ASHESI

ASHESI UNIVERSITY COLLEGE

Analysis of Data Mining Techniques and Algorithms for Healthcare Application Using Cervical Cancer as a Case Study

UNDERGRADUATE THESIS

B.Sc Computer Science

Cynthia Naela Priscilia Ngaffo Gouanfo

April 2018

ASHESI UNIVERSITY COLLEGE

Analysis of Data Mining Techniques and Algorithms for Healthcare Application Using Cervical Cancer as a Case Study

UNDERGRADUATE THESIS

Undergraduate Thesis submitted to the Department of Computer Science, Ashesi University College in partial fulfilment of the requirements for the award of Bachelor of Science degree in Computer Science

Cynthia Naela Priscilia Ngaffo Gouanfo

April 2018

DECLARATION

I hereby declare that this thesis is the result of my own original work and that no part of it has been presented for another degree in this university or elsewhere.

Candidate's Signature:

…………………………………………………………………………………………

Candidate's Name:

…………………………………………………………………………………………

Date:

…………………………………………………………………………………………

I hereby declare that preparation and presentation of this thesis were supervised in accordance with the guidelines on supervision of thesis laid down by Ashesi University College.

Supervisor's Signature:

…………………………………………………………………………………………

Supervisor's Name:

…………………………………………………………………………………………

Date:

…………………………………………………………………………………………

Acknowledgements

I would like to thank my supervisor, Dr. Elena Rosca who aided me to decide this area of research, and whose encouragement and expert advice helped me undertake this project. I would like to extend my sincerest gratitude to Dr. Lorenzo Torresani for his insightful feedback and in-depth explanation of how different classifiers could be used in my research.

Finally, many thanks go to Prof. Patrick Petignat and his team of researchers at the Faculty of Medicine, University of Geneva, Switzerland, for providing the cervical images that were used to train and test the classifiers implemented in this project.

Abstract

Cervical cancer is the most common cause of cancer among African women. It is a preventable disease and can be treated if identified at early stages. Given the lack of adequate health care services and the costly nature of colposcopies in Africa, it is difficult to get an early diagnosis. The development of smartphone-based diagnostic tools like MobileODT – with which pictures of the cervix are taken and sent to doctors for diagnosis – promises to address the expensive nature of colposcopy and Pap test; still, the diagnosis of these images is prone to human errors. This project aimed to recommend an algorithm that best classifies cervical images into cancerous and non-cancerous, in order to aid medical officials to give a better diagnosis. K-Nearest Neighbour (KNN), Convolutional Neural Network (CNN) and Support Vector Machine (SVM) were analysed and compared based on their classification accuracy, sensitivity and specificity and how these results varied after applying Principal Component Analysis (PCA) on the dataset. KNN, CNN and SVM models obtained classification accuracies of 68.75%, 83.3% and 66.37% respectively while PCA-KNN and PCA-SVM models had classification accuracies of 78.12% and 62.7% respectively.

Table of Contents

List of Tables

List of Figures

CHAPTER 1: INTRODUCTION

1.1 Background: Data mining

The field of Knowledge Discovery in Database (KDD) and Data mining has been one that keeps rising, especially since data mining gained popularity in the 90s (Coenen, 2011). It has evolved such that it is difficult to imagine a future without it because of the promising new directions it holds for the future. When we talk about data mining, we often refer to the processes involved, the techniques, the algorithms surrounding it and its applications. These terms are sometimes used interchangeably depending on the context in which they are used but they mean different things though they are closely related.

Data mining processes refer to the stages or steps taken to discover knowledge from datasets, for example the clustering of data, while a data mining technique is a general approach or guidance to solving a data mining problem. Decision trees, clustering classification are some examples of data mining techniques. A data mining algorithm on the other hand is a well-defined approach of how to correctly produce an output. Often times, implementing an algorithm requires combining a number of data mining techniques. These algorithms are continuously applied in the healthcare sector to efficiently solve problems, for example, the Fourier transform used in medical imaging, RSA (the encryption algorithm) used to transmit medical digital data, BLAST (Basic Local Alignment Search Tool) a search algorithm used in the analysis of gene and protein sequences (Thamilselvana & Sathiaseelan, 2015).

1.2 Background: Healthcare Industry

Healthcare involves activities like diagnosis, treatment and prevention of disease, injury and other physical and mental impairments in humans. The healthcare industry can be viewed as a place with rich information as they produce enormous volumes of data. However, these large volumes of data can most times become useless or underused if not processed the right way. Data mining provides an avenue for a search for new and valuable information from these large volumes of data. Data mining in healthcare is being utilized largely to predict the occurrence of different diseases in addition to helping doctors diagnose infections in order to aid them in making their clinical decision.

1.2.1 Cervical Cancer: Diagnoses and Causes

Many smartphone-based diagnostic tools are being developed to ease the diagnosis process of individuals, especially those at long distances from hospitals. An example of such tool is MobileODT developed by an Israeli company to aid in cervical cancer screening (Champlin, Bell, & Schocken 2017). Cervical cancer is a type of cancer that affects cervix, transmitted by a virus named human papillomavirus (HPV). It begins when the cells at the lining of the cervix (i.e. the lower part of the uterus) multiply uncontrollably. Diagnoses of this disease are done through a Pap test or/and a colposcopy. A Pap test involves swabbing the cervix and sending the swab for laboratory examinations in order to detect pre-cancer or cancer tissues. Most times, a colposcopy is done after abnormal Pap test results are obtained. Colposcopy involves the patient lying in a dorsal lithotomy position and doctor examining the cervix using a magnifying instrument (The American Cancer Society, 2017).

Some of the contributing factors to cervical cancer include: obesity, lack of vegetables and fruits in one's diet, oral contraceptives, smoking, weakened the immune system, etc. It is the fifth most common cancer (after lung cancer) that affects women worldwide. The World Health Organization suggests that "most sexually active … women will be infected at some point with HPV, [with some even being] repeatedly infected" (WHO, 2017). The American cancer society estimates that there will be about 12,820 new diagnoses and about 4,210 deaths by the end of 2017 in the United States of America (The American Cancer Society, 2017).

Even with these figures, cervical cancer is a relatively unusual disease in the US. In previous years, there have been 7 in 100,000 women diagnosed each year, amounting to about 3 deaths in North America, compared 34 in 100,000 diagnoses and 23 deaths each year in Africa (WHO, 2017). These statistics are just those for known cases on the African continent; many cases are still unknown as they have never been diagnosed. An explanation as to why the US has lower death rates relative to Africa is that they have been able to maintain frequent screening such as Pap smear tests to diagnose the disease at its early stage. Hence, the US has been able to better contain the disease unlike most African countries where Pap test and colposcopies are expensive, and access to cancer treatment is very limited (WHO, 2017).

1.3 Problem and Aim of Research

With the introduction of tools like MobileODT in Africa, specifically Kenya, gynaecologists are now able to take a magnified image of a woman's cervix by attaching an enhanced visual assessment (EVA) to the smartphone (Champlin, Bell, & Schocken 2017). The challenge here however is ensuring that health workers make accurate diagnoses from these images. As such, this research attempts to recommend an algorithm that efficiently classifies cervical cancer images taken with phones in order to ease the diagnosis process and reduce human error involved, hence, making it more efficient. To achieve this, three main algorithms, K-Nearest Neighbor (KNN), Convolutional Neural Networks (CNN), and Support Vector

Machines (SVM), were compared based on their classification accuracy, sensitivity, and specificity results and how they changed with the introduction of PCA on the image dataset.

CHAPTER 2: LITERATURE REVIEW

Data Mining is the "nontrivial extraction of implicit, previously unknown, and potentially useful information from data" (Bhatnagar, Gupta, & Wasan, 2001). Though there has been intense research in the fields of Knowledge Discovery in Database (KDD) and Data mining, there are several practical and theoretical issues in these fields. In a study by Bhatnagar *et. al.*, the authors commented that often times, the KDD process is executed "as a series of disconnected and disjoint steps" (Bhatnagar *et al.*, 2001) where the user has to be continuously involved at every step of the process. This therefore brings up the need to develop a complete KDD system which requires user intervention only at first and last step of the process.

In addition to this, the absence of a "windowing" data mining technique that enables data miners to monitor at predefined intermediate steps. This enables faster corrective action in case the results are not up to expectation. Other issues include the slowness of data mining algorithms such as incremental algorithms that are used on ever-growing databases but whose current speed will be unacceptable a few years (Bhatnagar *et al*., 2001). Due to such issues, it is important that programmers propose algorithms with efficient and reasonable runtimes.

2.1 Data mining techniques and Algorithms

In 2011, Coenen mentioned even with the availability of data mining processes, the present concern revolves around the effective collection and processing of data so that the different techniques could be applied efficiently. Additionally, the post-processing stage of the outcome is also a worry, how to effectively visualize the end results so that it is well comprehended by the concerned audience in the absence of an expert. Also, data mining still necessitates the development of efficient methods and algorithms to handle even larger datasets with greater variety (Coenen, 2011).

Moving on to data mining algorithms, Wu *et al.* attempted to provide 10 best-known algorithms as identified by the IEEE International Conference on Data Mining (ICDM) in December 2006. These algorithms include C4.5, k-Means, SVM, Apriori, EM, PageRank, AdaBoost, KNN, Naive Bayes, and CART. These 10 algorithms cover the data mining techniques of classification, clustering, statistical learning, association analysis, and link mining. In order to arrive at this ranking, the authors used a quantitative approach with included 5 steps. Firstly, they asked the ACM, KDD Innovation Award and IEEE ICDM Research Contributions Award winners to propose 10 best-known algorithms in data mining with relevant publications mentioning them. A total of 50 publications were later analysed, then narrowed to 18 publications which were organized in 10 topics: association analysis, classification, clustering, statistical learning, bagging and boosting, sequential patterns, integrated mining, rough sets, link mining, and graph mining (Wu *et al*., 2008).

2.1.1 Datamining Algorithms in Healthcare

Data mining technology has been adapted to different fields in the world. Various organisations ranging from business to medical have applied its techniques and algorithms and found excellent results. Still, it is important to avoid the blind application of data mining which can lead to the discovery of meaningless or misleading patterns. With respect to the medical field, Brusic and Zeleznikow draw our attention KDD and data mining in biological databases. In doing this, the authors present a different idea on the two fields. They mention that there is a difference between KDD and data mining in that data mining is a simple step (a core step) in KDD which in itself is a process (Brusic & Zeleznikow, 1999).

The KDD process involves 10 steps: learning the application domain, creating a target data set, data cleaning and pre-processing, data reduction, choosing the function of data mining, choosing the data mining algorithm, data mining, and interpretation, using discovered knowledge and evaluating the KDD purpose (Brusic & Zeleznikow, 1999). Nevertheless, it seems as if in the present age, the terminology Knowledge Discovery in Database has somewhat been vacated and that of Data mining has come to stand as what KDD used to mean. The researchers went ahead to state that it is important to have some important considerations when analyzing biological datasets such as the growth and complexity of genomic data.

In 2012, Kharya pointed out some of the data mining techniques frequently used in the medical domain. The research threw light into data mining techniques that are used for various diseases. The techniques featured in the research are classification, clustering, association and neural networks. In addition to these techniques, the data mining algorithms mostly used include Decision Tree Algorithms, BBN Bayesian networks, Bayesian Ying Yang (BYY), Genetic Algorithm, K-Nearest Neighbour, Naïve Bayesian, Apriori Algorithm, and Support Vector Machine. These algorithms were used on datasets for diseases such as conventional pathology, coronary heart, lymphoma, cancer and diabetes.

Furthermore, another study surveyed the development of data mining techniques using articles ranging from 2005 until 2015. The timeframe was chosen based on when there have been widespread of data mining techniques being used in the healthcare. After comparing the two data mining models, Predictive and Descriptive, Predictive was the most used model by 94% (Jothi, Rashid, & Husain, 2015). The authors were in accord with research shown above on techniques used in the medical field. Among the data mining techniques developed in recent years, the highlight was made on the data mining methods of generalization, characterization, classification, clustering, association, evolution, pattern matching, data visualization and meta-rule guided mining.

Once again, attention is drawn to the particularity of healthcare data sets, that is, they are highly imbalanced and contain a lot of missing values. Wasan *et al.* in their research agree with this observation as they describe medical data as noisy and heterogeneous with missing values. As a result, there's a need for the development of hybrid models to resolve these issues so as to obtain the highest accuracy.

2.1.2 Image Mining Algorithms in Healthcare

Image mining is significantly becoming an area of increasing demand for developing real-world vision systems. In the medical field for example, the use of images taken by smartphones has become an easier way of getting diagnosed with a disease after it has been sent to an expert. As these images are stored, their databases increase by the minute and it is very important to mine these images to get meaningful information. Various algorithms have been proposed for image mining, and this section seeks to explore these techniques as proposed by different kinds of literature.

Thamilselvan and Sathiaseelan used a comparative approach to study image classification algorithms. They considered two indications; classification accuracy and the kappa coefficient. The kappa coefficient ranges from 0 to 1 and measures how true a classification result is with respect to randomly assigned values. When the kappa coefficient equals to 0, there is no relation between the image of reference and the classified one, while classification accuracy refers to the probability that image is correctly classified (Thamilselvana & Sathiaseelan, 2015).

From the study, the Support Vector Machine (SVM) showed an accuracy of 97.15 % while K-Nearest Neighbour (KNN) showed 96.94% accuracy in image classification. From this, it would be safe to conclude that SVM and KNN show higher accuracy in image classification and these chances are more likely to remain unchanged when introduced in a hybrid algorithm. The authors went ahead to make a bold conclusion that "SVM and KNN are the most predominant data mining algorithms in image classification" (Thamilselvana & Sathiaseelan, 2015).

In an attempt to extend their study, the authors considered algorithms not as individuals, but as hybrids. By their study, classification involves two fundamental stages; effective representation of an image and error-free classification of the image. The authors did a brilliant job as they proposed five hybrid algorithms for image mining. These hybrid algorithms are a combination of two or more existing algorithms for optimized results. These algorithms were modelled such that they use the two fundamental steps of image classification as proposed in the study, as well as the training and testing processing.

The algorithms include: Genetic Algorithm and Support Vector Machine (GA- SVM), Extreme K-Means Algorithm and Effective Extreme Learning Machine (EKM-EELM), Naive Bayes and Support Vector Machine (NB-SVM), Decision Tree and Naive Bayes (DT-NB), and Support Vector Machine and Classification regression tree (SVM-CART) (Thamilselvana, 2016). When the algorithms' performances were compared, NB-SVM had the highest accuracy percentage of 98% followed by EKM-EELM, DT-NB, GA- SVM and SVM-CART with 90%, 90%, 89% and 84% respectively. Nevertheless, it would have been better if these performances were tested on the same database.

While NB-SVM was tested on Retinal Image Dataset, EKM-EELM, DT-NB, GA- SVM and SVM-CART were tested on Face Image, Breast cancer, Cancer Image and Human Face Image datasets respectively. Testing all of them on the same dataset would remove the doubt of whether the data set on which it was tested had a role to play the algorithm's accuracy. Nevertheless, it would interesting to look into the hybrid of Genetic Algorithm (GA) and Extreme K-Nearest Algorithm (KNN) since from their previous study, they consistently show a high accuracy percentage with both indicators (90vs90 and 87vs93 for classification accuracy and kappa coefficient respectively).

K-Nearest Neighbor is a well-known algorithm for pattern recognition. However, the KNN algorithm has a high computation cost which stands as a drawback and its performance highly relies on the training dataset. Nevertheless, Suguna and Thanushkodi found an excellent way of producing a hybrid algorithm of Genetic Algorithm and Extreme K-Means Algorithm in 2010, which they called Genetic KNN (GKNN), and achieved some excellent results. The fundamental thought here is that similarities of k-neighbours are computed at each iteration and dataset set is classified based on this after which the accuracy is obtained. The process was repeated a number of times in order to obtain a high accuracy. The performance was then compared with traditional algorithms like KNN and SVM across five different medical data. Their proposed hybrid algorithm is particularly interesting because it not only improves the classification accuracy but also makes KNN simpler.

In a more recent study, Yu *et al.* proposed improving image quality classification using deep learning. Deep learning is a machine learning technique that learns features and tasks directly from all forms of data. They implemented an algorithm (SM-Alexnet-SVM) that combines unsupervised features from saliency map (SM) and supervised features from convolutional neural networks (CNN) Alexnet (Yu *et al.*, 2017). The combination was used for classification and tested in a Support Vector Machine (SVM) to automatically detect and classify images. The algorithm's performance was compared to other methods across a large retinal fundus image dataset.

The proposed algorithm was successful in outperforming other methods with its high accuracy. It had an accuracy of 95.42% compared to HOG-Alexnet-SVM, Alexnet-SVM, HOG-SVM and SM-SVM with accuracies of 94.93%, 94.80%, 89.64 and 80.41% respectively (HOG stands for Histogram of Oriented Gradients). An advantage of this algorithm is that it could be used on a variety of medical images. However, it is hard to determine how efficient this algorithm would be when it comes to determining whether an image belongs to an infected patient or not. This is because in the study, it was used to classify images into high quality vs poor quality images where the parameters used are quite different from that of a picture representing infection or not.

The management of large datasets is of utmost importance in order to effectively interpret their contents. This means that methods to reduce their dimensionality while maintaining the underlining or main characteristics of the datasets are sort out for. Principal Component Analysis (PCA) is one of the oldest and widely methods to perform such task (Jolliffe & Cadima, 2016). The aim of PCA is to minimize the dimensionality of a dataset while preserving the dataset's variability – that is, finding new variables or a set of optimal variables that best represents the original dataset. It serves as a technique for data compression, removal of redundancies and feature extraction.

However, some complications arise when using this technique. An example includes getting a set of variables with different measurement units. It can be adapted to many forms depending on the situation. Some of the applications of PCA include image processing, for example texture analysis, and face verification. Since hybrid algorithms have proven to be very efficient, combining this approach of dimension reduction to a dataset before applying other algorithms discussed above would increase performance.

2.2 Common Gap

Even though most of the algorithms in the literature review have high accuracy percentages, applying them to this study is not feasible because they have a high computational cost, an example is K-Nearest Neighbour (KNN) algorithm. The Deep learning technique requires a large data set. Using Deep learning on a small dataset will not guarantee accurate results.

2.3 Summary

On the whole, data mining is often times used cross fields for predictive or descriptive purposes. In all cases, it is important that process is done thoroughly so that misleading information be not the output. This will be done with the use of data mining techniques as they help you formulate an algorithm. The prominent data mining techniques used across fields are clustering, association and neural networks, and those of data mining algorithms are (but not limited to) Support Vector Machine, K-Nearest Neighbour and Naïve Bayesian.

Most times because of the complex nature of image datasets, hybrid algorithms are often preferred because of their high accuracy. However, with respect to biological datasets, one should not underestimate its heterogeneous, fuzzy and voluminous nature. In retrospect, this research aims at using principal component analysis and proposing a machine learning image mining algorithm through supervised learning. This algorithm is intended to classify cervical cancer images in order to minimize the inaccuracies involved in the diagnosis of patients by doctors. This will be done while keeping into account the fuzzy nature of healthcare dataset and the need to propose a faster algorithm as suggested by the literature.

CHAPTER 3: METHODOLOGY

3.1 Overview of Dataset Origin

The dataset used in this project was obtained from a previous research, cervical cancer screening in low-resource settings: a smartphone image application as an alternative to colposcopy, where the researchers evaluated the quality of cervix images taken with smartphones. The aim of this evaluation was to assess the feasibility and usability of a mobile application for cervical cancer screening as an alternative to colposcopy in low and middle-income countries. The study was carried out in 2015 in Madagascar on 56 HPV-positive women, who underwent the selection process, between the ages of 30 and 65. Pictures of the cervix were taken with a Samsung Galaxy S5 using an application, **Exam**, specifically designed to obtain high-quality images. Consecutive images were captured in a precise sequence – native, VIA, VILI – after participants were positioned in lithotomic position. The picture was later sent to three gynaecologists, experts in colposcopy, to independently assess the pictures' quality. For each picture, they were asked questions like:

According to the images, the cervix appears to you: ------------ Normal or Suspect

Quality of images for diagnosis is: -------------- Sufficient or Insufficient

Quality of images was determined according to criteria: diagnostic utility, sharpness, focus and zoom. From the consensus, 93.27% (194/208) of images were judged as being of good quality. Given the output obtained, we can conclude that the images are a true representation of the human anatomy of the cervix with a high degree of confidence, hence, we can proceed with classifying the pictures.

3.2 Dataset Description

The dataset consists of 125 pictures of HPV-positive women. Each picture has 3 different images of a patient's cervix taken after different applications. **NATIVE** represents the picture of the cervix before any substance was added to it. **VIA** represents the picture of the cervix after application of acid acetic, and **VILI** after the application of Lugol's iodine. The dataset was given with an excel file which contains gynaecologists' aggregate response as they diagnosed the cervical images. A sample image in the dataset can be seen below:

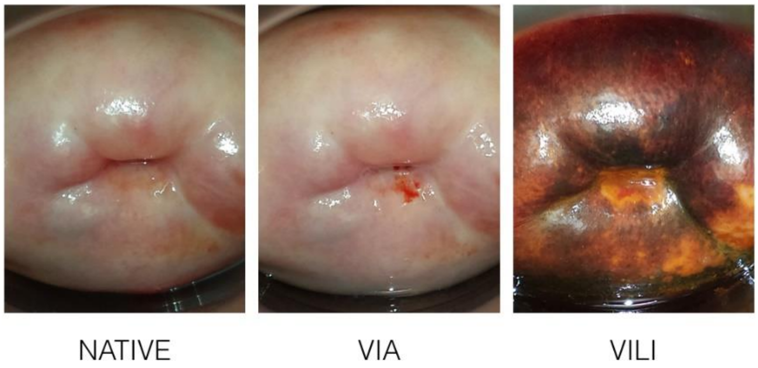


Figure 1. Sample image in the dataset. NATIVE represents a cervical image in its original state, VIA and VILI represents the cervix after the application of acid acetic and Lugol's iodine respectively

Having HPV does not mean you have or will get cervical cancer, as such, even though all the women are HPV-positive, they do not all have cervical cancer. In the dataset, we have 5 categories under which these women were diagnosed and Table 1 shows the number of images for each category that are present in the dataset. **Non-cancerous** represents images that were diagnosed as being void of cancer while **Cancerous** are those that have been infected.

Depending on the extent of abnormal cells founds, dysplasia – abnormal cells on the cervix caused by HPV virus – could be categorized as CIN 1, CIN 2, or CIN 3. **CIN 1** means that about one-third of the cervical cells are abnormal, it does not require treatment because the abnormal cells disappear after a while. **CIN 2** means about two-thirds of the cells are abnormal while **CIN 3** means almost all the cells are abnormal or pre-cancerous but not yet considered a cancer patient.

Type	Non-cancerous (Negative)	CIN1	CIN2	CIN3	Cancerous (Positive)
No of images	101	5	6	8	5
Percentage	80.8%	4%	4.8%	6.4%	4%

Table 1. Distribution of images in the dataset across the different categories

3.3 Selected Algorithms for Classification

Classification is a supervised learning approach in which the computer program learns from the data input given to it and uses this learning to classify new observations. Algorithms that will be analysed for the purpose of classifying cervical images are

i. K – Nearest Neighbour (KNN)
ii. Convolutional Neural Network (CNN)
iii. Support Vector Machines (SVM)

These algorithms were chosen because from the literature, they have been the most prominent algorithms used for image classification and have given high classification accuracies of greater than 70%.

3.4 Performance Metrics for Evaluating Algorithms

Apart from classification accuracy, other performance metrics would be used to evaluate, analyse and compare the performance of the different algorithms. These metrics include sensitivity, and specificity.

Sensitivity: is the ability of the algorithm to correctly detect a cervical cancer image as one (true positive). A non-cervical cancer image that is said by the algorithm to have cervical cancer is referred to as a false positive.

$$Sensitivity = \frac{number\ of\ cervical\ cancer\ images\ detected}{Total\ number\ of\ cervical\ cancer\ images}$$

Specificity: is the ability of the algorithms to correctly detect a non-cervical cancer image as one (true negative). A cervical cancer image that is said by the algorithm to not have cervical cancer is referred to as a false negative.

$$Specificity = \frac{number\ of\ non-cervical\ cancer\ images\ detected}{Total\ number\ of\ non-cervical\ cancer\ images}$$

In addition to an algorithm having a high classification accuracy in general, the aim is that the algorithm also has a high sensitivity and specificity value, as they indicate that most images were correctly classified in their respective categories.

3.5 Analysis

The analysis was done after implementing the different algorithms in **Python**. For this study, we used major libraries: **Scikit-Learn** and **TensorFlow** and a pre-trained model

MobileNet. The analysis was done in major steps: Data Cleaning, Data Processing, PCA and Machine Learning. The diagram below describes the major steps taken to carry out the study.

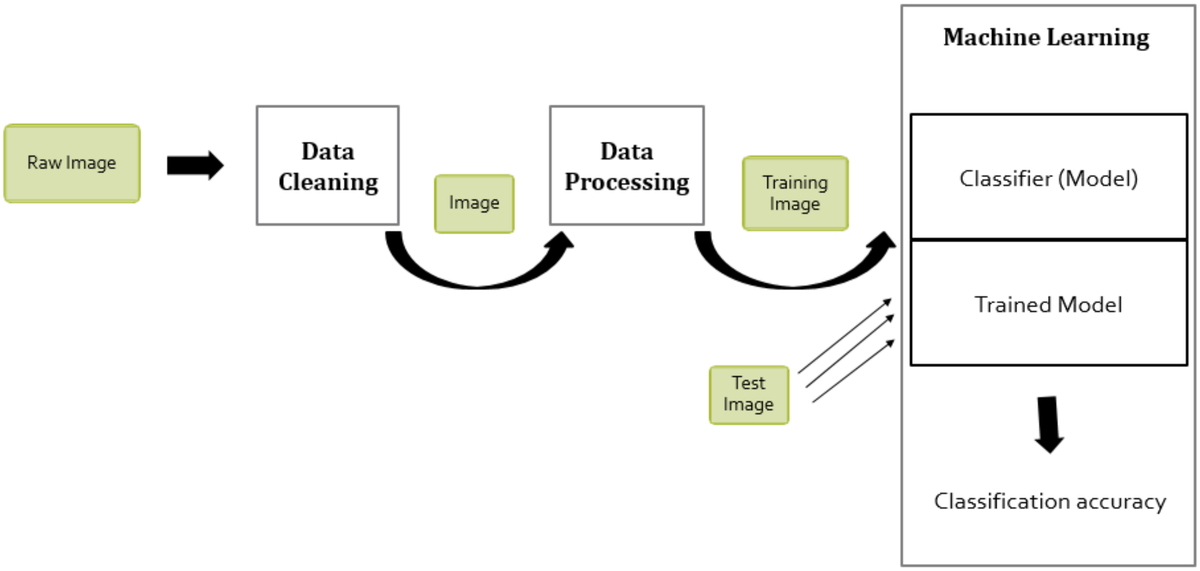


Figure 2. Flow diagram showing steps for an algorithm e.g. KNN

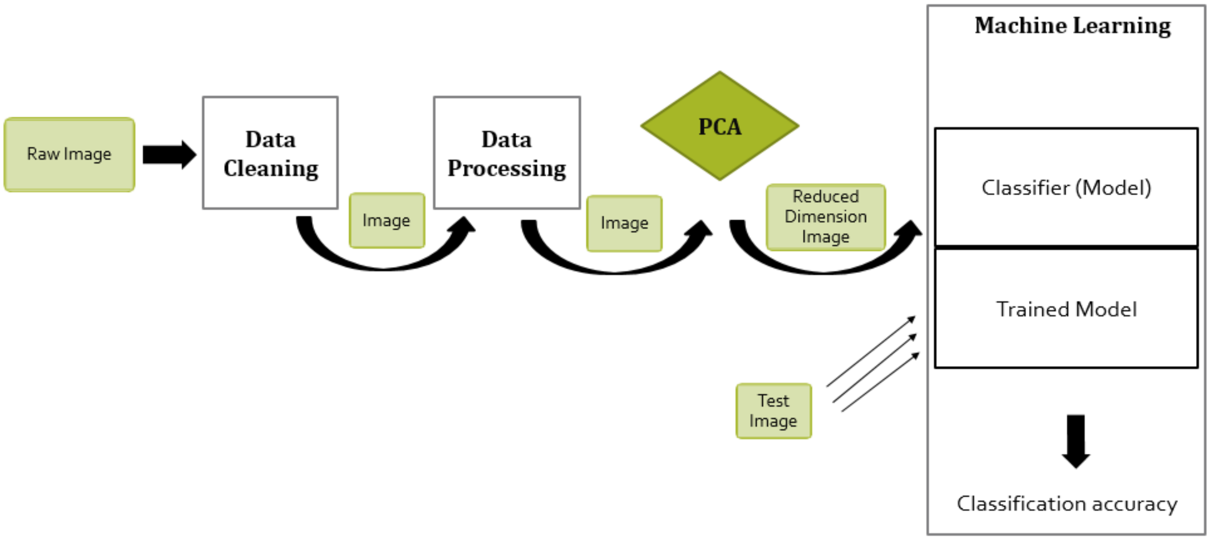


Figure 3. Flow diagram showing steps for a hybrid algorithm e.g. PCA-KNN

3.5.1 Data Cleaning

An initial cleaning is done where the images are cropped because only images with label "NATIVE" are used in this study. This is done with the assumption if a patient is taking a picture of their cervix to send for diagnoses, they would not have acetic acid nor Lugol's iodine at their disposal to obtain the corresponding VIA or VILI image respectively. After cropping the images, they all had different sizes ranging from *231* x *235* to *313* x *398* pixels.

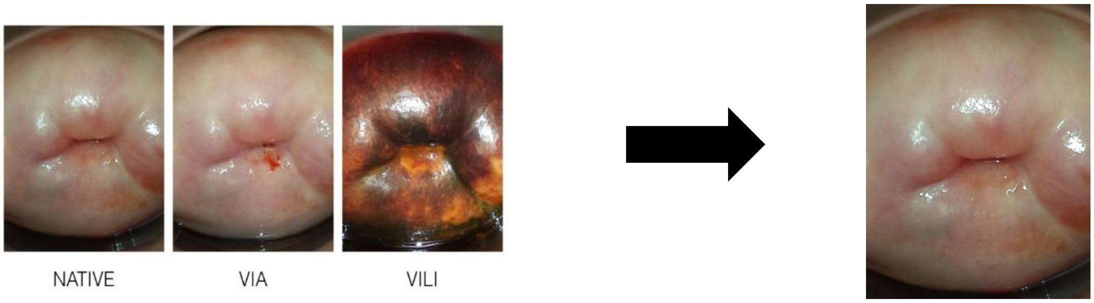


Figure 4. Data cleaning output after cropping of images

Examining the distribution of images across the different categories, the true positive (cancerous) set is too small - making just 4% of the dataset. Maintaining this would create biases in the performance of any selected algorithm as the algorithm would "learn" more on non-cancerous than cancerous images. The next best alternative is to introduce CIN 1, CIN 2 and CIN 3 images into the positive set. *Table 2* shows the number of images in each category after introducing CIN images in the positive set.

Type	Non-cancerous (Negative)	Cancerous (Positive)
No of images	101	24
Percentage	80.8%	19.2%

Table 2. Distribution of images in the dataset across the different categories after the introduction of CIN images in the positive set

Evidently, the difference between the two categories is still very high, but the cancerous set has increased by 15%. This significant increase could reduce biases relatively, and results from the diagnoses through the algorithm could also be used as a preventive care, thereby reducing the incidence of cancer.

3.5.2 Data Processing

The implementation of each algorithm required that the images be processed, and this processing may vary from one algorithm to another depending on the library that was used to implement the algorithm.

i. KNN

Step 1: Using a script, each image in the dataset was labelled as either negative or positive. This is essential because all the images were then put into a single folder and were differentiated by their labels.

Step 2: The images were resized to a fixed width and height of *32* x *32* to obtain a uniform size across the dataset.

Step 3: The RGB pixel intensity of the images was flattened into a single list of numbers. This means that the *32* x *32* image is given three channels for each Red, Green and Blue component respectively to obtain a feature vector – list of numbers that quantify the contents of an image – of $32 \times 32 \times 3 = 3072$ numbers. Each image now has its own feature vector.

ii. CNN

Step 1: Change each image's resolution to 224px. Even though higher resolution image takes more processing time, it results in better classification accuracy.

Step 2: Split the dataset such that 75% of images for training and the rest is for testing.

iii. SVM

Step 1: Split the dataset such that 75% of images for training (X) and the rest is for testing (Y).

Step 2: The images were resized to a fixed width and height of 32×32 to obtain a uniform size across the dataset.

3.5.3 Principal Component Analysis (PCA)

Principal Component Analysis is used in combination with each of the algorithms – obtaining a hybrid algorithm, for example PCA-KNN – to compare performance with and without PCA. After processing the images (for example, if we want a PCA-KNN hybrid, processing of the image follows the KNN processing as explained above), PCA is applied on the image before the machine learning step. The assumption (which is tested at the end of the study) is that reducing the dimensionality of the images before training the model would make the model learn essential features. This will mean that the performance of the hybrid algorithm is expected to be higher than the individual algorithm.

Step 1: Partition the data into training and testing, 75% for training and the remaining for testing.

Step 2: Scale the dataset's features using StandardScaler from Scikit-Learn. This scales the data to have zero mean and a unit variance.

Step 3: Transform each image in the dataset to its equivalence with a minimum number of principal components such that a percentage of the image variance is retained. For this study, we decided to reduce the dimensionality while retaining 95% of the image variance.

3.5.4 Machine Learning

At this stage, we built a Binary Classification Model since the algorithm seeks to predict a binary outcome, either an image is cancerous or non-cancerous. The training and testing of the model also take place at this stage. We used a supervised learning approach to train the model, and we expect the algorithm to correctly classify an image based on the knowledge obtained from training data. Below is the pseudocode for each of the algorithms we considered.

i) KNN

```
Step 1: Process the data using KNN data processing described in the previous section and store each feature vector in a list.
Step 2: From all image names, get the labels and store in another list.
Step 3: Split data into 75% training (X) and the rest for testing (Y).
Step 4:  Learn the features and pattern of the training set.
Step 5: Classify the images in testing set Y by computing the Euclidean distances between images in Y and 4 labelled
```

```
neighbours in X (K= 4). The labelled neighbour in X to which an image in Y has the smallest Euclidean distance (is closest to) takes the label of that neighbour, that is, if an unknown image is closest to a cancerous image, the image is then classified as a cancerous image. This is the step that tests the model.
Step 6: Return the average accuracy obtained in classifying all the images in the test set, Y.
```

ii) PCA-KNN

```
Step 1: Process the data using KNN data processing described in the section and store each feature vector in a list.
Step 2: Process the feature vectors in the list using the PCA approach described in the previous section.
Step 3: Pass the transformed output obtained to Step 4 of KNN pseudocode.
Step 4: Do Step 5 of KNN pseudocode.
Step 5: Do Step 6 of KNN pseudocode.
```

iii) CNN

This algorithm was implemented on TensorFlow, using a pre-trained convolutional neural network model, MobileNet, for image classification built by Google developers. MobileNet has been pre-trained on the ImageNet, an image database of over 14 million different images like dogs, cats and houses. To benefit from this pre-trained model, we are using the concept of transfer learning where we make use of the knowledge gained while solving one problem to solving a different but related problem. MobileNet gained knowledge in how to efficiently learn 14 million images and classify them, and we would be applying this knowledge gain to efficiently learn cervical cancer images and classify them.

3.5.4.1 Fine tuning

Given that the new dataset of 125 images is smaller than the original dataset used to train the pre-trained model, transfer learning is essential and for it to be complete, we needed to fine tune the pre-trained model to suit our needs. In fine-tuning, we trained only the final layers of the model to avoid overfitting, keeping all other layers fixed. That is, we removed the final layers of the pre-trained model, add new layers and retrain only the new layers. We then test this fine-tuned model with the test set and return the classification accuracy.

iv) SVM

```
Step 1: Process the data using SVM data processing described in the previous section and store each feature vector in a list.
Step 2:  Using a kernel function, draw a linear boundary between pixels of the images in the training set and learn the features and patterns (support vectors) on each side of the boundary.
Step 3: Classify the images in testing set Y by looking at the two sets of support vectors to determine which set the new image’s pixel falls in. This is the step that tests the model.
```

v) PCA-SVM

```
Step 1: Process the data using SVM data processing described in the previous section and store each feature vector in a list.
Step 2: Process the feature vectors in the list using the PCA approach described in the previous section.
Step 3: Pass the transformed output obtained to Step 2 of SVM pseudocode.
Step 4:  Do Step 3 of SVM pseudocode.
```

CHAPTER 4: RESULTS

After training the different algorithms and testing their performance, the following results were obtained.

Algorithm	Classification Accuracy	Sensitivity	Specificity
KNN	68.75%	83.3%	65.38%
PCA-KNN	78.12%	83.3%	77.69%
CNN	83.3%	100%	84.6%
SVM	66.37%	89.3%	21.05%
PCA-SVM	62.7%	68.2%	39.02%

Table 3. Classification accuracy, sensitivity and specificity of different algorithms

CNN algorithm obtained a high classification accuracy followed by PCA-KNN, KNN, SVM and PCA-SVM with 83.3%, 78.12%, 68.75%, 66.37% and 62.7% respectively. With sensitivity, high percentages were obtained by CNN, SVM, KNN, PCA-KNN and PCA-SVM with 100%, 89.3%, 83.3%, 83.3% and 68.2% respectively; while algorithms with high specificity results were CNN, PCA-KNN, KNN, PCA-SVM and SVM with 84.6%, 77.69%, 65.38%, 39.02% and 21.05% respectively.

Based on the results obtained, performing PCA on the dataset before training the model increased the classification accuracy of KNN but reduced that of SVM. KNN classification accuracy increased by 9.37% percent after including PCA while SVM classification accuracy decreased by 3.67%. SVM algorithm's sensitivity decreased by 21.1%, from 89.3% to 68.2%, after the introduction of PCA while the specificity increased by 17.97%, from 39.02% to 21.05%. KNN algorithm on the other hand experienced an increase of 12.31% in specificity while sensitivity remained constant.

Post – PCA	**KNN**	**SVM**
Classification accuracy	+ 9.37%	-3.67%
Sensitivity	00%	-21.1%
Specificity	+12.31%	+17.97%

Table 4. Performance metrics variation after the introduction of PCA to KNN and SVM

4.1 Discussion

CNN obtained the highest results across the different performance metric, these results suggest that CNN performed better than all the algorithms in classifying cancerous and non-cancerous images in their right categories correctly. This outperformance is as a result of the algorithm's architecture. Its architecture is made up of many layers where the earlier layers learn generic features, like edges and colour blobs, of an image and the later layers become more specific to the details and content of the image. This in-depth learning makes the CNN classifier more intelligent than others thereby resulting in better results.

In both KNN and SVM algorithms, the images had *32* x *32* = ***1024*** components; applying PCA on the images reduced the dimensionality of the images to about ***205*** components. The algorithm had few things to learn due to the reduction in the number of components, however, the effect of this dimensionality reduction was different on both algorithms. The hypothesis that PCA would help increase the performance of the algorithm was correct with KNN classification accuracy and specificity, but false with SVM classification accuracy and sensitivity. This apparent lack of correlation can be attributed to implementation mistakes. Although performance was not ideal, CNN has proven to be the right algorithm for cervical image classification.

4.2 Shortcoming of research

It is plausible that a number of limitations in this research could have influenced the results obtained. To begin with, the size of the dataset used to train and test the algorithms is not one that is recommended. Some drawbacks of having a small data set include over-fitting of training and/or test set, and outliers could be misinterpreted to be adequate representations of the dataset. Given that the difference between the positive and negative image dataset was very high, 61.6%, there is some likelihood that this might have affected the performance of the algorithms used in this study. As a result, an argument like SVM had a high sensitivity of 89.3% and low specificity of 21.05% because the number of true positives was significantly small, hence a single image had more weight could be binding.

CHAPTER 5: CONCLUSION

The central problem addressed in this research is that of ensuring that health workers make accurate diagnoses from cervical these images taken with smartphones. This research attempted to solve this by analyzing different algorithms in order to recommend one that would efficiently classify cervical cancer images in order to make the diagnoses of this disease efficient and reliable.

After choosing K-Nearest Neighbor (KNN), Convolutional Neural Networks (CNN), and Support Vector Machines (SVM) based on their past performance of obtaining classification accuracies of over 70%, CNN outperformed all the algorithms on the three performance metrics by obtaining a classification accuracy of 83.3%, sensitivity of 100% and specificity of 84.6%. Introducing Principal Component Analysis (PCA) improved the performance of KNN algorithm but had a negative effect on SVM algorithm's performance. With this output it is difficult to endorse or rebuff the theory that generally improves the performance of algorithms, nevertheless, it is possible that the research's limitation could have influenced the results obtained. Finally, further work needs to overcome the research's limitations so as to obtain reliable and unquestionable results.

5.1 Direction for future work

The results obtained suggest the following directions for future research:

i. PCA-CNN

Implementing a PCA-CNN hybrid algorithm or PCA with many other algorithms and comparing the performance of the algorithm before and after PCA would be useful in confirming whether PCA improves the performance of algorithms. This project combined PCA with 2 algorithms and this sample size is not adequate to reach a conclusion.

ii. Test Time Augmentation

Test time augmentation is an approach used to further improve prediction accuracy. This approach predicts the class of a test image while taking into consideration 4 or more random transform (rotations) of the test image. An average of the different predictions is taken in order to determine which class the test image belongs to. This will be particularly beneficial in this project because an image is viewed at multiple angles before being classified in a class, hence making the result more reliable.

REFERENCES

Ahmad, P., Qamar, S., & Qasim Afser Rizvi, S. (2015). Techniques of Data Mining In Healthcare: A Review. *International Journal of Computer Applications*, *120*(15), 38–50. https://doi.org/10.5120/21307-4126

Antonie, M., Coman, A., & Zaiane, O. R. (2001). Application of Data Mining Techniques for Medical Image Classification. *Proceedings of the Second International Workshop on Multimida Data Mining (MDM/KDD'2001)*, 94–101. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.23.9742&rep=rep1&type=pdf

Baitharu, T. R., & Pani, S. K. (2016). Analysis of Data Mining Techniques for Healthcare Decision Support System Using Liver Disorder Dataset. *Procedia Computer Science*, *85*(Cms), 862–870. https://doi.org/10.1016/j.procs.2016.05.276

Bhatnagar, V., Gupta, S. K., & Wasan, S. K. (2001). On mining of data. *IETE Journal of Research*, *47*(1–2), 5–17. https://doi.org/10.1080/03772063.2001.11416198

Brusic, V., & Zeleznikow, J. (1999). Knowledge discovery and data mining in biological databases. *The Knowledge Engineering Review*, *14*(3), 257–277. https://doi.org/10.1017/S0269888999003069

Coenen, F. (2011). Data mining: past, present and future. *The Knowledge Engineering Review*, *26*(1), 25–29. https://doi.org/10.1017/S0269888910000378

Colantonio, A., Di Pietro, R., Ocello, A., & Verde, N. V. (2012). Visual role mining: A picture is worth a thousand roles. *IEEE Transactions on Knowledge and Data Engineering*, *24*(6), 1120–1133. https://doi.org/10.1109/TKDE.2011.37

Computing, M. (2016). Performance Analysis of Data Mining Algorithms for Medical Image

Classification, *5*(3), 604–609. https://doi.org/10.13140/RG.2.1.1603.6245

Doukas, C., Maglogiannis, I., & Chatziioannou, A. (2009). An open web services - based framework for data mining of biomedical image data. *2009 9th International Conference on Information Technology and Applications in Biomedicine*, (November), 1–5. https://doi.org/10.1109/ITAB.2009.5394403

Engineering, C. (2015). A Survey on Data Mining Techniques in, *54*(51), 887–892.

Gallay, C., Girardet, A., Viviano, M., Catarino, R., Benski, A. C., Tran, P. L., … Petignat, P. (2017). Cervical cancer screening in low-resource settings: A smartphone image application as an alternative to colposcopy. *International Journal of Women's Health*, *9*, 455–461. https://doi.org/10.2147/IJWH.S136351

Hand, D., Hand, D., Mannila, H., Mannila, H., Smyth, P., & Smyth, P. (2001). *Principles of data mining*. *Drug safety : an international journal of medical toxicology and drug experience* (Vol. 30). https://doi.org/10.2165/00002018-200730070-00010

Ilayaraja, M., & Meyyappan, T. (2015). Efficient Data Mining Method to Predict the Risk of Heart Diseases Through Frequent Itemsets. *Procedia Computer Science*, *70*, 586–592. https://doi.org/10.1016/j.procs.2015.10.040

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *374*(2065), 20150202. http://doi.org/10.1098/rsta.2015.0202

Jothi, N., Rashid, N. A., & Husain, W. (2015). Data Mining in Healthcare - A Review. *Procedia Computer Science*, *72*, 306–313. https://doi.org/10.1016/j.procs.2015.12.145

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017).

Machine Learning and Data Mining Methods in Diabetes Research. *Computational and Structural Biotechnology Journal*, *15*, 104–116. https://doi.org/10.1016/j.csbj.2016.12.005

Kharya, S. (2012). Survey of Data mining techniques used in healthcare domain. *Jitta*, *2*(2), 55–66. https://doi.org/10.5121/ijist.2016.6206

Leung, K., Cunha, A., Toga, A. W., & Parker, D. S. (2014). Developing image processing meta-algorithms with data mining of multiple metrics. *Computational and Mathematical Methods in Medicine*, *2014*(iii). https://doi.org/10.1155/2014/383465

Li, D., Dong, X., Liu, L., & Xiang, D. (2008). A new cloud detection algorithm for FY-2C images over China. *Proceedings - 1st International Workshop on Knowledge Discovery and Data Mining, WKDD*, 289–292. https://doi.org/10.1109/WKDD.2008.61

Liu, X. (1996). Intelligent data analysis : issues and challenges. *The Knowledge Engineering Review*. Retrieved from http://journals.cambridge.org/abstract_S0269888900008055

Liu, Y., Xia, J., Shi, C. X., & Hong, Y. (2009). An improved cloud classification algorithm for China's FY-2C multi-channel images using artificial neural network. *Sensors*, *9*(7), 5558–5579. https://doi.org/10.3390/s90705558

Matyja, D. (2007). Applications of data mining algorithms to analysis of medical data, (August).

Miles N. Wernick, Yang, Y., Brankov, J. G., Yourganov, G., & Strother, S. C. (2014). Machine Learning in Medical Imaging. *IEEE Signal Process Mag.*, *27*(4), 25–38. https://doi.org/10.1109/MSP.2010.936730.Machine

Ordonez, C., & Omiecinski, E. (1998). Image mining: A new approach for data mining. *Image (Rochester, N.Y.)*, 1–21. Retrieved from http://smartech.gatech.edu/handle/1853/6632

Ramageri, M. (2010). Data Mining Techniques and Applications. *Indian Journal of Computer Science and Engineering*, *1*(4), 301–305.

Ran, L., Zhang, Y., Wei, W., & Zhang, Q. (2017). A Hyperspectral Image Classification Framework with Spatial Pixel Pair Features. *Sensors*, *17*(10), 2421. https://doi.org/10.3390/s17102421

Sudhir, R. (2011). A Survey on Image Mining Techniques: Theory and Applications. *Computer Engineering and Intelligent Systems*, *2*(6), 44–53. Retrieved from http://iiste.org/Journals/index.php/CEIS/article/view/514

Suguna, N., & Thanushkodi, K. (2010). An Improved k-Nearest Neighbor Classification Using Genetic Algorithm. *International Journal of Computer Science Issues*, *7*(4), 18–21.

Thamilselvan, P., & Sathiaseelan, J. G. R. (2015). Image Classification using Hybrid Data Mining Algorithms – A Review. *IEEE International Conference on Innovations in Information Embedded and Communication Systems*, 71–76. https://doi.org/10.1109/ICIIECS.2015.7192922

Thamilselvana, P., & Sathiaseelan, J. G. R. (2015). A Comparative Study of Data Mining Algorithms for Image Classification. *International Journal of Education and Management Engineering*, *5*(2), 1–9. https://doi.org/10.5815/ijeme.2015.02.01

Ullah, Z., Fayaz, M., & Iqbal, A. (2016). Critical Analysis of Data Mining Techniques on Medical Data. *International Journal of Modern Education and Computer Science*, *8*(2), 42–48. https://doi.org/10.5815/ijmecs.2016.02.05

Wasan, S. K., Bhatnagar, V., & Kaur, H. (2006). The impact of data mining techniques on medical diagnostics. *Data Science Journal*, *5*(October), 119–126.

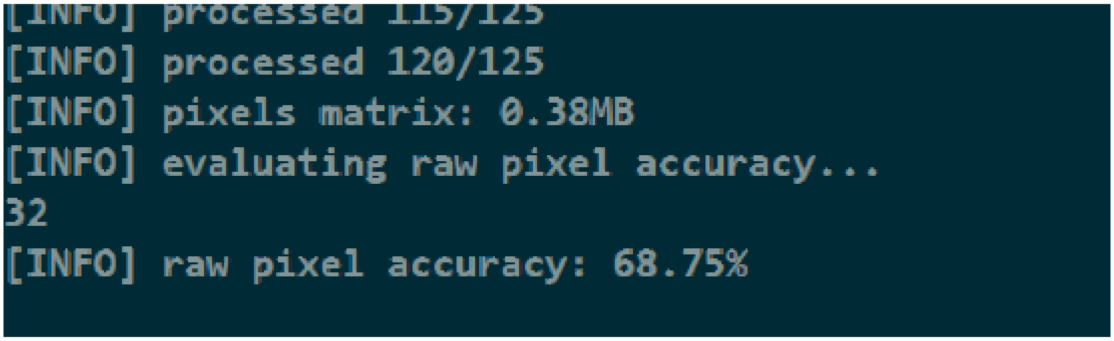
https://doi.org/10.2481/dsj.5.119

Wong, P. C. (1999). Visual Data Mining. *IEEE Computer Graphics and Applications*, *19*(5), 20–21. https://doi.org/10.1109/MCG.1999.788794

Wu, X., Kumar, V., Ross, Q. J., Ghosh, J., Yang, Q., Motoda, H., … Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, *14*(1), 1–37. https://doi.org/10.1007/s10115-007-0114-2

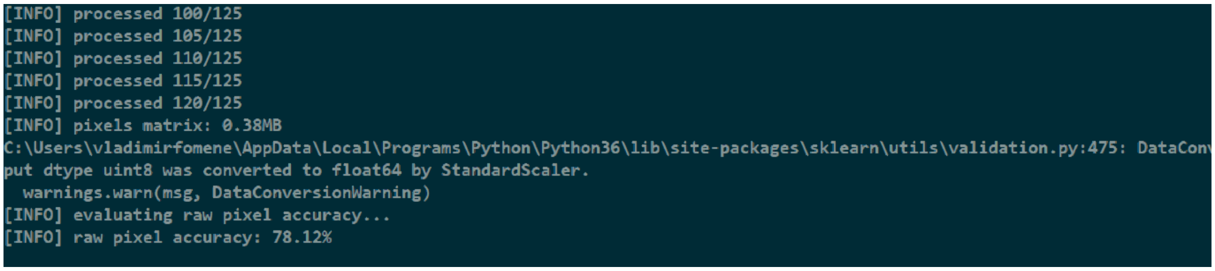
Yu, F., Sun, J., Li, A., Cheng, J., Wan, C., & Liu, J. (2017). Image Quality Classification for DR Screening Using Deep Learning, 664–667.

APPENDIX



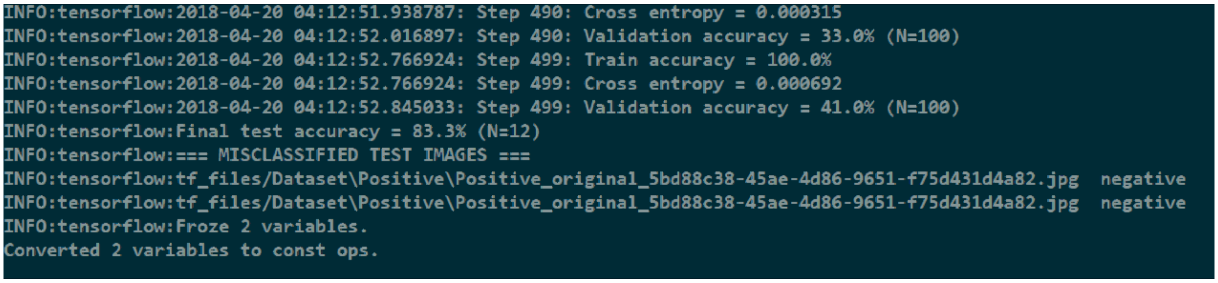
```
[INFO] processed 115/125
[INFO] processed 120/125
[INFO] pixels matrix: 0.38MB
[INFO] evaluating raw pixel accuracy...
32
[INFO] raw pixel accuracy: 68.75%
```

Figure 5. KNN classification accuracy result



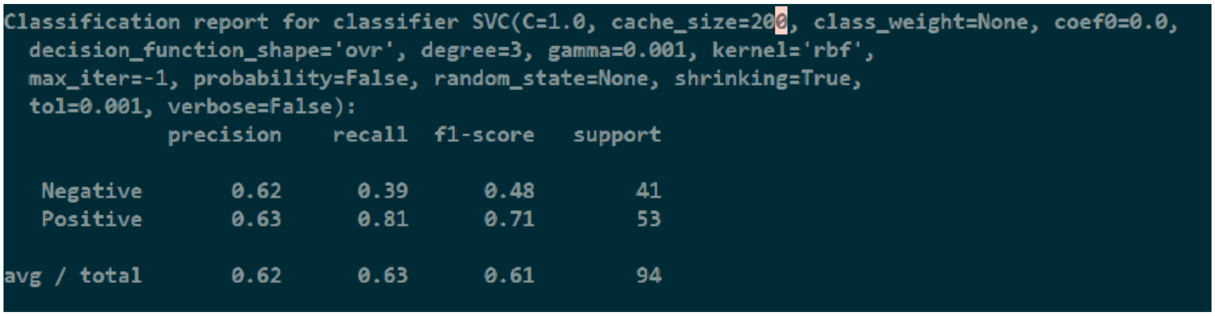
```
[INFO] processed 100/125
[INFO] processed 105/125
[INFO] processed 110/125
[INFO] processed 115/125
[INFO] processed 120/125
[INFO] pixels matrix: 0.38MB
C:\Users\vladimirfomene\AppData\Local\Programs\Python\Python36\lib\site-packages\sklearn\utils\validation.py:475: DataConv
put dtype uint8 was converted to float64 by StandardScaler.
  warnings.warn(msg, DataConversionWarning)
[INFO] evaluating raw pixel accuracy...
[INFO] raw pixel accuracy: 78.12%
```

Figure 6. PCA-KNN classification accuracy result



```
INFO:tensorflow:2018-04-20 04:12:51.938787: Step 490: Cross entropy = 0.000315
INFO:tensorflow:2018-04-20 04:12:52.016897: Step 490: Validation accuracy = 33.0% (N=100)
INFO:tensorflow:2018-04-20 04:12:52.766924: Step 499: Train accuracy = 100.0%
INFO:tensorflow:2018-04-20 04:12:52.766924: Step 499: Cross entropy = 0.000692
INFO:tensorflow:2018-04-20 04:12:52.845033: Step 499: Validation accuracy = 41.0% (N=100)
INFO:tensorflow:Final test accuracy = 83.3% (N=12)
INFO:tensorflow:=== MISCLASSIFIED TEST IMAGES ===
INFO:tensorflow:tf_files/Dataset\Positive\Positive_original_5bd88c38-45ae-4d86-9651-f75d431d4a82.jpg  negative
INFO:tensorflow:tf_files/Dataset\Positive\Positive_original_5bd88c38-45ae-4d86-9651-f75d431d4a82.jpg  negative
INFO:tensorflow:Froze 2 variables.
Converted 2 variables to const ops.
```

Figure 7. CNN classification accuracy result

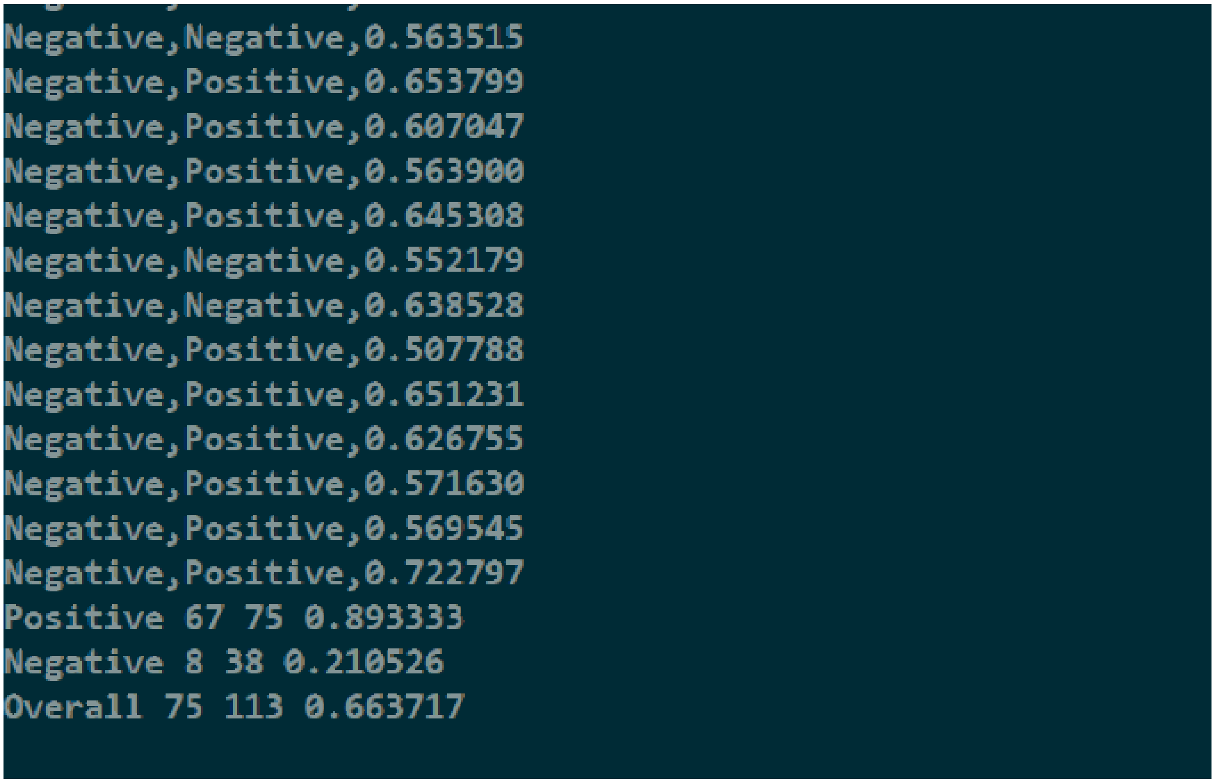


```
Classification report for classifier SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0,
  decision_function_shape='ovr', degree=3, gamma=0.001, kernel='rbf',
  max_iter=-1, probability=False, random_state=None, shrinking=True,
  tol=0.001, verbose=False):
             precision    recall  f1-score   support

   Negative       0.62      0.39      0.48        41
   Positive       0.63      0.81      0.71        53

avg / total       0.62      0.63      0.61        94
```

Figure 8. PCA-SVM classification accuracy result



```
Negative,Negative,0.563515
Negative,Positive,0.653799
Negative,Positive,0.607047
Negative,Positive,0.563900
Negative,Positive,0.645308
Negative,Negative,0.552179
Negative,Negative,0.638528
Negative,Positive,0.507788
Negative,Positive,0.651231
Negative,Positive,0.626755
Negative,Positive,0.571630
Negative,Positive,0.569545
Negative,Positive,0.722797
Positive 67 75 0.893333
Negative 8 38 0.210526
Overall 75 113 0.663717
```

Figure 9. SVM classification accuracy result