Exploring the Performance of Feature Dimensionality Reduction Technique Using Malware Dataset.

Azaabi Cletus
School of Sciences, University of Energy and Natural Resources, Sunyani, Ghana
Email: cleinhim@yahoo.com
Alex Opoku (PhD)[2]
School of Sciences, University of Energy and Natural Resources, Sunyani, Ghana
Email: Alex.opoku@uenr.edu.gh
Benjamin Weyory (PhD.)[2]
School of Sciences, University of Energy and Natural Resources, Sunyani, Ghana
Email: Benjamin.weyori@uenr.edu.gh

Abstract
Features play a critical role in the machine learning or predictive modelling in general and in malware detection in particular. Machine learning models or algorithms start and die on the bases of the features used in the training and testing phases. To ensure optimum predictive capability of a model and to reduce computational resources, irrelevant, duplicates and unwanted features need to be transformed into lower dimensional space for improved prediction. The paper experimented three feature dimensionality reduction techniques (data-dependent, data-independent and graphic based) dimensionality reduction strategies for the prediction of malware. We implemented a data-dependent (Principal Component Analysis), data-independent (Hashing trick) and graph-based (Uniform Manifold Approximation & Projection) UMAP using malware diagnostic dataset. After an initial experiment on the original dataset with all features using four classifiers, the best performing classifier (SVM) was selected for the implementation. Features were reduced to 16 and the performance of the model was evaluated. The performance of the dimensionality reduction techniques on the dataset was evaluated using Accuracy and false positive Rate (FPR). The performances of the three techniques were compared. The result demonstrate that, the data independent technique (HT) outperformed the other two with accuracy (99.1%), FPR(1.2%) as against the two data-dependent and graph-based at accuracy(98.7%) and False positive (1.7%) respectively. The paper concluded that, data-independent dimensionality reduction technique (HT) produces superior malware detection accuracy, lower FPR with the malware dataset, and consequently presents a high potential for malware detection and classification.

Keywords:
Features, malware, transformation, accuracy, false positive.

1. Introduction

Malware attacks are among the highest threat facing cybersecurity stakeholders globally [1], [2], [3], [4], [5]. These authors suggest that it is among the leading cause of data breach and exposures at individual, corporate and at national levels. The ramification of a system breach is always disastrous including reputational loss, financial loss, data exfiltration [6]. Consequently, early detection of a malware attack is of great essence to prevent malware exploitation and or mitigation to ensure the Confidentiality, Integrity and Availability (CIA) and other security goals and resources.

Over the years, efforts at malware detection has resulted in signature-based and heuristic techniques including static analysis, dynamic and hybrid analysis or detection techniques [7], [8], [9] and [10]. However, static analysis techniques are fast, but fails to reveal the behaviour of the malware. In response to this, the use of dynamic analysis techniques such as code analysis, dynamic analysis and memory or forensic analysis has been explored. Nevertheless, malware Authors responded by providing variants of the malware such that the dynamic signature-based scanners are unable to detect [11]. He indicated that, this results in malware variants such as oligormophic (encryption key can be changed instead of encryption where the key is constant) malware, metamorphic (malware changes form but same identity) malware and polymorphic (develop multiples number of descriptors) malware and encryption techniques. These transformed malware are able to outwit even sometimes the dynamic techniques. In response to this, the use of hybrid static and dynamic paradigm has also been explored [12], [13]. Notwithstanding the various efforts at malware detection, the lack of adaptability, memorability, learnability of the signature-based detection remains a problem as malware authors adopt innovative obfuscation techniques such as dead code insertion, data transposition, and others to evade detection. These techniques produce low detection accuracy, high false positive and false negative rates which lead to wrong decision making, poor triaging and poor incident response management [14].

To overcome these problems and to introduce adaptivity to malware detection, the use of machine learning has been proposed and explored [15], [16], [17], [18.

Manuscript received June 5, 2022
Manuscript revised June 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.6.87

However Machine learning is highly if not entirely dependent on features or attributes for training and testing a model. Features are the attributes of a given instance that is being modelled. Malware features therefore are the attributes that can be used to train a machine learning model to be able to generalized to unseen malware [19], [20].

Thus, for efficient and effective machine learning, the need for relevant malware dataset with good features is of essence. Malware dataset with redundant and irrelevant features is computationally time consuming, costly, and lead to poor judgement of the predictive model during training and testing phases. In addition, very large features constructed during the feature construction or generation stage will ultimately result in increases in the volumes of dataset requiring very high storage space. The larger the features the higher the chances that the features are correlated and duplicated which negatively affect the model performance [21]. Thus, one way of pruning features to reduce columns and only present relevant features with high variances is feature reduction or dimensionality reduction or feature extraction.

Feature transformation techniques or dimensionality reduction techniques produce new columns using some hiding structures inherent in the original dataset that produces entirely a new dataset and new columns using algorithms. These new columns contain enough variances of the data. Thus, by eliminating the irrelevant features, the new columns contain so much semblance that they are able to accurately predict malware from benign ware.

Feature transformation or dimensionality reduction techniques can be data-dependent (e.g. PCA), data-independent (e.g. Hashing trick or features hashing) and Graph-based (e.g. UMAP) [22]. Thus, this paper investigated the performance of these three dimensionality reduction techniques for malware detection and or prediction using a malware dataset.

To achieve this aim, we set the following objectives:

a. To determine the predictive performance of data-dependent(PCA) feature transformation or dimensionality reduction technique on malware dataset
b. To determine the predictive performance of data-independent(HT) feature transformation or dimensionality reduction technique on malware dataset
c. To determine the predictive performance of graph-based (UMAP) feature transformation or dimensionality reduction on malware dataset.
d. Propose an improved and efficient feature transformation or dimensionality reduction technique and a machine-learning algorithm for malware prediction.

The paper contributed to knowledge by highlighting the performance of PCA, HT, and UMAP feature reduction techniques on malware dataset. Demonstrated that data-independent Feature reduction technique such as HT produced high prediction accuracy and low false positive rate in detecting malware.

We structured the rest of the paper as follows: section 2 discuss related literature, section 3, discusses feature dimensionality reduction of transformation techniques, section 4, is the methodology of the study, section 5, presents the results and discussion, section 6, is the conclusion of the study.

2. Related Literature

In recent times, the use of features and the need for feature transformation or feature dimensionality reduction techniques has gained traction in both industry and academia especially in malware detection using machine learning. Several studies involving the use of feature transformation or dimensionality reduction has been explored:

[23]explored malignant masses found in mammograms using feature extraction or dimensionality reduction technique. They used Gray Level Co-occurrence Matrix (GLCM) for texture feature extraction based on three hybrid techniques. They used Support Vector Machine (SVM) and two datasets images for their approach. The y concluded that their proposed method performed better over the multi-resolution feature extraction or transformation techniques in based on the number of extracted features and a superior performance on using the Area under the Curve (AUC) performance measure.

[24] conducted an experiment on emotion based recognition task using principal component (PCA) and t-statistical to reduce features or dimensionality of the extracted features from EEG emotional signals. They applied their proposed technique on a dataset known as SJTU and classified using four algorithms; SVM, k-Nearest Neighbor(KNN), Artificial Neural Networks(ANN), Linear Discriminant Analysis(LDA). They concluded that their proposed method outperformed the other emotion recognition methods.

Similarly, [25] espoused a method called Spectral Segmentation and Integration (SSI) as a feature reduction technique for hyperspectral images. They used the mean operator for integration to extract new features in minimal number as compared with the original. PSO algorithm was used to merge spectral curves in the signature to reduce the dimensionality of the images and increase accuracy. They contend that, their approach outperformed NWFE, DAFE, PCA, SELD, BCC, CBFE and PCA.

[26] proposed a system for the reduction of features from fault-based vibration signals using three feature extraction methods; Fourier frequency transform Spectrum(FTFS), Local mean decomposition(LMD) and envelopment analysis. Their results demonstrate that, the envelop analysis and LMD lead to extraction of cancerous

and non-cancerous breast tumors. The CNN and SVM used lead to efficient classification of ultrasound images with high sensitivity and model accuracy.

[27] conducted studies on fault detection using discriminative graph regularized auto-encoder (DGAE) to develop a feature extraction technique aimed at manually developing features. They integrated a neural network with graph to learn and used NN as a classifier. They contend that in comparison with other fault detection methods, their approach archived better performance.

[28] considered empirical mode decomposition(EMD) to propose two feature extraction/transformation methods of mammogram images by dividing the image into group with different frequency. The first method was based on bi-dimensional empirical mode decomposition (BEMD) and extracted features through GLCM and gray level matrix features. The second was a modified version of the earlier one (MBEMD). Two classifiers SVM and LDA were used and archived a study performance.

However, notwithstanding the efforts at using feature extraction or the dimensionality reduction techniques in extant literature, the need for a better performing feature reduction technique for malware detection needs consideration to ensure efficient performance. Thus, we explore data-dependent dimensionality reduction technique (PCA), data-independent dimensionality reduction (feature Hashing Trick) and graph- based technique (UMAP).

Methodology

We obtained a malware dataset from VirusTotal.com and Malware.com for the experiment. it was made up of 30 input features and with 569 observations. The data was prepared by removing duplicates, errors, mission values, data type conversion and so on. It was randomized to disorder any form of pattern in the data and visualized to see inherent relationships if any such as imbalances, and being bias alert. The data was also divided in the usual train-test split process for the experiment. We trained a SVM model with the original features and optimized for the three techniques. We used the three techniques to reduce the features into 16 for each transformation or reduction technique. The model (SVM) was tested on PCA, Hashing Trick or feature hashing and UMAP respectively and their performance compared.

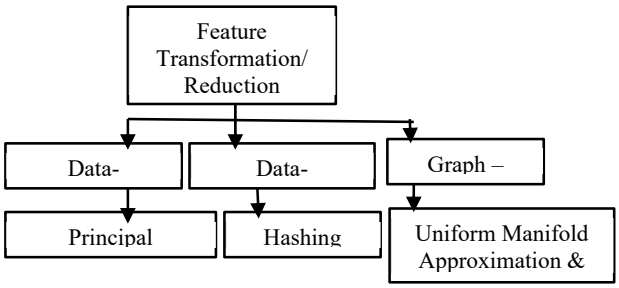


Figure 1. Diagram of Dimensionality Reduction Techniques

3.1. Data-Dependent Method (PCA)

These try to construct a projection function in a form of a matrix for the given dataset. This requires the entire dataset or almost part of it. The problem with these methods or techniques is that keeping the entire data in memory is highly impractical[20]. For the purpose of this study, we used the PCA. It takes given datasets with many correlated features and try to project them into an lower coordinate system with fewer correlation. The resulting features that are uncorrelated as called Principal Components and they capture many of the variances that is needed for prediction. If we start with data of matrix of size n x d, with n= number of observation, d=number of original features, this matrix is projected into another matrix of size n x k, where k less than d. this new matrix give rise to a new columns has the capacity to maximize the variance in the dataset. Hence, the first principal component does most of the data variations, followed by the second etc.

Figure 2. Concept of PCA

PCA is one of the most popular unsupervised and straightforward method that seek to reduce the space dimensionality by finding and transposing from the higher dimensional space into a lower dimension.

3.2. Data-Independent Technique (Hashing Trick)

These are mainly using the principle of random projections and are good at evolving data streams because they generate projection matrices or functions that transform data into low dimensional feature space from independent input data.

When too many features are generated for an algorithm, it tend to affect the performance of the model due to duplicates and other noise that do not actually add to the predictive value of the model. Hence such features need to be pruned [20], [17]. We used the hashing technique, which is also called feature hashing to reduce the features obtained into manageable number for experimental purposes. This is a technique in which large features are compressed to the required features by assigning required features that is enough for the machine learner. The system may lead to collision as more features are compressed. However, the degradation from this seems insignificant. The algorithm for the feature hashing technique is as shown in the table.

Table 1. Algorithm of Feature Hashing

```
** create a vector array of zeros and new features
Def apply feature hashing (feature_vector_size=1000)
** iterate over all feature set in the feature dictionary
For key in feature_dict:
** get the index into the new feature array
Array_index=hash (key) %vector_size
**add the value of the feature to the new feature array
**at the index we got using the hashing trick.
New_features([array_index]+=feature_dict[key]
Return_new_features
```

3.3. Graph-Based Methods

They are also data dependent techniques or methods that works by constructing a graph that is based on similarity of instance and function on this representation. Uniform Manifold Approximation and Projection (UMAP) is an example [29]. They suggest that it is new method that is similar to t-SNE, which is a current discovery with strong mathematical base. It start by construction open balls in the instances and building complexes. To obtain the space reduction, a representation in a lower-space that closely depicts the topological structure found in the original feature space and with fuzzy representation. It has better visualization, and faster than t-SNE. A newer version using batch-incremental strategy has been proposed [22].

3.4. Measures of Performance.

To be able to evaluate any model in and experiment, there is need for measure of performance to objectively ascertain how the proposed models perform on the task [17]. For the purpose of this study, we measured the accuracy and False Positive (FP).

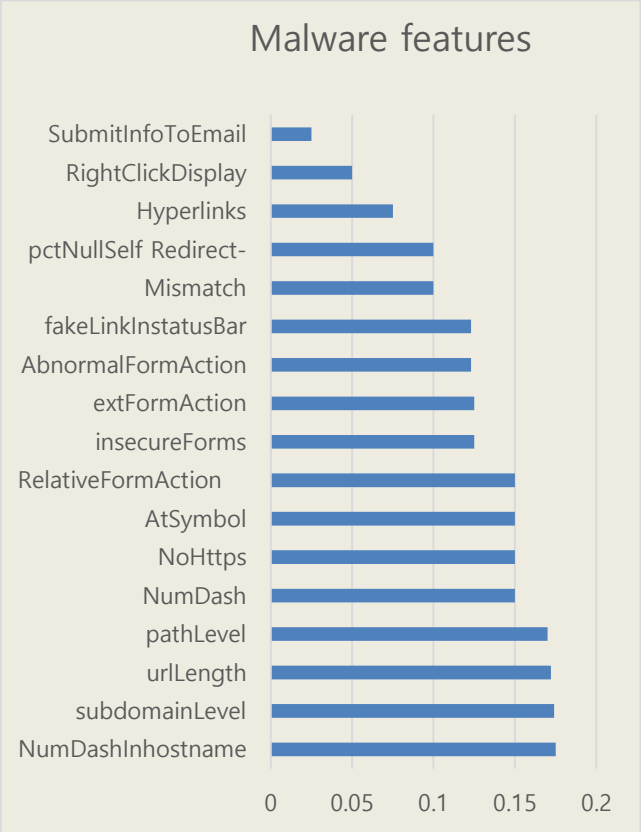
Accuracy= TP+TN/N……1

False Positive Rate (FPR) = FP/ (TN+FP)……..(2)

3.4.1. Features of the Malware Dataset

As discussed earlier, the dataset was made up of 30 features as the main features of the dataset. We reduced this features into 16 using the various feature transformation or dimensionality reduction strategy. The relevant 16 features and their contribution in prediction is shown in table k and figure x respectively. The three feature reduction methods were then applied for prediction and the results of their performance recorded.

Features	*Data type*	*Description*
NumDashInhostname	*numeric*	*looks for number of dash characteristic in hostname*
subdomainLevel	*numeric*	*checks the number of siubdomain levels*
urlLength	*numeric*	*the length of url*
pathLevel	*numeric*	*checks the depth of the url*
NumDash	*numeric*	*number of characters with dash*
NoHttps	*numeric*	*checks for the presence of https in the url*
AtSymbol	*Boolean*	*checks for @ sign in the url*
RelativeFormAction	*Boolean*	*check if action has relative url*
insecureForms	*Boolean*	*check if form action for content without https protocol*
extFormAction	*Boolean*	*check the action form for external url*
AbnormalFormAction	*Boolean*	*checks for abnormal url form actions*
fakeLinkInstatusBar	*Boolean*	*check html source code for for javascript commands for mouseover to display fake url in the status bar*
frequentDomainName-	*Boolean*	*checks*

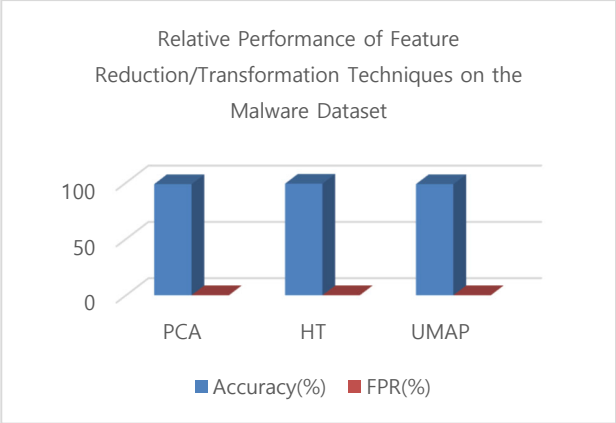


4. Results and Discussion of the study

This section details the experimental result of the study and discussion how the findings relates to literature and the implications of the study. Thus, we present the performance of the SVM on the data-dependent dimensionality reduction technique (PCA), Data-independent dimensionality reduction technique (Feature hashing technique or feature hashing trick, and graph-based feature reduction or transformation technique (UMAP). The result of the three techniques are as shown in table MMM.

Feature Transformation Technique	No. of features	Accuracy	False Positive
PCA	16	0.987	0.017
Hashing Trick (HT)	16	0.991	0.012

From table k which compares the relative performance of the PCA, HT and UMAP feature transformation or reduction techniques, the results show that all the techniques performs relatively well with the dataset. However, the HT which is a data-independent reduction technique performed slightly higher in both performance measures; 99.1% accuracy and 1.2% false Positive rate while the other two showed no variation in performance. This is in line with [22] who posited that, both graph-based and data-dependent feature reduction techniques are equivalent and that they both handle features and dataset in similar fashion. Also, the relatively high performance of the SVM model on all the techniques might demonstrate the fact the, a well extracted features and relevant reduction techniques produces columns that captures the key variances in the data and leads to efficiency of a model. This view is supported by [19][22],



From figure 4, it shown clearly that all feature showed relative excellent performance. However, the HT feature produced superior results in both accuracy at 99.1% and FPR 1.2% with the other two performing the same at accuracy of 98.7% and FPR of 1.7%.

Thus, in summary, the results is in line with the argument that, data-dependent feature transformation methods such as PCA are the classical data reduction techniques and are data-dependent and therefore produce good accuracy in combination with other data mining streams. On the data-independent reduction techniques are naturally adaptable to the evolving environment, do not suffer from the problem of scalability and are extremely faster as compared to other reduction techniques [22], [21].

5. Conclusion

The paper explored the relative performance of PCA, HT, and UMAP feature transformation or Dimensionality Reduction techniques on Malware dataset. The features of the dataset were reduced to 16 features and SVM applied to determine the predictive capacity of each of the three techniques on the dataset. The experimental results of the study shows that, PCA, HT and UMAP all performed relatively well on the dataset. However, the HT, which is a data-independent reduction method produce superior

results compared with the other two. The paper concluded that higher feature dimensionality influences the performance of models such as computational time, memory constraints and the predictive accuracy of a model. Thus, knowing which dimensionality reduction technique, yields higher predictive accuracy with extremely low false positives will lead to improved malware detection and feature engineering. Consequently, the use of the Data-independent technique such as HT presents a huge potential for feature engineering in general and feature transformation or dimensionality reduction methods in particular in malware detection. However, the experimental results good performance by all the techniques, this might be due to the smaller data size. Future effort will employ the approach and apply a larger data size with increased reduction techniques to observe the phenomenon.

Acknowledgment

This work did not receive any external funding. However, personal funds and the support of my supervisors was used for the work

References

[1] AV-Test Institute. Malware statistics. Retrieved online at www.av-test institute, 20th November, 2021

[2] Bassett, G., Hylender, C.D., Langloise, P., Pinto, A., and Widup, S. (DBIR Team). Verizon Data Breach Investigations Report 2021. Results and Analysis. https://www.verizon.com/business/resources/report. Retrieved 26/12/21

[3] Fatima Salahdine & Naima Kaabouch . Social Engineering Attacks: A survey. Future Internet, MDPI, 2018 .doi.10.3390/fi11040089

[4] Hussein Aldawood, Geofrey Skinner. Reviewing cyber security Social Engineering Training and Awareness Programs-Pitfalls and Ongoing Issues. Future Internet, 2019. MDPI. https://doi: 10.3390/fi11030073.

[5] Albladi, S.M., Weir, R.S. Predicting individuals' vulnerability to social engineering in social entworks. Cybersecurity, springer open, 2020. https://doi.org/10.1186/s42400=020-00047-5

[6] Ana Kovacevic, Sonja, D. Radenkovit. SAWIT: Security Awareness Improvement Tool in Workplace. Applied Sciences, 2020. MDPI. Doi: 10.3390/app10093065

[7] Yun, Zhou, Peichao Wang . An ensemble learning approach for xss attack detection with domain knowledge and threat intelligence. Computers and security. Elsevier, 2019.

[8] Anusha Damodaran,Fabil Di Troia, Corrado, AaronVisagio, Thomas, H. Austin and Mark Stamp. A comparison of static dynamic and hybrid analysis for malware detection. Com Virol Tech, 2015. doi10.1007/s11416-015-026-z.

[9] Hemant Dhamija, Ajay, K. Dhamija . Malware detection using Machine Learning Classification Algorithms. International Journal of Computational Intelligence Research. Research India Publications, 2020

[10] Monnappa, K. A . Learning malware analysis: explore the concepts, tools and the techniques. Birmingham, Mumbai., 2018.

[11] Jagsir Singh and Jaswinder Singh. . Challenges of malware Analysis. Obfuscation Techniques. International journal of information security science, vol.7. no. 3, 2018

[12] S. Valluripally, A. Gulhane, R. Mittra, K.A. Hoque and C. Prasad, "Attack Tress for security and privacy in social virtual reality environment, in proceedings of the IEEE Annual consumer communication and networking conference, Las Vegas, NV, USA, January, 2020.

[13] N.Wongwiwachai, P. Pongkham, and K. Sripanidkulchai. "Comprehensive detection of vulnerable personal information leaks in android applications." In proceedings of IEEE Conference on computer communications workshop,Toronto, Canada, July, 2020.

[14] Chris Pace and Ahlberg. The threat intelligence Handbook: A practical Guide for security Teams to Unlocking the Power of Intellignece. Cyber Edge Group Ltd, 2020

[15] A. Shalaginov, S. Banin, A. Dehghantanha and K. Franke. Machine Learning Aided Malware Analysis: A survey and a tutorial. Springer international publishing AG, Part of springer nature, 2018.

[16] Neil, Balram,; George, Hsieh and Christian, Mcfall, . Static malware analysis using machine learning algorithms on APT1 dataset with string and PE features. International conference on computational science and computational intelligence (CSCI), 2019.

[17] Saxe J, and H. Sanders. Malware data science: attack detection and attribution. Starch press, 2018

[18] Cho, Do ,Xuan; Hoa Dinh Nguyen , Tisenko, Victor, Nikolaevich. (IJACSA) International Journal of Advance Computer Science and Application.Vol 11, No. 1, 2020.

[19] Pablo Duboue . The art of feature engineering. Essentials for machine learning. Textualization software Ltd, 2020.

[20] Sinan Ozdemir and Divya Susarla. Feature engineering made easy. Identifying unique features from your dataset in order to build a powerful machine learning systems. Birminham, Mumbai, 2018.

[21] T. Admassu, R. L. Tulasi, V. Elanangai, N. K. Kumar. Exploring the perfoemance of feature selection method using breast cancer dataset. Indian journal of electrical and computer science, vol. 25, no. 1, January, 2022, pp.232-237.

[22] Bahri et al. 2020. Efficient Batch-incremental classification using umap for evolving data streams. In IDA , pp, 4053, 2020.

[23] M.A. Berbar " hybrid methods for feature extraction for breast masses classification, "Egyptian informatics journal, Vol.19, no.1.pp.63-73.2018.

[24] M.A. Rahman, M.F. Hossain, and R. Ahmmed, " employing PCA and t-statisticcal approach for feature extraction and classification from multichannel EEG signal". Egyptian Informatics Journal, vol. 21,no.1,pp.23-35, 2020

[25] Yun, Zhou & Peichao Wang. An ensemble learning approach for XSS attack detection with domain Knowledge and threat intelligence. Elsevier. Computers & Security, 2019.

[26] C. Chu, Z. Zuo-xi, K. Xin-Rong and G. Yun-Zhi " The research of machinery fault feature extraction methods based on vibration signal, IFAC-Papersonline, Vol.51, no.17, pp.346-352, 2018.

[27] Y.Li, Y.Chai, H.Zhou and H.Yin, " A novel feature extraction method based on discriminative graph regularized autoencoder for fault diagnosis", IFAC-Paperonline, vol.52, no,24.pp. 272-277, 2019.
[28] V. Nagarajan, E.C.Britto, and S.M.Veeraputhiran, " feature extraction based on empirical mode decomposition for automatic mass classification of mammogram images," medicine in Novel technology and devices, vol1, p.100004,2019.
[29] L. McInns, Healy, J., Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. 2018.

Azaabi Cletus (Corresponding Author) had a B.ed Computer science from the university of Cape Coast (2007), PGD (Management Information Systems), GIMPA (2010), MSc. Information Technology, KNUST (2016). Currently a PhD. Student, UENR, all in Ghana. Azaabi Cletus is Lecturer at St. John Boscos College of Education. Research interest is in cybersecurity.
Email: cleinhim@yahoo.com

Alex Opoku(PhD) is a Senior Lecturer in the Department of Mathematics, University of Energy and Natural Resources, Sunyani, Ghana. He is currently the Head of Quality Assurance of the university and has many publications in his name.
Email: Alex.opoku@uenr.edu.gh

Benjamin Asubam Weyori(PhD) is a senior Lecturer and the immediate Past Head of department of Computer Science and Informatics at the university of Energy and Natural Resources, Sunyani, Ghana. He has many publications in his name in peer-reviewed journals.
Email: Benjamin.weyori@uenr.edu.gh