 6 | 92.806 |
| PCA | 15 | KMEANS | 6 | 92.806 |
| PCA | 11 | KMEANS | 9 | 92.806 |

**Fig 30.** Results: Top 15 Malware Detection Accuracies

While examining the accuracy of malware detection, that is, classifying all files as either benign or malicious while ignoring differences between malware families. A detection accuracy of 100% was achieved using Non-negative matrix factorization with 3 features for the dimensionality reduction stage of the pipe-line and the KMeans clustering algorithm with 11 clusters for the classification stage of the pipeline. It should be noted that, the logic behind using more than just 2 clusters to represent benign or malicious was based off the idea that if a file fits better into a cluster other than the benign cluster then there is a non-zero risk associated with the file. The best result using only two clusters was significantly lower at only 6.47%, using the same inputs from the dimensionality reduction stage.

### Effect of Decomposition Algorithm Choice on Detection Rate

|  |  |
| --- | --- |
| **Decomposition Algorithm** | **Average Clustering Accuracy** |
| NMF | 55.759 |
| PCA | 52.385 |

**Fig 31.** Results: Decomposition Algorithm Average Malware Detection Accuracy

For the purposes of malware detection, Non-Negative Matrix factorization proved to be the superior decomposition algorithm, averaging approximately 55.75% correct detection rate, in comparison with the 52.38% averaged by using Principle Component Analysis.

### Effect of Decomposition Number of Features on Detection Rate

|  |  |
| --- | --- |
| **No. Features** | **Av. Detection Accuracy** |
| 1 | 40.408 |
| 2 | 51.888 |
| 3 | 49.940 |
| 4 | 49.640 |
| 5 | 47.602 |
| 6 | 56.924 |
| 7 | 51.499 |
| 8 | 59.802 |
| 9 | 54.736 |
| 10 | 52.488 |
| 11 | 58.843 |
| 12 | 53.927 |
| 13 | 55.695 |
| 14 | 63.549 |
| 15 | 53.177 |
| 16 | 58.633 |
| 17 | 58.933 |
| 18 | 55.606 |

**Fig 32.** Results: Number of Features vs Average Malware Detection Accuracy

In general, there was a weak positive correlation between the number of components used from the data-mined features. Initially, there was a considerable increase in detection accuracy as the number of components increased but it is not a clear cut trend as this eventually levelled off.

### Effect of Clustering Algorithm Choice on Detection Rate

|  |  |
| --- | --- |
| **Clustering Algorithm** | **Average Clustering Accuracy** |
| KMEANS | 57.365 |
| MEANSHIFT | 17.846 |

**Fig 33.** Results: Clustering Algorithm Average Malware Detection Accuracy

In terms of clustering algorithm choice, the KMeans algorithm proved undeniably superior to the MeanShift algorithm, with the algorithms averaging approximately 57.36% and 17.84% correct detection rate, respectively.

### Effect of Number of Clusters on Detection Rate

|  |  |
| --- | --- |
| **No. Clusters** | **Av. Detection Accuracy** |
| 2 | 6.475 |
| 3 | 27.758 |
| 4 | 33.653 |
| 5 | 54.229 |
| 6 | 48.351 |
| 7 | 62.406 |
| 8 | 67.758 |
| 9 | 69.345 |
| 10 | 74.341 |
| 11 | 76.699 |
| 12 | 79.197 |

**Fig 34.** Results: Number of Clusters vs Average Malware Detection Accuracy

A strong positive correlation between number of clusters and clustering accuracy was observed when testing detection rates. This could be partially explained by the logic used to calculate the detection accuracy, where, if there are more clusters there is a higher chance of a file being labeled as not in the benign cluster. Although this is correct, the inherent higher chance can be discounted as the KMeans algorithm will label data according to which cluster it fits best into. That is, even if we are using more clusters, if a file is labeled as being in any cluster other than the benign cluster, then the data shows more similarities to malicious files than it does benign files.

## Evaluation of Malware Family Categorisation

| **Decomposition** | **No. Features** | **Clustering** | **No. Clusters** | **Clustering Accuracy** |
| --- | --- | --- | --- | --- |
| NMF | 16 | KMEANS | 9 | 46.763 |
| NMF | 16 | KMEANS | 8 | 46.043 |
| NMF | 15 | KMEANS | 9 | 43.165 |
| NMF | 14 | KMEANS | 12 | 41.727 |
| NMF | 14 | KMEANS | 11 | 41.727 |
| PCA | 5 | KMEANS | 8 | 40.288 |
| PCA | 9 | KMEANS | 8 | 40.288 |
| PCA | 9 | KMEANS | 7 | 40.288 |
| PCA | 8 | KMEANS | 8 | 40.288 |
| NMF | 15 | KMEANS | 6 | 40.288 |
| PCA | 10 | KMEANS | 8 | 40.288 |
| PCA | 14 | KMEANS | 8 | 40.288 |
| NMF | 16 | KMEANS | 10 | 40.288 |
| PCA | 7 | KMEANS | 7 | 40.288 |
| PCA | 7 | KMEANS | 8 | 40.288 |

**Fig 35.** Results: Top 15 Malware Clustering Accuracies

When measuring the ability of the system to categorise malware into families, the highest accuracy achieved was only 46.76%. With 8 families of malware in our dataset and a number of benign files for comparison, we observed that the KMeans algorithm with 9 clusters proved to be the best fit. For dimensionality reduction the NMF algorithm was the best performer when used with 16 features from the data-mined files.

### Effect of Decomposition Algorithm Choice on Clustering Accuracy

|  |  |
| --- | --- |
| **Decomposition Algorithm** | **Average Clustering Accuracy** |
| NMF | 19.644 |
| PCA | 22.192 |

**Fig 36.** Results: Decomposition Algorithm Average Malware Clustering Accuracy

When measuring clustering accuracy, Principle Component Analysis proved marginally better for dimensionality reduction in comparison to Non-Negative Matrix Factorisation, with each of the algorithms averaging approximately 19.64% and 22.19% accuracy, respectively.

### Effect of Number of Features on Clustering Accuracy

|  |  |
| --- | --- |
| **No. Features** | **Av. Clustering Accuracy** |
| 1 | 8.933 |
| 2 | 15.168 |
| 3 | 21.703 |
| 4 | 22.362 |
| 5 | 23.861 |
| 6 | 22.602 |
| 7 | 22.722 |
| 8 | 22.542 |
| 9 | 22.332 |
| 10 | 21.343 |
| 11 | 21.043 |
| 12 | 21.283 |
| 13 | 21.733 |
| 14 | 21.523 |
| 15 | 23.321 |
| 16 | 25.210 |
| 17 | 19.814 |
| 18 | 19.035 |

**Fig 37.** Results: Number of Features vs Average Malware Clustering Accuracy

The number of components used in the pipeline showed positive correlation with clustering accuracy from 1 – 5 components, but quickly levelled off after this point. While the number of components doesn’t seem to have an insignificant effect on clustering accuracy, based on the above results it’s fair to say that is not the main source of accuracy in the system.

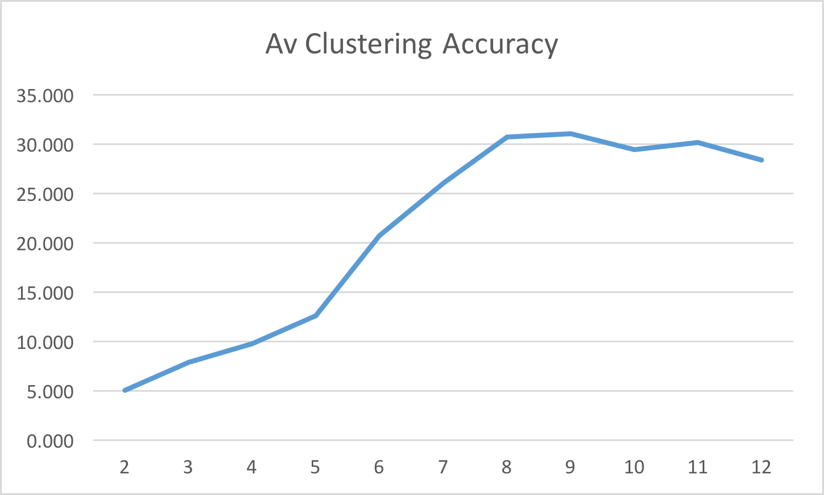
### Effect of Clustering Algorithm Choice on Clustering Accuracy

|  |  |
| --- | --- |
| **Clustering Algorithm** | **Average Clustering Accuracy** |
| KMEANS | 22.048 |
| MEANSHIFT | 8.493 |

**Fig 38.** Results: Clustering Algorithm Average Malware Clustering Accuracy

As was the case with detection accuracy, a large difference in clustering accuracy was observed between our two chosen algorithms, with KMeans averaging more than 2.5 times more accurate than MeanShift.

### Effect of Number of Clusters on Clustering Accuracy

****

|  |  |
| --- | --- |
| **No. Clusters** | **Av Clustering Accuracy** |
| 2 | 5.036 |
| 3 | 7.854 |
| 4 | 9.736 |
| 5 | 12.580 |
| 6 | 20.687 |
| 7 | 26.067 |
| 8 | 30.689 |
| 9 | 31.055 |
| 10 | 29.416 |
| 11 | 30.176 |
| 12 | 28.397 |

**Fig 39.** Results: Number of Clusters vs Average Malware Clustering Accuracy

Similar to the trend we observed in detection accuracy and number of clusters, in terms of categorizing malware into families, there is a positive correlation between the number of clusters and clustering accuracy. In this case we see the accuracy level off after 9 clusters.

### Effectiveness of the System

At the outset of the project, visualization was an important part of the proposed pipeline, while the tool can be used to visualize 2-dimensional data, the best results observed for both clustering and detection accuracy were obtained using greater than 2 features from the data-set. While the visualization functionality is still useful for experimentation, with the particular data-set analysed in this project, visualization was not practical. The visualization functionality is still

In terms of processing time when data-mining files, the system handles all but the biggest PE files in our data-set in less than 2 seconds, while the initial target was less than 5 seconds.

The ability of the system to automatically perform all permutations of the available algorithms and their inputs proved to be very effective, allowing rapid iterations with different feature sets.

The system has proven to be very effective at categorizing malicious and benign files for the purposes of malware detection, achieving 100% accuracy with the NMF and KMeans algorithms. Exceeding the initial target of 95%.

When using the system for categorizing malware into families it was less effective that originally hoped, achieving only 46.76% accuracy using the NMF and KMeans algorithms. This accuracy rate was considerably lower than the initial target of 80%.

# Conclusion

In this project, it was proposed that; using heuristics data-mined from WIN32 Portable Executable Files, malware could be effectively categorized. As part of this experimentation, a tool was created which can automatically run every possible permutation of the available decomposition algorithms and clustering algorithms and present results to the user. The tool also allows the user to visualize 2-dimensional data in a number of different formats.

Reflecting on the performance of the current system, we have proved that the heuristics of Portable Executable files are highly effective at categorizing a file as malicious or benign, achieving a 100% detection accuracy when the system is used with the NMF decomposition algorithm and 3 features as an input to the KMeans clustering algorithm with 11 clusters. In regards to categorizing malicious files into families, the system has proven to be mostly ineffective, achieving only 46.76% accuracy when clustering. It is possible that, with further research and experimentation, this accuracy could be improved, but at the current time the proposed system is not accurate enough for professional use.

# **References**

[1] Statista, “IoT: number of connected devices worldwide 2012-2020 | Statista,” 2015. [Online]. Available: https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/. [Accessed: 14-Apr-2017].

[2] D. Peraković, M. Periša, and I. Cvitić, “Analysis of the IoT impact on volume of DDoS attacks,” 2015.

[3] M. Sikorski and A. Honig, *Practical Malware Analysis*. San Francisco: No Starch Press, 2012.

[4] M. Christodorescu, S. Jha, S. A. Seshia, D. Song, and R. E. Bryant, “Semantics-aware malware detection,” in *2005 IEEE SYMPOSIUM ON SECURITY AND PRIVACY, PROCEEDINGS*, 2005, pp. 32–46.

[5] M. Egele, T. Scholte, E. Kirda, and S. Barbara, “A survey on automated dynamic malware analysis techniques and tools,” *ACM Comput. Surv.*, vol. V, no. 2, pp. 1–49, 2011.

[6] B. Lau and V. Svajcer, “Measuring virtual machine detection in malware using DSD tracer,” *J Comput Virol*, vol. 6, pp. 181–195, 2010.

[7] T. Petsas, G. Voyatzis, E. Athanasopoulos, M. Polychronakis, and S. Ioannidis, “Rage Against the Virtual Machine: Hindering Dynamic Analysis of Android Malware,” in *Proceedings of the Seventh European Workshop on System Security*, 2014, p. 5:1--5:6.

[8] S. Gadhiya, K. Bhavsar, and P. D. Student, “Techniques for Malware Analysis,” *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 3, no. 4, pp. 2277–128, 2013.

[9] M. Zubair Shafiq, S. Momina Tabish, F. Mirza, and M. Farooq, “‘PE-Miner: Mining Structural Information to Detect Malicious Executables in Realtime’ in Recent Advances in Intrusion Detection,” Springer Science + Business Media, 2009, pp. 121–141.

[10] M. Christodorescu and S. Jha, “Static analysis of executables to detect malicious patterns,” *SSYM’03 Proc. 12th Conf. USENIX Secur. Symp.*, vol. 12, pp. 12–12, 2003.

[11] A. Moser, C. Kruegel, and E. Kirda, “Limits of Static Analysis for Malware Detections,” *Acsac*, pp. 421–430, 2007.

[12] “Microsoft Portable Executable and Common Object File Format Specification,” 2015. [Online]. Available: http://download.microsoft.com/download/9/c/5/9c5b2167-8017-4bae-9fde-d599bac8184a/pecoff\_v83.docx.

[13] R. Lyda and J. Hamrock, “Using Entropy Analysis to Find Encrypted and Packed Malware,” *IEEE Secur. Priv.*, vol. 5, pp. 40–45, 2007.

[14] K. Raman, “Selecting Features to Classify Malware,” *InfoSec Southwest 2012*, pp. 1–5, 2012.

[15] J. Yonts, “Attributes of Malicious Files,” 2012. [Online]. Available: https://uk.sans.org/reading-room/whitepapers/malicious/attributes-malicious-files-33979. [Accessed: 08-Dec-2016].

[16] “ProcessLibrary.com - The Online Resource For Process Information!” [Online]. Available: http://www.processlibrary.com/en/. [Accessed: 24-Apr-2017].

[17] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press, 2012.

[18] S. Wold, K. H. Esbensen, and P. Geladi, “Principal Component Analysis,” *Chemom. Intell. Lab. Syst.*, vol. 7439, no. August, pp. 37–52, 1987.

[19] D. Lee and H. Seung, “Algorithms for non-negative matrix factorization,” *Adv. Neural Inf. Process. Syst.*, no. 1, pp. 556–562, 2001.

[20] T. M. Kodinariya and P. R. Makwana, “Review on determining number of Cluster in K-Means Clustering,” *Int. J. Adv. Res. Comput. Sci. Manag. Stud.*, vol. 1, no. 6, pp. 2321–7782, 2013.

[21] J. A. Hartigan and M. A. Wong, “Algorithm AS 136: A K-Means Clustering Algorithm,” *Source J. R. Stat. Soc. Ser. C (Applied Stat.*, vol. 28, no. 1, pp. 100–108, 1979.

[22] Y. Cheng, “Mean Shift, Mode Seeking, and Clustering,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 8, 1995.

# **Appendices**

## Appendix 1. Meeting Minutes





## Appendix 2. PE32 Class Full UML



## Appendix 3. peparse.py Sample Command Line Output

PE HEADER FEATURE EXTRACTION:

FILENAME: /Binaries/1.dll

FileSize: 163840 bytes

CodeDataRatio: 0.02631578947368421

\*\*\*FILE HEADER\*\*\*

Machine: 014c

NumberOfSections: 4

TimeDateStamp: 4d0e2fe6

PointerToSymbolTable: 00000000

NumberOfSymbols: 00000000

SizeOfOptionalHeader: 00e0

Characteristics: 210e

\*\*\*OPTIONAL HEADER STANDARD FIELDS\*\*\*

Magic: 010b

MajorLinkerVersion: 06

MinorLinkerVersion: 00

SizeOfCode: 00001000

SizeOfInitializedData: 00026000

SizeOfUninitializedData: 00000000

AddressOfEntryPoint: 000012fa

BaseOfCode: 00001000

BaseOfData: 00002000

\*\*\*OPTIONAL HEADER WINDOWS SPECIFIC\*\*\*

ImageBase: 10000000

SectionAlignment: 00001000

FileAlignment: 00001000

MajorOperatingSystemVersion: 0004

MinorOperatingSystemVersion: 0000

MajorImageVersion: 0000

MinorImageVersion: 0000

MajorSubsystemVersion: 0004

MinorSubsystemVersion: 0000

Win32VersionValue: 00000000

SizeOfImage: 00028000

SizeOfHeaders: 00001000

CheckSum: 00000000

Subsystem: 0002

DllCharacteristics: 0000

SizeOfStackReserve: 00100000

SizeOfStackCommit: 00001000

SizeOfHeapReserve: 00100000

SizeOfHeapCommit: 00001000

LoaderFlags: 00000000

NumberOfRvaAndSizes: 00000010

\*\*\*OPTIONAL HEADER DATA DICTIONARIES\*\*\*

ExportTableRVA: 000021b0

ExportTableSize: 00023e16

ImportTableRVA: 8284

ImportTableSize: 00000050

ResourceTableRVA: 00000000

ResourceTableSize: 00000000

ExceptionTableRVA: 00000000

ExceptionTableSize: 00000000

CertificateTableRVA: 00000000

CertificateTableSize: 00000000

BaseRelocationTableRVA: 00027000

BaseRelocationTableSize: 00000064

DebugRVA: 00000000

DebugSize: 00000000

ArchitectureRVA: 00000000

ArchitectureSize: 00000000

GlobalPtrRVA: 00000000

GlobalPtrSize: 00000000

TLSTableRVA: 00000000

TLSTableSize: 00000000

LoadConfigTableRVA: 00000000

LoadConfigTableSize: 00000000

BoundImportRVA: 00000000

BoundImportSize: 00000000

IATRVA: 00002000

IATSize: 0000005c

DelayImportDescriptorRVA: 00000000

DelayImportDescriptorSize: 00000000

CLRRuntimeHeaderRVA: 00000000

CLRRuntimeHeaderSize: 00000000

\*\*\*SECTION HEADERS\*\*\*

SECTION 0

Section 0 Name: .text

Section 0 VirtualSize: 0000039e

Section 0 VirtualAddress: 00001000

Section 0 SizeOfRawData: 00001000

Section 0 PointerToRawData: 00001000

Section 0 PointerToRelocations: 00000000

Section 0 PointerToLinenumbers: 00000000

Section 0 NumberOfRelocations: 0000

Section 0 NumberOfLinenumbers: 0000

Section 0 Characteristics: 60000020

Section 0 Entropy: 0.2375355877988614

Section 0 Md5: df7870bf6248d38832a649bc47dc28b1

SECTION 1

Section 1 Name: .rdata

Section 1 VirtualSize: 00023fc6

Section 1 VirtualAddress: 00002000

Section 1 SizeOfRawData: 00024000

Section 1 PointerToRawData: 00002000

Section 1 PointerToRelocations: 00000000

Section 1 PointerToLinenumbers: 00000000

Section 1 NumberOfRelocations: 0000

Section 1 NumberOfLinenumbers: 0000

Section 1 Characteristics: 40000040

Section 1 Entropy: 0.0032470259978673757

Section 1 Md5: 79bafdde51876f3e85f20d0596df7504

SECTION 2

Section 2 Name: .data

Section 2 VirtualSize: 0000006c

Section 2 VirtualAddress: 00026000

Section 2 SizeOfRawData: 00001000

Section 2 PointerToRawData: 00026000

Section 2 PointerToRelocations: 00000000

Section 2 PointerToLinenumbers: 00000000

Section 2 NumberOfRelocations: 0000

Section 2 NumberOfLinenumbers: 0000

Section 2 Characteristics: c0000040

Section 2 Entropy: 0.013294945794440415

Section 2 Md5: a505ae879dbdea80cb9c759f2cf47010

SECTION 3

Section 3 Name: .reloc

Section 3 VirtualSize: 00000204

Section 3 VirtualAddress: 00027000

Section 3 SizeOfRawData: 00001000

Section 3 PointerToRawData: 00027000

Section 3 PointerToRelocations: 00000000

Section 3 PointerToLinenumbers: 00000000

Section 3 NumberOfRelocations: 0000

Section 3 NumberOfLinenumbers: 0000

Section 3 Characteristics: 42000040

Section 3 Entropy: 0.03251353209648683

Section 3 Md5: 887c282d5df814c8d3f28966548c4ff0

\*\*\*IMPORTS\*\*\*

IMPORTING: KERNEL32.dll

IMPORTING: WS2\_32.dll

IMPORTING: MSVCRT.dll

## Appendix 4 Full Table of results for Malware Detection Accuracy

| **Decomposition** | **No. Features** | **Clustering** | **No. Clusters** | **Detection Accuracy** |
| --- | --- | --- | --- | --- |
| NMF | 3 | KMEANS | 11 | 100.000 |
| NMF | 10 | KMEANS | 10 | 99.281 |
| PCA | 10 | KMEANS | 12 | 97.842 |
| NMF | 14 | KMEANS | 5 | 96.403 |
| PCA | 5 | KMEANS | 8 | 93.525 |
| NMF | 11 | KMEANS | 7 | 93.525 |
| NMF | 16 | KMEANS | 6 | 93.525 |
| PCA | 16 | KMEANS | 4 | 93.525 |
| PCA | 6 | KMEANS | 3 | 93.525 |
| PCA | 9 | KMEANS | 8 | 92.806 |
| PCA | 8 | KMEANS | 7 | 92.806 |
| PCA | 16 | KMEANS | 7 | 92.806 |
| PCA | 14 | KMEANS | 6 | 92.806 |
| PCA | 15 | KMEANS | 6 | 92.806 |
| PCA | 11 | KMEANS | 9 | 92.806 |
| PCA | 12 | KMEANS | 9 | 92.806 |
| NMF | 17 | KMEANS | 10 | 92.806 |
| PCA | 3 | KMEANS | 7 | 92.086 |
| PCA | 13 | KMEANS | 7 | 92.086 |
| PCA | 7 | KMEANS | 6 | 92.086 |
| PCA | 11 | KMEANS | 6 | 92.086 |
| PCA | 4 | KMEANS | 7 | 92.086 |
| PCA | 14 | KMEANS | 9 | 92.086 |
| PCA | 12 | KMEANS | 11 | 92.086 |
| PCA | 15 | KMEANS | 12 | 92.086 |
| NMF | 17 | KMEANS | 5 | 92.086 |
| PCA | 13 | KMEANS | 8 | 91.367 |
| NMF | 16 | KMEANS | 11 | 91.367 |
| NMF | 18 | KMEANS | 11 | 91.367 |
| PCA | 9 | KMEANS | 7 | 90.647 |
| NMF | 9 | KMEANS | 10 | 90.647 |
| PCA | 18 | KMEANS | 6 | 90.647 |
| PCA | 12 | KMEANS | 10 | 90.647 |
| NMF | 12 | KMEANS | 8 | 90.647 |
| NMF | 8 | KMEANS | 5 | 90.647 |
| PCA | 1 | KMEANS | 3 | 90.647 |
| NMF | 6 | KMEANS | 3 | 90.647 |
| PCA | 14 | KMEANS | 3 | 90.647 |
| PCA | 17 | KMEANS | 12 | 89.928 |
| NMF | 17 | KMEANS | 11 | 89.928 |
| NMF | 14 | KMEANS | 12 | 89.209 |
| NMF | 16 | KMEANS | 7 | 89.209 |
| NMF | 12 | KMEANS | 10 | 89.209 |
| NMF | 8 | KMEANS | 7 | 89.209 |
| NMF | 14 | KMEANS | 8 | 89.209 |
| NMF | 2 | KMEANS | 12 | 89.209 |
| PCA | 9 | KMEANS | 11 | 88.489 |
| PCA | 6 | KMEANS | 8 | 88.489 |
| PCA | 12 | KMEANS | 12 | 88.489 |
| PCA | 2 | KMEANS | 9 | 88.489 |
| NMF | 7 | KMEANS | 4 | 88.489 |
| PCA | 2 | KMEANS | 7 | 88.489 |
| NMF | 1 | KMEANS | 7 | 88.489 |
| NMF | 4 | KMEANS | 12 | 87.770 |
| PCA | 16 | KMEANS | 11 | 87.770 |
| NMF | 11 | KMEANS | 9 | 87.770 |
| NMF | 10 | KMEANS | 11 | 87.770 |
| NMF | 18 | KMEANS | 10 | 87.770 |
| NMF | 11 | KMEANS | 8 | 87.770 |
| NMF | 17 | KMEANS | 9 | 87.770 |
| NMF | 2 | KMEANS | 8 | 87.770 |
| NMF | 8 | KMEANS | 4 | 87.770 |
| NMF | 8 | KMEANS | 6 | 87.770 |
| NMF | 18 | KMEANS | 6 | 87.770 |
| PCA | 8 | KMEANS | 8 | 87.050 |
| PCA | 3 | KMEANS | 11 | 87.050 |
| PCA | 14 | KMEANS | 10 | 87.050 |
| PCA | 15 | KMEANS | 10 | 87.050 |
| PCA | 10 | KMEANS | 10 | 87.050 |
| NMF | 14 | KMEANS | 7 | 87.050 |
| PCA | 6 | KMEANS | 9 | 86.331 |
| PCA | 18 | KMEANS | 9 | 86.331 |
| NMF | 4 | KMEANS | 8 | 86.331 |
| PCA | 7 | KMEANS | 10 | 86.331 |
| PCA | 7 | KMEANS | 11 | 86.331 |
| NMF | 5 | KMEANS | 11 | 86.331 |
| NMF | 6 | KMEANS | 12 | 86.331 |
| NMF | 8 | KMEANS | 12 | 86.331 |
| PCA | 6 | KMEANS | 6 | 85.612 |
| NMF | 6 | KMEANS | 8 | 85.612 |
| NMF | 5 | KMEANS | 8 | 85.612 |
| NMF | 4 | KMEANS | 7 | 85.612 |
| NMF | 3 | KMEANS | 12 | 85.612 |
| NMF | 7 | KMEANS | 10 | 85.612 |
| NMF | 10 | KMEANS | 12 | 85.612 |
| NMF | 12 | KMEANS | 9 | 85.612 |
| PCA | 6 | KMEANS | 11 | 85.612 |
| PCA | 2 | KMEANS | 5 | 85.612 |
| PCA | 3 | KMEANS | 5 | 85.612 |
| PCA | 1 | KMEANS | 4 | 85.612 |
| PCA | 18 | KMEANS | 4 | 85.612 |
| NMF | 8 | KMEANS | 9 | 84.892 |
| NMF | 12 | KMEANS | 11 | 84.892 |
| PCA | 7 | KMEANS | 12 | 84.892 |
| PCA | 10 | KMEANS | 11 | 84.892 |
| PCA | 16 | KMEANS | 12 | 84.892 |
| PCA | 6 | KMEANS | 12 | 84.892 |
| PCA | 8 | KMEANS | 11 | 84.892 |
| NMF | 9 | KMEANS | 12 | 84.892 |
| NMF | 5 | KMEANS | 7 | 84.173 |
| NMF | 6 | KMEANS | 7 | 84.173 |
| PCA | 9 | KMEANS | 4 | 84.173 |
| NMF | 16 | KMEANS | 9 | 83.453 |
| NMF | 16 | KMEANS | 12 | 83.453 |
| NMF | 11 | KMEANS | 11 | 83.453 |
| NMF | 18 | KMEANS | 12 | 83.453 |
| NMF | 15 | KMEANS | 12 | 83.453 |
| PCA | 4 | KMEANS | 5 | 83.453 |
| NMF | 16 | KMEANS | 8 | 82.734 |
| NMF | 9 | KMEANS | 9 | 82.734 |
| NMF | 7 | KMEANS | 8 | 82.734 |
| NMF | 14 | KMEANS | 10 | 82.734 |
| PCA | 7 | KMEANS | 5 | 82.734 |
| PCA | 13 | KMEANS | 5 | 82.734 |
| PCA | 16 | KMEANS | 5 | 82.734 |
| NMF | 8 | KMEANS | 11 | 82.014 |
| PCA | 14 | KMEANS | 12 | 82.014 |
| NMF | 11 | KMEANS | 6 | 82.014 |
| NMF | 12 | KMEANS | 12 | 82.014 |
| NMF | 13 | KMEANS | 7 | 82.014 |
| NMF | 13 | KMEANS | 9 | 82.014 |
| NMF | 11 | KMEANS | 5 | 82.014 |
| PCA | 5 | KMEANS | 10 | 80.576 |
| NMF | 11 | KMEANS | 12 | 80.576 |
| NMF | 15 | KMEANS | 7 | 80.576 |
| NMF | 17 | KMEANS | 7 | 80.576 |
| NMF | 15 | KMEANS | 9 | 79.856 |
| NMF | 15 | KMEANS | 6 | 79.856 |
| PCA | 18 | KMEANS | 12 | 79.856 |
| PCA | 11 | KMEANS | 11 | 79.856 |
| NMF | 13 | KMEANS | 8 | 79.856 |
| NMF | 18 | KMEANS | 8 | 79.856 |
| NMF | 1 | KMEANS | 12 | 79.856 |
| NMF | 14 | KMEANS | 6 | 79.856 |
| PCA | 13 | KMEANS | 12 | 79.137 |
| NMF | 17 | KMEANS | 8 | 79.137 |
| PCA | 2 | KMEANS | 12 | 78.417 |
| NMF | 9 | KMEANS | 5 | 78.417 |
| NMF | 14 | KMEANS | 11 | 77.698 |
| PCA | 17 | KMEANS | 5 | 77.698 |
| NMF | 11 | KMEANS | 10 | 76.978 |
| NMF | 3 | KMEANS | 10 | 76.978 |
| NMF | 15 | KMEANS | 8 | 76.978 |
| NMF | 17 | KMEANS | 6 | 76.978 |
| NMF | 9 | KMEANS | 11 | 76.259 |
| NMF | 8 | KMEANS | 10 | 76.259 |
| NMF | 8 | KMEANS | 8 | 76.259 |
| NMF | 9 | KMEANS | 8 | 76.259 |
| PCA | 1 | KMEANS | 10 | 76.259 |
| NMF | 7 | KMEANS | 11 | 75.540 |
| NMF | 2 | KMEANS | 11 | 75.540 |
| NMF | 5 | KMEANS | 10 | 74.820 |
| PCA | 10 | KMEANS | 9 | 74.820 |
| PCA | 13 | KMEANS | 9 | 74.820 |
| PCA | 7 | KMEANS | 9 | 74.820 |
| NMF | 17 | KMEANS | 4 | 74.820 |
| PCA | 17 | KMEANS | 8 | 74.101 |
| NMF | 6 | KMEANS | 10 | 74.101 |
| PCA | 9 | KMEANS | 10 | 74.101 |
| PCA | 8 | KMEANS | 5 | 74.101 |
| NMF | 17 | KMEANS | 3 | 74.101 |
| PCA | 10 | KMEANS | 8 | 73.381 |
| PCA | 14 | KMEANS | 8 | 73.381 |
| NMF | 2 | KMEANS | 10 | 73.381 |
| PCA | 4 | KMEANS | 8 | 72.662 |
| PCA | 14 | KMEANS | 7 | 72.662 |
| PCA | 8 | KMEANS | 6 | 72.662 |
| PCA | 12 | KMEANS | 6 | 72.662 |
| PCA | 13 | KMEANS | 6 | 72.662 |
| NMF | 13 | KMEANS | 11 | 71.942 |
| PCA | 18 | KMEANS | 10 | 71.942 |
| PCA | 5 | KMEANS | 7 | 71.223 |
| PCA | 4 | KMEANS | 11 | 71.223 |
| PCA | 4 | KMEANS | 10 | 71.223 |
| NMF | 15 | KMEANS | 11 | 70.504 |
| PCA | 3 | KMEANS | 12 | 70.504 |
| NMF | 10 | KMEANS | 7 | 70.504 |
| PCA | 11 | KMEANS | 12 | 69.784 |
| NMF | 15 | KMEANS | 5 | 69.784 |
| PCA | 8 | KMEANS | 12 | 69.065 |
| PCA | 17 | KMEANS | 10 | 69.065 |
| NMF | 4 | KMEANS | 5 | 69.065 |
| NMF | 13 | KMEANS | 12 | 68.345 |
| PCA | 9 | KMEANS | 12 | 68.345 |
| PCA | 4 | KMEANS | 12 | 68.345 |
| PCA | 5 | KMEANS | 12 | 67.626 |
| PCA | 3 | KMEANS | 10 | 67.626 |
| PCA | 11 | KMEANS | 7 | 66.906 |
| NMF | 3 | KMEANS | 9 | 66.906 |
| NMF | 2 | KMEANS | 9 | 66.906 |
| PCA | 2 | KMEANS | 11 | 66.187 |
| PCA | 16 | KMEANS | 10 | 66.187 |
| PCA | 2 | KMEANS | 10 | 66.187 |
| PCA | 3 | KMEANS | 9 | 66.187 |
| NMF | 18 | KMEANS | 9 | 66.187 |
| NMF | 2 | KMEANS | 7 | 66.187 |
| NMF | 5 | KMEANS | 12 | 65.468 |
| PCA | 15 | KMEANS | 11 | 65.468 |
| NMF | 14 | KMEANS | 9 | 64.748 |
| PCA | 17 | KMEANS | 11 | 64.748 |
| PCA | 14 | KMEANS | 11 | 64.748 |
| NMF | 4 | KMEANS | 11 | 64.748 |
| NMF | 7 | KMEANS | 12 | 64.029 |
| NMF | 13 | KMEANS | 10 | 64.029 |
| NMF | 17 | KMEANS | 12 | 64.029 |
| NMF | 11 | KMEANS | 4 | 64.029 |
| NMF | 14 | KMEANS | 4 | 63.309 |
| PCA | 5 | KMEANS | 11 | 62.590 |
| PCA | 13 | KMEANS | 11 | 62.590 |
| NMF | 18 | KMEANS | 7 | 62.590 |
| PCA | 15 | KMEANS | 9 | 61.871 |
| NMF | 10 | KMEANS | 4 | 61.871 |
| PCA | 2 | KMEANS | 8 | 61.151 |
| PCA | 6 | KMEANS | 10 | 61.151 |
| NMF | 9 | KMEANS | 6 | 61.151 |
| NMF | 16 | KMEANS | 10 | 60.432 |
| NMF | 3 | KMEANS | 8 | 60.432 |
| NMF | 3 | KMEANS | 7 | 60.432 |
| PCA | 13 | KMEANS | 10 | 60.432 |
| PCA | 18 | KMEANS | 11 | 60.432 |
| NMF | 15 | KMEANS | 10 | 59.712 |
| NMF | 3 | KMEANS | 6 | 59.712 |
| PCA | 8 | KMEANS | 10 | 59.712 |
| NMF | 13 | KMEANS | 3 | 59.712 |
| NMF | 18 | KMEANS | 4 | 59.712 |
| PCA | 17 | KMEANS | 9 | 58.993 |
| NMF | 11 | KMEANS | 3 | 58.993 |
| NMF | 7 | KMEANS | 9 | 56.115 |
| NMF | 1 | KMEANS | 11 | 56.115 |
| NMF | 6 | KMEANS | 11 | 55.396 |
| NMF | 6 | KMEANS | 9 | 55.396 |
| NMF | 9 | KMEANS | 7 | 55.396 |
| PCA | 11 | KMEANS | 10 | 54.676 |
| NMF | 10 | KMEANS | 9 | 54.676 |
| NMF | 5 | KMEANS | 9 | 52.518 |
| NMF | 2 | KMEANS | 6 | 51.799 |
| PCA | 11 | KMEANS | 8 | 51.079 |
| PCA | 5 | KMEANS | 9 | 51.079 |
| NMF | 2 | KMEANS | 5 | 51.079 |
| PCA | 4 | KMEANS | 9 | 50.360 |
| NMF | 7 | KMEANS | 7 | 49.640 |
| PCA | 8 | KMEANS | 9 | 49.640 |
| PCA | 16 | KMEANS | 9 | 49.640 |
| NMF | 4 | KMEANS | 10 | 49.640 |
| PCA | 9 | KMEANS | 9 | 49.640 |
| NMF | 4 | KMEANS | 9 | 48.921 |
| NMF | 12 | KMEANS | 6 | 48.921 |
| NMF | 10 | KMEANS | 8 | 48.201 |
| PCA | 6 | KMEANS | 5 | 48.201 |
| NMF | 10 | KMEANS | 5 | 47.482 |
| PCA | 5 | KMEANS | 5 | 47.482 |
| NMF | 10 | KMEANS | 6 | 46.763 |
| NMF | 12 | KMEANS | 7 | 46.763 |
| NMF | 13 | KMEANS | 6 | 46.763 |
| NMF | 13 | KMEANS | 5 | 46.043 |
| NMF | 12 | KMEANS | 4 | 46.043 |
| PCA | 16 | KMEANS | 8 | 45.324 |
| NMF | 12 | KMEANS | 5 | 45.324 |
| PCA | 1 | KMEANS | 11 | 45.324 |
| PCA | 1 | KMEANS | 12 | 45.324 |
| PCA | 3 | KMEANS | 8 | 44.604 |
| PCA | 10 | KMEANS | 7 | 44.604 |
| PCA | 15 | KMEANS | 7 | 44.604 |
| NMF | 6 | KMEANS | 6 | 44.604 |
| NMF | 5 | KMEANS | 6 | 44.604 |
| NMF | 4 | KMEANS | 6 | 44.604 |
| NMF | 13 | KMEANS | 4 | 44.604 |
| PCA | 1 | KMEANS | 9 | 44.604 |
| NMF | 1 | KMEANS | 9 | 44.604 |
| NMF | 1 | KMEANS | 10 | 44.604 |
| PCA | 6 | KMEANS | 7 | 43.885 |
| PCA | 12 | KMEANS | 8 | 43.885 |
| NMF | 15 | KMEANS | 4 | 43.885 |
| NMF | 18 | KMEANS | 5 | 43.885 |
| PCA | 7 | KMEANS | 7 | 43.165 |
| PCA | 7 | KMEANS | 8 | 43.165 |
| PCA | 15 | KMEANS | 8 | 43.165 |
| PCA | 18 | KMEANS | 8 | 43.165 |
| PCA | 3 | KMEANS | 6 | 43.165 |
| PCA | 9 | KMEANS | 6 | 43.165 |
| PCA | 4 | KMEANS | 6 | 43.165 |
| PCA | 17 | KMEANS | 7 | 42.446 |
| PCA | 10 | KMEANS | 6 | 42.446 |
| NMF | 16 | KMEANS | 5 | 42.446 |
| PCA | 12 | KMEANS | 7 | 41.727 |
| PCA | 18 | KMEANS | 7 | 41.727 |
| PCA | 5 | KMEANS | 6 | 41.727 |
| PCA | 16 | KMEANS | 6 | 41.727 |
| PCA | 17 | KMEANS | 6 | 41.727 |
| NMF | 16 | KMEANS | 4 | 41.727 |
| NMF | 12 | KMEANS | 3 | 41.727 |
| PCA | 18 | KMEANS | 5 | 41.007 |
| NMF | 15 | KMEANS | 3 | 39.568 |
| NMF | 16 | KMEANS | 3 | 39.568 |
| NMF | 10 | KMEANS | 3 | 39.568 |
| NMF | 14 | KMEANS | 3 | 38.849 |
| PCA | 10 | KMEANS | 5 | 38.849 |
| PCA | 2 | KMEANS | 6 | 37.410 |
| PCA | 9 | KMEANS | 5 | 35.971 |
| PCA | 14 | KMEANS | 5 | 35.971 |
| PCA | 12 | KMEANS | 5 | 35.252 |
| PCA | 11 | KMEANS | 5 | 33.813 |
| PCA | 1 | KMEANS | 7 | 30.935 |
| NMF | 7 | KMEANS | 6 | 30.216 |
| PCA | 15 | KMEANS | 5 | 30.216 |
| PCA | 1 | KMEANS | 8 | 28.777 |
| NMF | 1 | KMEANS | 8 | 28.058 |
| NMF | 3 | KMEANS | 5 | 23.741 |
| NMF | 5 | KMEANS | 5 | 23.741 |
| NMF | 6 | KMEANS | 5 | 23.741 |
| NMF | 1 | KMEANS | 6 | 23.741 |
| PCA | 1 | KMEANS | 5 | 23.741 |
| NMF | 1 | KMEANS | 5 | 23.741 |
| PCA | 1 | KMEANS | 6 | 23.022 |
| PCA | 1 | MEANSHIFT | 7 | 22.302 |
| NMF | 1 | MEANSHIFT | 7 | 22.302 |
| NMF | 4 | MEANSHIFT | 7 | 21.583 |
| NMF | 2 | MEANSHIFT | 7 | 21.583 |
| NMF | 3 | MEANSHIFT | 7 | 20.863 |
| NMF | 7 | KMEANS | 5 | 20.863 |
| PCA | 5 | MEANSHIFT | 6 | 20.863 |
| PCA | 6 | MEANSHIFT | 6 | 20.863 |
| PCA | 7 | MEANSHIFT | 6 | 20.863 |
| PCA | 8 | MEANSHIFT | 6 | 20.863 |
| PCA | 9 | MEANSHIFT | 6 | 20.863 |
| PCA | 10 | MEANSHIFT | 6 | 20.863 |
| PCA | 11 | MEANSHIFT | 6 | 20.863 |
| PCA | 12 | MEANSHIFT | 6 | 20.863 |
| PCA | 13 | MEANSHIFT | 6 | 20.863 |
| PCA | 14 | MEANSHIFT | 6 | 20.863 |
| PCA | 15 | MEANSHIFT | 6 | 20.863 |
| PCA | 16 | MEANSHIFT | 6 | 20.863 |
| PCA | 17 | MEANSHIFT | 6 | 20.863 |
| PCA | 18 | MEANSHIFT | 6 | 20.863 |
| PCA | 4 | KMEANS | 4 | 20.863 |
| NMF | 5 | MEANSHIFT | 8 | 20.144 |
| NMF | 6 | MEANSHIFT | 8 | 20.144 |
| PCA | 8 | KMEANS | 4 | 20.144 |
| PCA | 12 | KMEANS | 4 | 20.144 |
| PCA | 13 | KMEANS | 4 | 20.144 |
| PCA | 14 | KMEANS | 4 | 20.144 |
| PCA | 15 | KMEANS | 4 | 20.144 |
| PCA | 17 | KMEANS | 4 | 20.144 |
| PCA | 2 | MEANSHIFT | 6 | 19.424 |
| PCA | 3 | MEANSHIFT | 6 | 19.424 |
| NMF | 17 | MEANSHIFT | 6 | 19.424 |
| NMF | 1 | KMEANS | 4 | 18.705 |
| PCA | 2 | KMEANS | 4 | 18.705 |
| NMF | 2 | KMEANS | 4 | 18.705 |
| PCA | 3 | KMEANS | 4 | 18.705 |
| PCA | 5 | KMEANS | 4 | 18.705 |
| PCA | 6 | KMEANS | 4 | 18.705 |
| PCA | 7 | KMEANS | 4 | 18.705 |
| PCA | 10 | KMEANS | 4 | 18.705 |
| PCA | 11 | KMEANS | 4 | 18.705 |
| NMF | 18 | MEANSHIFT | 7 | 17.266 |
| NMF | 5 | KMEANS | 4 | 16.547 |
| NMF | 9 | KMEANS | 4 | 16.547 |
| NMF | 7 | MEANSHIFT | 7 | 16.547 |
| NMF | 3 | KMEANS | 4 | 15.827 |
| NMF | 4 | KMEANS | 4 | 15.827 |
| NMF | 6 | KMEANS | 4 | 15.827 |
| NMF | 10 | MEANSHIFT | 4 | 11.511 |
| NMF | 11 | MEANSHIFT | 4 | 11.511 |
| NMF | 12 | MEANSHIFT | 4 | 11.511 |
| NMF | 13 | MEANSHIFT | 4 | 11.511 |
| PCA | 4 | MEANSHIFT | 5 | 10.791 |
| NMF | 14 | MEANSHIFT | 4 | 10.791 |
| NMF | 15 | MEANSHIFT | 4 | 10.791 |
| NMF | 16 | MEANSHIFT | 4 | 10.791 |
| NMF | 8 | MEANSHIFT | 4 | 10.072 |
| NMF | 9 | MEANSHIFT | 4 | 10.072 |
| NMF | 1 | KMEANS | 3 | 10.072 |
| PCA | 2 | KMEANS | 3 | 10.072 |
| NMF | 2 | KMEANS | 3 | 10.072 |
| PCA | 3 | KMEANS | 3 | 10.072 |
| NMF | 3 | KMEANS | 3 | 10.072 |
| PCA | 4 | KMEANS | 3 | 10.072 |
| NMF | 4 | KMEANS | 3 | 10.072 |
| PCA | 5 | KMEANS | 3 | 10.072 |
| NMF | 5 | KMEANS | 3 | 10.072 |
| PCA | 7 | KMEANS | 3 | 10.072 |
| NMF | 7 | KMEANS | 3 | 10.072 |
| PCA | 8 | KMEANS | 3 | 10.072 |
| NMF | 8 | KMEANS | 3 | 10.072 |
| PCA | 9 | KMEANS | 3 | 10.072 |
| NMF | 9 | KMEANS | 3 | 10.072 |
| PCA | 10 | KMEANS | 3 | 10.072 |
| PCA | 11 | KMEANS | 3 | 10.072 |
| PCA | 12 | KMEANS | 3 | 10.072 |
| PCA | 13 | KMEANS | 3 | 10.072 |
| PCA | 15 | KMEANS | 3 | 10.072 |
| PCA | 16 | KMEANS | 3 | 10.072 |
| PCA | 17 | KMEANS | 3 | 10.072 |
| PCA | 18 | KMEANS | 3 | 10.072 |
| NMF | 18 | KMEANS | 3 | 10.072 |
| PCA | 1 | KMEANS | 2 | 6.475 |
| NMF | 1 | KMEANS | 2 | 6.475 |
| PCA | 2 | KMEANS | 2 | 6.475 |
| NMF | 2 | KMEANS | 2 | 6.475 |
| PCA | 3 | KMEANS | 2 | 6.475 |
| NMF | 3 | KMEANS | 2 | 6.475 |
| PCA | 4 | KMEANS | 2 | 6.475 |
| NMF | 4 | KMEANS | 2 | 6.475 |
| PCA | 5 | KMEANS | 2 | 6.475 |
| NMF | 5 | KMEANS | 2 | 6.475 |
| PCA | 6 | KMEANS | 2 | 6.475 |
| NMF | 6 | KMEANS | 2 | 6.475 |
| PCA | 7 | KMEANS | 2 | 6.475 |
| NMF | 7 | KMEANS | 2 | 6.475 |
| PCA | 8 | KMEANS | 2 | 6.475 |
| NMF | 8 | KMEANS | 2 | 6.475 |
| PCA | 9 | KMEANS | 2 | 6.475 |
| NMF | 9 | KMEANS | 2 | 6.475 |
| PCA | 10 | KMEANS | 2 | 6.475 |
| NMF | 10 | KMEANS | 2 | 6.475 |
| PCA | 11 | KMEANS | 2 | 6.475 |
| NMF | 11 | KMEANS | 2 | 6.475 |
| PCA | 12 | KMEANS | 2 | 6.475 |
| NMF | 12 | KMEANS | 2 | 6.475 |
| PCA | 13 | KMEANS | 2 | 6.475 |
| NMF | 13 | KMEANS | 2 | 6.475 |
| PCA | 14 | KMEANS | 2 | 6.475 |
| NMF | 14 | KMEANS | 2 | 6.475 |
| PCA | 15 | KMEANS | 2 | 6.475 |
| NMF | 15 | KMEANS | 2 | 6.475 |
| PCA | 16 | KMEANS | 2 | 6.475 |
| NMF | 16 | KMEANS | 2 | 6.475 |
| PCA | 17 | KMEANS | 2 | 6.475 |
| NMF | 17 | KMEANS | 2 | 6.475 |
| PCA | 18 | KMEANS | 2 | 6.475 |
| NMF | 18 | KMEANS | 2 | 6.475 |

## Appendix 5 Full Table of Results for Clustering Accuracy

| **Decomposition** | **No. Features** | **Clustering** | **No. Clusters** | **Clustering Accuracy** |
| --- | --- | --- | --- | --- |
| NMF | 16 | KMEANS | 9 | 46.763 |
| NMF | 16 | KMEANS | 8 | 46.043 |
| NMF | 15 | KMEANS | 9 | 43.165 |
| NMF | 14 | KMEANS | 12 | 41.727 |
| NMF | 14 | KMEANS | 11 | 41.727 |
| PCA | 5 | KMEANS | 8 | 40.288 |
| PCA | 9 | KMEANS | 8 | 40.288 |
| PCA | 9 | KMEANS | 7 | 40.288 |
| PCA | 8 | KMEANS | 8 | 40.288 |
| NMF | 15 | KMEANS | 6 | 40.288 |
| PCA | 10 | KMEANS | 8 | 40.288 |
| PCA | 14 | KMEANS | 8 | 40.288 |
| NMF | 16 | KMEANS | 10 | 40.288 |
| PCA | 7 | KMEANS | 7 | 40.288 |
| PCA | 7 | KMEANS | 8 | 40.288 |
| PCA | 15 | KMEANS | 8 | 40.288 |
| PCA | 18 | KMEANS | 8 | 40.288 |
| PCA | 9 | KMEANS | 11 | 39.568 |
| NMF | 11 | KMEANS | 10 | 39.568 |
| NMF | 15 | KMEANS | 10 | 39.568 |
| PCA | 17 | KMEANS | 9 | 39.568 |
| NMF | 10 | KMEANS | 10 | 38.849 |
| NMF | 16 | KMEANS | 7 | 38.849 |
| PCA | 6 | KMEANS | 8 | 38.849 |
| PCA | 17 | KMEANS | 8 | 38.849 |
| NMF | 13 | KMEANS | 12 | 38.849 |
| PCA | 3 | KMEANS | 7 | 38.129 |
| PCA | 13 | KMEANS | 7 | 38.129 |
| NMF | 9 | KMEANS | 10 | 38.129 |
| NMF | 8 | KMEANS | 9 | 38.129 |
| NMF | 12 | KMEANS | 11 | 38.129 |
| NMF | 9 | KMEANS | 9 | 38.129 |
| NMF | 8 | KMEANS | 11 | 38.129 |
| NMF | 9 | KMEANS | 11 | 38.129 |
| NMF | 5 | KMEANS | 10 | 38.129 |
| PCA | 4 | KMEANS | 8 | 38.129 |
| NMF | 13 | KMEANS | 11 | 38.129 |
| NMF | 14 | KMEANS | 9 | 38.129 |
| PCA | 16 | KMEANS | 8 | 38.129 |
| PCA | 3 | KMEANS | 8 | 38.129 |
| PCA | 6 | KMEANS | 7 | 38.129 |
| PCA | 12 | KMEANS | 8 | 38.129 |
| PCA | 8 | KMEANS | 7 | 37.410 |
| PCA | 16 | KMEANS | 7 | 37.410 |
| NMF | 12 | KMEANS | 10 | 37.410 |
| NMF | 4 | KMEANS | 12 | 37.410 |
| PCA | 14 | KMEANS | 7 | 37.410 |
| NMF | 15 | KMEANS | 11 | 37.410 |
| PCA | 10 | KMEANS | 7 | 37.410 |
| PCA | 15 | KMEANS | 7 | 37.410 |
| PCA | 17 | KMEANS | 7 | 37.410 |
| PCA | 12 | KMEANS | 7 | 37.410 |
| PCA | 18 | KMEANS | 7 | 37.410 |
| PCA | 14 | KMEANS | 6 | 36.691 |
| PCA | 15 | KMEANS | 6 | 36.691 |
| PCA | 7 | KMEANS | 6 | 36.691 |
| PCA | 11 | KMEANS | 6 | 36.691 |
| PCA | 13 | KMEANS | 8 | 36.691 |
| PCA | 6 | KMEANS | 6 | 36.691 |
| NMF | 16 | KMEANS | 12 | 36.691 |
| NMF | 8 | KMEANS | 10 | 36.691 |
| NMF | 7 | KMEANS | 11 | 36.691 |
| PCA | 8 | KMEANS | 6 | 36.691 |
| PCA | 12 | KMEANS | 6 | 36.691 |
| PCA | 13 | KMEANS | 6 | 36.691 |
| NMF | 5 | KMEANS | 12 | 36.691 |
| NMF | 7 | KMEANS | 9 | 36.691 |
| PCA | 3 | KMEANS | 6 | 36.691 |
| PCA | 9 | KMEANS | 6 | 36.691 |
| PCA | 10 | KMEANS | 6 | 36.691 |
| PCA | 5 | KMEANS | 6 | 36.691 |
| PCA | 16 | KMEANS | 6 | 36.691 |
| PCA | 17 | KMEANS | 6 | 36.691 |
| PCA | 4 | KMEANS | 7 | 35.971 |
| NMF | 16 | KMEANS | 11 | 35.971 |
| NMF | 6 | KMEANS | 8 | 35.971 |
| PCA | 7 | KMEANS | 12 | 35.971 |
| PCA | 10 | KMEANS | 11 | 35.971 |
| PCA | 18 | KMEANS | 12 | 35.971 |
| NMF | 6 | KMEANS | 10 | 35.971 |
| PCA | 5 | KMEANS | 11 | 35.971 |
| NMF | 6 | KMEANS | 11 | 35.971 |
| PCA | 11 | KMEANS | 8 | 35.971 |
| NMF | 7 | KMEANS | 7 | 35.971 |
| NMF | 6 | KMEANS | 6 | 35.971 |
| PCA | 11 | KMEANS | 9 | 35.252 |
| PCA | 14 | KMEANS | 9 | 35.252 |
| PCA | 6 | KMEANS | 9 | 35.252 |
| PCA | 18 | KMEANS | 9 | 35.252 |
| PCA | 16 | KMEANS | 12 | 35.252 |
| PCA | 10 | KMEANS | 9 | 35.252 |
| PCA | 13 | KMEANS | 9 | 35.252 |
| PCA | 5 | KMEANS | 7 | 35.252 |
| PCA | 5 | KMEANS | 12 | 35.252 |
| PCA | 17 | KMEANS | 11 | 35.252 |
| PCA | 13 | KMEANS | 11 | 35.252 |
| NMF | 3 | KMEANS | 8 | 35.252 |
| NMF | 5 | KMEANS | 9 | 35.252 |
| PCA | 5 | KMEANS | 9 | 35.252 |
| PCA | 8 | KMEANS | 9 | 35.252 |
| PCA | 16 | KMEANS | 9 | 35.252 |
| PCA | 4 | KMEANS | 6 | 35.252 |
| PCA | 17 | KMEANS | 12 | 34.532 |
| NMF | 8 | KMEANS | 7 | 34.532 |
| NMF | 4 | KMEANS | 8 | 34.532 |
| PCA | 7 | KMEANS | 9 | 34.532 |
| PCA | 9 | KMEANS | 12 | 34.532 |
| PCA | 2 | KMEANS | 11 | 34.532 |
| NMF | 3 | KMEANS | 7 | 34.532 |
| NMF | 5 | KMEANS | 6 | 34.532 |
| PCA | 12 | KMEANS | 12 | 33.813 |
| NMF | 11 | KMEANS | 11 | 33.813 |
| NMF | 7 | KMEANS | 8 | 33.813 |
| PCA | 5 | KMEANS | 10 | 33.813 |
| PCA | 9 | KMEANS | 10 | 33.813 |
| PCA | 15 | KMEANS | 9 | 33.813 |
| NMF | 4 | KMEANS | 10 | 33.813 |
| NMF | 4 | KMEANS | 9 | 33.813 |
| PCA | 13 | KMEANS | 12 | 33.094 |
| PCA | 11 | KMEANS | 7 | 33.094 |
| PCA | 15 | KMEANS | 11 | 33.094 |
| NMF | 6 | KMEANS | 9 | 33.094 |
| PCA | 4 | KMEANS | 9 | 33.094 |
| PCA | 9 | KMEANS | 9 | 33.094 |
| PCA | 10 | KMEANS | 12 | 32.374 |
| PCA | 12 | KMEANS | 9 | 32.374 |
| PCA | 18 | KMEANS | 6 | 32.374 |
| PCA | 7 | KMEANS | 10 | 32.374 |
| NMF | 8 | KMEANS | 8 | 32.374 |
| PCA | 8 | KMEANS | 12 | 32.374 |
| PCA | 16 | KMEANS | 11 | 31.655 |
| PCA | 7 | KMEANS | 11 | 31.655 |
| PCA | 4 | KMEANS | 11 | 31.655 |
| PCA | 16 | KMEANS | 10 | 31.655 |
| PCA | 2 | KMEANS | 8 | 31.655 |
| PCA | 12 | KMEANS | 10 | 30.935 |
| PCA | 17 | KMEANS | 10 | 30.935 |
| PCA | 12 | KMEANS | 11 | 30.216 |
| PCA | 2 | KMEANS | 9 | 30.216 |
| PCA | 3 | KMEANS | 11 | 30.216 |
| PCA | 14 | KMEANS | 10 | 30.216 |
| PCA | 15 | KMEANS | 10 | 30.216 |
| PCA | 6 | KMEANS | 12 | 30.216 |
| NMF | 5 | KMEANS | 7 | 30.216 |
| PCA | 4 | KMEANS | 12 | 30.216 |
| PCA | 3 | KMEANS | 10 | 30.216 |
| PCA | 2 | KMEANS | 10 | 30.216 |
| PCA | 13 | KMEANS | 10 | 30.216 |
| PCA | 11 | KMEANS | 10 | 30.216 |
| NMF | 5 | KMEANS | 8 | 29.496 |
| PCA | 8 | KMEANS | 11 | 29.496 |
| NMF | 9 | KMEANS | 12 | 29.496 |
| PCA | 11 | KMEANS | 11 | 29.496 |
| PCA | 3 | KMEANS | 9 | 29.496 |
| PCA | 15 | KMEANS | 12 | 28.777 |
| PCA | 18 | KMEANS | 11 | 28.777 |
| NMF | 3 | KMEANS | 6 | 28.777 |
| NMF | 4 | KMEANS | 7 | 28.058 |
| NMF | 17 | KMEANS | 8 | 28.058 |
| PCA | 4 | KMEANS | 10 | 28.058 |
| PCA | 8 | KMEANS | 10 | 28.058 |
| PCA | 10 | KMEANS | 10 | 27.338 |
| PCA | 18 | KMEANS | 10 | 27.338 |
| PCA | 6 | KMEANS | 10 | 27.338 |
| NMF | 3 | KMEANS | 12 | 26.619 |
| NMF | 6 | KMEANS | 7 | 26.619 |
| PCA | 2 | KMEANS | 12 | 26.619 |
| PCA | 3 | KMEANS | 12 | 26.619 |
| NMF | 5 | KMEANS | 11 | 25.899 |
| NMF | 7 | KMEANS | 10 | 25.899 |
| NMF | 3 | KMEANS | 10 | 25.899 |
| PCA | 14 | KMEANS | 11 | 25.899 |
| NMF | 12 | KMEANS | 8 | 25.180 |
| PCA | 14 | KMEANS | 12 | 25.180 |
| NMF | 7 | KMEANS | 12 | 25.180 |
| NMF | 11 | KMEANS | 12 | 24.460 |
| NMF | 13 | KMEANS | 8 | 24.460 |
| NMF | 15 | KMEANS | 8 | 24.460 |
| NMF | 18 | KMEANS | 7 | 24.460 |
| NMF | 10 | KMEANS | 9 | 24.460 |
| NMF | 3 | KMEANS | 11 | 23.741 |
| NMF | 11 | KMEANS | 9 | 23.741 |
| NMF | 6 | KMEANS | 12 | 23.741 |
| NMF | 8 | KMEANS | 12 | 23.741 |
| NMF | 10 | KMEANS | 12 | 23.741 |
| NMF | 12 | KMEANS | 9 | 23.022 |
| NMF | 11 | KMEANS | 6 | 23.022 |
| NMF | 18 | KMEANS | 8 | 23.022 |
| NMF | 10 | KMEANS | 7 | 23.022 |
| PCA | 11 | KMEANS | 12 | 23.022 |
| NMF | 3 | KMEANS | 9 | 23.022 |
| NMF | 12 | KMEANS | 12 | 22.302 |
| NMF | 15 | KMEANS | 7 | 22.302 |
| NMF | 2 | KMEANS | 11 | 22.302 |
| NMF | 11 | KMEANS | 7 | 21.583 |
| NMF | 10 | KMEANS | 11 | 21.583 |
| NMF | 13 | KMEANS | 7 | 21.583 |
| NMF | 17 | KMEANS | 7 | 21.583 |
| NMF | 18 | KMEANS | 9 | 21.583 |
| NMF | 4 | KMEANS | 11 | 21.583 |
| NMF | 10 | KMEANS | 6 | 21.583 |
| NMF | 17 | KMEANS | 10 | 20.863 |
| NMF | 18 | KMEANS | 12 | 20.863 |
| NMF | 10 | KMEANS | 8 | 20.863 |
| NMF | 18 | KMEANS | 10 | 20.144 |
| NMF | 13 | KMEANS | 10 | 20.144 |
| NMF | 11 | KMEANS | 8 | 19.424 |
| NMF | 17 | KMEANS | 9 | 19.424 |
| NMF | 13 | KMEANS | 9 | 19.424 |
| NMF | 14 | KMEANS | 5 | 18.705 |
| NMF | 18 | KMEANS | 11 | 18.705 |
| PCA | 6 | KMEANS | 11 | 18.705 |
| NMF | 4 | KMEANS | 5 | 18.705 |
| NMF | 17 | KMEANS | 12 | 18.705 |
| NMF | 13 | KMEANS | 5 | 18.705 |
| NMF | 16 | KMEANS | 5 | 18.705 |
| NMF | 14 | KMEANS | 8 | 17.986 |
| NMF | 2 | KMEANS | 8 | 17.986 |
| NMF | 15 | KMEANS | 12 | 17.266 |
| NMF | 14 | KMEANS | 10 | 17.266 |
| NMF | 11 | KMEANS | 5 | 17.266 |
| NMF | 9 | KMEANS | 8 | 17.266 |
| NMF | 2 | KMEANS | 10 | 17.266 |
| NMF | 1 | KMEANS | 11 | 17.266 |
| NMF | 10 | KMEANS | 5 | 17.266 |
| NMF | 12 | KMEANS | 7 | 17.266 |
| NMF | 4 | KMEANS | 6 | 17.266 |
| NMF | 16 | KMEANS | 6 | 16.547 |
| NMF | 14 | KMEANS | 7 | 16.547 |
| NMF | 1 | KMEANS | 12 | 16.547 |
| NMF | 2 | KMEANS | 9 | 16.547 |
| NMF | 13 | KMEANS | 6 | 16.547 |
| NMF | 2 | KMEANS | 12 | 15.108 |
| NMF | 14 | KMEANS | 6 | 15.108 |
| NMF | 9 | KMEANS | 5 | 15.108 |
| NMF | 15 | KMEANS | 5 | 15.108 |
| NMF | 9 | KMEANS | 7 | 15.108 |
| NMF | 12 | KMEANS | 6 | 15.108 |
| NMF | 12 | KMEANS | 5 | 15.108 |
| NMF | 15 | KMEANS | 4 | 15.108 |
| NMF | 3 | MEANSHIFT | 7 | 15.108 |
| NMF | 8 | KMEANS | 5 | 14.388 |
| NMF | 17 | KMEANS | 11 | 14.388 |
| NMF | 7 | KMEANS | 4 | 14.388 |
| NMF | 8 | KMEANS | 4 | 14.388 |
| NMF | 17 | KMEANS | 6 | 14.388 |
| NMF | 11 | KMEANS | 4 | 14.388 |
| NMF | 14 | KMEANS | 4 | 14.388 |
| NMF | 10 | KMEANS | 4 | 14.388 |
| NMF | 12 | KMEANS | 4 | 14.388 |
| NMF | 13 | KMEANS | 4 | 14.388 |
| NMF | 16 | KMEANS | 4 | 14.388 |
| NMF | 4 | MEANSHIFT | 7 | 14.388 |
| NMF | 7 | KMEANS | 5 | 14.388 |
| NMF | 5 | MEANSHIFT | 8 | 14.388 |
| NMF | 6 | MEANSHIFT | 8 | 14.388 |
| NMF | 5 | KMEANS | 4 | 14.388 |
| NMF | 9 | KMEANS | 4 | 14.388 |
| NMF | 3 | KMEANS | 4 | 14.388 |
| NMF | 4 | KMEANS | 4 | 14.388 |
| NMF | 6 | KMEANS | 4 | 14.388 |
| NMF | 15 | KMEANS | 3 | 13.669 |
| NMF | 16 | KMEANS | 3 | 13.669 |
| NMF | 8 | KMEANS | 6 | 12.950 |
| PCA | 4 | KMEANS | 5 | 12.950 |
| NMF | 9 | KMEANS | 6 | 12.950 |
| NMF | 13 | KMEANS | 3 | 12.950 |
| NMF | 11 | KMEANS | 3 | 12.950 |
| NMF | 12 | KMEANS | 3 | 12.950 |
| NMF | 10 | KMEANS | 3 | 12.950 |
| NMF | 14 | KMEANS | 3 | 12.950 |
| PCA | 2 | KMEANS | 6 | 12.950 |
| NMF | 7 | KMEANS | 6 | 12.950 |
| NMF | 3 | KMEANS | 5 | 12.950 |
| NMF | 5 | KMEANS | 5 | 12.950 |
| NMF | 6 | KMEANS | 5 | 12.950 |
| NMF | 17 | KMEANS | 5 | 12.230 |
| PCA | 2 | KMEANS | 7 | 12.230 |
| PCA | 7 | KMEANS | 5 | 12.230 |
| PCA | 13 | KMEANS | 5 | 12.230 |
| PCA | 16 | KMEANS | 5 | 12.230 |
| NMF | 18 | KMEANS | 4 | 12.230 |
| PCA | 9 | KMEANS | 5 | 12.230 |
| PCA | 14 | KMEANS | 5 | 12.230 |
| PCA | 11 | KMEANS | 5 | 12.230 |
| PCA | 15 | KMEANS | 5 | 12.230 |
| PCA | 2 | KMEANS | 5 | 11.511 |
| PCA | 3 | KMEANS | 5 | 11.511 |
| PCA | 17 | KMEANS | 5 | 11.511 |
| PCA | 12 | KMEANS | 5 | 11.511 |
| NMF | 18 | KMEANS | 6 | 10.791 |
| NMF | 17 | KMEANS | 4 | 10.791 |
| NMF | 18 | KMEANS | 5 | 10.791 |
| PCA | 1 | KMEANS | 10 | 10.072 |
| NMF | 2 | KMEANS | 7 | 10.072 |
| NMF | 2 | KMEANS | 6 | 10.072 |
| PCA | 1 | KMEANS | 9 | 10.072 |
| NMF | 1 | KMEANS | 9 | 10.072 |
| NMF | 1 | KMEANS | 10 | 10.072 |
| PCA | 8 | KMEANS | 5 | 9.353 |
| NMF | 17 | KMEANS | 3 | 9.353 |
| PCA | 1 | KMEANS | 11 | 9.353 |
| PCA | 1 | KMEANS | 12 | 9.353 |
| PCA | 10 | KMEANS | 5 | 9.353 |
| PCA | 1 | KMEANS | 7 | 9.353 |
| NMF | 1 | KMEANS | 7 | 8.633 |
| NMF | 2 | KMEANS | 5 | 8.633 |
| PCA | 6 | KMEANS | 5 | 8.633 |
| PCA | 5 | KMEANS | 5 | 8.633 |
| PCA | 18 | KMEANS | 5 | 8.633 |
| NMF | 1 | KMEANS | 6 | 8.633 |
| PCA | 1 | KMEANS | 6 | 8.633 |
| PCA | 1 | MEANSHIFT | 7 | 8.633 |
| NMF | 1 | MEANSHIFT | 7 | 8.633 |
| NMF | 2 | MEANSHIFT | 7 | 8.633 |
| PCA | 5 | MEANSHIFT | 6 | 8.633 |
| PCA | 6 | MEANSHIFT | 6 | 8.633 |
| PCA | 7 | MEANSHIFT | 6 | 8.633 |
| PCA | 8 | MEANSHIFT | 6 | 8.633 |
| PCA | 9 | MEANSHIFT | 6 | 8.633 |
| PCA | 10 | MEANSHIFT | 6 | 8.633 |
| PCA | 11 | MEANSHIFT | 6 | 8.633 |
| PCA | 12 | MEANSHIFT | 6 | 8.633 |
| PCA | 13 | MEANSHIFT | 6 | 8.633 |
| PCA | 14 | MEANSHIFT | 6 | 8.633 |
| PCA | 15 | MEANSHIFT | 6 | 8.633 |
| PCA | 16 | MEANSHIFT | 6 | 8.633 |
| PCA | 17 | MEANSHIFT | 6 | 8.633 |
| PCA | 18 | MEANSHIFT | 6 | 8.633 |
| PCA | 2 | MEANSHIFT | 6 | 8.633 |
| PCA | 3 | MEANSHIFT | 6 | 8.633 |
| NMF | 18 | MEANSHIFT | 7 | 8.633 |
| PCA | 16 | KMEANS | 4 | 7.914 |
| PCA | 1 | KMEANS | 4 | 7.914 |
| PCA | 18 | KMEANS | 4 | 7.914 |
| PCA | 9 | KMEANS | 4 | 7.914 |
| PCA | 1 | KMEANS | 5 | 7.914 |
| NMF | 1 | KMEANS | 5 | 7.914 |
| PCA | 4 | KMEANS | 4 | 7.914 |
| PCA | 8 | KMEANS | 4 | 7.914 |
| PCA | 12 | KMEANS | 4 | 7.914 |
| PCA | 13 | KMEANS | 4 | 7.914 |
| PCA | 14 | KMEANS | 4 | 7.914 |
| PCA | 15 | KMEANS | 4 | 7.914 |
| PCA | 17 | KMEANS | 4 | 7.914 |
| NMF | 17 | MEANSHIFT | 6 | 7.914 |
| NMF | 1 | KMEANS | 4 | 7.914 |
| PCA | 2 | KMEANS | 4 | 7.914 |
| NMF | 2 | KMEANS | 4 | 7.914 |
| PCA | 3 | KMEANS | 4 | 7.914 |
| PCA | 5 | KMEANS | 4 | 7.914 |
| PCA | 6 | KMEANS | 4 | 7.914 |
| PCA | 7 | KMEANS | 4 | 7.914 |
| PCA | 10 | KMEANS | 4 | 7.914 |
| PCA | 11 | KMEANS | 4 | 7.914 |
| PCA | 1 | KMEANS | 8 | 7.194 |
| NMF | 1 | KMEANS | 8 | 7.194 |
| NMF | 8 | MEANSHIFT | 4 | 7.194 |
| NMF | 9 | MEANSHIFT | 4 | 7.194 |
| PCA | 6 | KMEANS | 3 | 6.475 |
| PCA | 1 | KMEANS | 3 | 6.475 |
| NMF | 6 | KMEANS | 3 | 6.475 |
| PCA | 14 | KMEANS | 3 | 6.475 |
| PCA | 4 | MEANSHIFT | 5 | 6.475 |
| NMF | 1 | KMEANS | 3 | 6.475 |
| PCA | 2 | KMEANS | 3 | 6.475 |
| NMF | 2 | KMEANS | 3 | 6.475 |
| PCA | 3 | KMEANS | 3 | 6.475 |
| NMF | 3 | KMEANS | 3 | 6.475 |
| PCA | 4 | KMEANS | 3 | 6.475 |
| NMF | 4 | KMEANS | 3 | 6.475 |
| PCA | 5 | KMEANS | 3 | 6.475 |
| NMF | 5 | KMEANS | 3 | 6.475 |
| PCA | 7 | KMEANS | 3 | 6.475 |
| NMF | 7 | KMEANS | 3 | 6.475 |
| PCA | 8 | KMEANS | 3 | 6.475 |
| NMF | 8 | KMEANS | 3 | 6.475 |
| PCA | 9 | KMEANS | 3 | 6.475 |
| NMF | 9 | KMEANS | 3 | 6.475 |
| PCA | 10 | KMEANS | 3 | 6.475 |
| PCA | 11 | KMEANS | 3 | 6.475 |
| PCA | 12 | KMEANS | 3 | 6.475 |
| PCA | 13 | KMEANS | 3 | 6.475 |
| PCA | 15 | KMEANS | 3 | 6.475 |
| PCA | 16 | KMEANS | 3 | 6.475 |
| PCA | 17 | KMEANS | 3 | 6.475 |
| PCA | 18 | KMEANS | 3 | 6.475 |
| NMF | 18 | KMEANS | 3 | 6.475 |
| NMF | 7 | MEANSHIFT | 7 | 5.755 |
| NMF | 10 | MEANSHIFT | 4 | 5.755 |
| NMF | 11 | MEANSHIFT | 4 | 5.755 |
| NMF | 12 | MEANSHIFT | 4 | 5.755 |
| NMF | 13 | MEANSHIFT | 4 | 5.755 |
| NMF | 14 | MEANSHIFT | 4 | 5.755 |
| NMF | 15 | MEANSHIFT | 4 | 5.755 |
| NMF | 16 | MEANSHIFT | 4 | 5.755 |
| PCA | 1 | KMEANS | 2 | 5.036 |
| NMF | 1 | KMEANS | 2 | 5.036 |
| PCA | 2 | KMEANS | 2 | 5.036 |
| NMF | 2 | KMEANS | 2 | 5.036 |
| PCA | 3 | KMEANS | 2 | 5.036 |
| NMF | 3 | KMEANS | 2 | 5.036 |
| PCA | 4 | KMEANS | 2 | 5.036 |
| NMF | 4 | KMEANS | 2 | 5.036 |
| PCA | 5 | KMEANS | 2 | 5.036 |
| NMF | 5 | KMEANS | 2 | 5.036 |
| PCA | 6 | KMEANS | 2 | 5.036 |
| NMF | 6 | KMEANS | 2 | 5.036 |
| PCA | 7 | KMEANS | 2 | 5.036 |
| NMF | 7 | KMEANS | 2 | 5.036 |
| PCA | 8 | KMEANS | 2 | 5.036 |
| NMF | 8 | KMEANS | 2 | 5.036 |
| PCA | 9 | KMEANS | 2 | 5.036 |
| NMF | 9 | KMEANS | 2 | 5.036 |
| PCA | 10 | KMEANS | 2 | 5.036 |
| NMF | 10 | KMEANS | 2 | 5.036 |
| PCA | 11 | KMEANS | 2 | 5.036 |
| NMF | 11 | KMEANS | 2 | 5.036 |
| PCA | 12 | KMEANS | 2 | 5.036 |
| NMF | 12 | KMEANS | 2 | 5.036 |
| PCA | 13 | KMEANS | 2 | 5.036 |
| NMF | 13 | KMEANS | 2 | 5.036 |
| PCA | 14 | KMEANS | 2 | 5.036 |
| NMF | 14 | KMEANS | 2 | 5.036 |
| PCA | 15 | KMEANS | 2 | 5.036 |
| NMF | 15 | KMEANS | 2 | 5.036 |
| PCA | 16 | KMEANS | 2 | 5.036 |
| NMF | 16 | KMEANS | 2 | 5.036 |
| PCA | 17 | KMEANS | 2 | 5.036 |
| NMF | 17 | KMEANS | 2 | 5.036 |
| PCA | 18 | KMEANS | 2 | 5.036 |
| NMF | 18 | KMEANS | 2 | 5.036 |