Eze et al., J Appl Bioinforma Comput Biol 2020, 9:4
DOI: 10.37532/jabcb.2020.9(4).176

Journal of Applied Bioinformatics & Computational Biology

Research Article

A SCITECHNOL JOURNAL

Evaluation of the Effect of Steganography on Medical Image Classification Accuracy

Peter Eze[1]*, Udaya Parampalli[2], Robin Evans[3] and Dongxi Liu[4]

Abstract

Predictors and features that are used in teleradiology and machine-based auto diagnosis in medicine are often not put into consideration while evaluating medical image Steganography algorithms. In this paper, the effect of embedded security data in automated diagnosis was evaluated using Support Vector Machine (SVM) image classification of Chest X-rays Scan of ***Normal*** and ***Pneumonia*** patients. The goal is to quantify and qualify disease classification parameters because of the addition of steganographic security data in to the image. Four textural image features: Contrast, Homogeneity Energy, and Entropy were used as medical image biomarkers. Their statistical properties for the disease conditions (normal or pneumonia) were profiled and used in SVM training. The evaluation parameters for the machine learning models include accuracy, specificity, recall, and precision. The baseline (before the addition of security data) performance was 86.14% accuracy, 82.18% recall, 90.10% specificity, and 89.25% precision while a typical performance after the addition of security data was 82.18% accuracy, 85.15% recall, 79.21% specificity, and 80.37% precision. We conclude that embedding strength, watermark payload, and region of embedding should be carefully selected to avoid changing the automated diagnostic outcome by a steganographic security algorithm applied to a medical image.

Keywords:

Biomarkers; Evaluation; Steganography; Medical image security

Introduction

According to Liang [1], a biomarker is an objectively measured characteristic indicator of either normal biological processes, pathogenic processes, or pharmacological response to medical treatment. Biomarker are generally classified based on how and where the characteristics are measured. There are radiological, molecular, histological, and physiological characteristic that could be measured as biomarkers [2]. Radiological characteristics are tissue properties measured from non-invasive radiological images or scans. These biomarkers are also called ***image biomarkers***. In general, medical procedures combine image biomarkers as well as other biomarkers for diagnosis and health monitoring, including during clinical trials of new drugs [2]. In contemporary times, medical diagnosis and monitoring processes are increasingly becoming automated [3]. As part of these processes, various medical image scans that originate from patients are stored, transmitted, and used both locally and remotely. These raise security and privacy concerns.

Security and privacy are achieved technologically through Cryptography, Information Hiding, and Intrusion Prevention Systems. The cryptographic means are widely used and better accepted in medicine than data hiding techniques such as steganography. However, the work of [4] outlined the shortcomings of the existing security measures such as firewalls, VPN, encryption, use of header files in images and the use of cryptographic hash. Some of these shortcomings include: only internal security is provided and easily bypassed by hackers (Firewall and VPN), brings about suspicion (encryption), can easily be changed and pirated in both encrypted and unencrypted form (file headers), cannot localise tampered region (cryptographic hash) and always searching for a match, requiring access to a central database (perceptual hashing). The limitations of the existing security measures have led to the demand for an extra layer of security, especially for multimedia data (image, audio and video). This extra layer of security has been known to be data hiding in digital watermarking and Steganography. They are particularly important for robust authentication, unsuspected image integrity checks, text data privacy, and copyright protection.

Steganography completely hides the data from suspicion and tries to maximize capacity and fidelity. With Steganography, the embedded message is of interest but not necessarily encoded in a robust manner. In digital watermarking, the cover is of interest to the receiver, and the embedded message only adds value to the cover either for the sender or for the receiver. The message is often encoded robustly.

As a steganographic technique often introduces extra data into an image, it is important to investigate how this possible image modification affects the performance of machine learning models developed using these stego images. This is the objective of this paper. Specifically, it quantified the effect of watermarking as well as image processing functions on the selected textural image biomarkers: Contrast, Energy, Homogeneity, and Entropy. We have chosen the Support Vector Machine (SVM) classification method. This is because, unlike the deep learning methods, it enables us to study the effect of each feature or biomarker separately. This enables our model to be utilised as a biomarker research tool as well as a disease prediction model. This provides greater insight both in this research and in order, kind of biomarker research study. Thus, we do not like to treat our prediction model as a black box. Instead, we want to know how each feature in the prediction model affects the diagnostic outcome.

This study is necessitated by the fact that the use of generic evaluation parameters originally utilised in Digital Signal Processing (DSP) is no longer adequate for health-related bio data. This practice has limited the option for Medical Image Watermarking (***MIW/S***) to reversible watermarking only. There is a little chance for ***MIW/S*** where medical ethics forbids modification of any sort, whether reversible or irreversible. Hence, we stress that more objective evaluation with medical- specific biomarkers and parameters will enhance the adoption of data hiding methods for medical data security for remote auto diagnosis.

Therefore, along the line of capacity-distortion (diagnostic distortion) performance analysis in the field of steganography, the

***Corresponding authors:** Peter Eze,School of Computing and Information Systems, University of Melbourne, 3010, Parkville, Australia, E-mail: peter.eze@unimelb.edu.au

Received: July 02, 2020 **Accepted:** August 05, 2020 **Published:** August 12, 2020

following question is answered: how do steganographic algorithms affect selected medical image biomarkers employed for auto-diagnosis? Answering this question is important for applications in which patient data are transmitted to remote Artificial Intelligence (AI) systems for automated diagnosis. An example is during this era of COVID-19 [5] pandemic when the health system is being overwhelmed in various nations of the world. We could employ machine learning algorithms for initial diagnosis, but the security of patients' health records will need to be preserved. The application of this evaluation technique has been incorporated in a wider medical image information hiding evaluation algorithm [6].

The rest of this paper is organized thus: Section ***Related Work*** covers existing works in both watermark evaluation in auto-diagnosis as well as some existing works that used textural features for disease classification. This was followed by our evaluation method in Section ***Proposed Method.*** Section ***Experimental Setup*** is a description of the Pneumonia diagnosis case study, the employed watermarking algorithm, and the experimental setup. The results from the Pneumonia dataset classification using SVM is presented in Section ***Results*** and then discussed in ***Discussion and Comparison*** with some comparisons with existing methods. The paper is concluded in ***Section Conclusion.***

Related Work

Though lots of works have been done on Medical Image watermarking (MIW) and Medical Image Steganography (MIS) [7-10]. The evaluation of the effect of water marking on diagnosis by the use of image biomarkers and machine learning models has been under-researched .Existing evaluations are largely limited to the use of parameters such as Peak Signal-to-Noise ratio(PSNR), Mean Square Error(MSE),Structural Similarity Index Measure (KLD), Kullback-Leibler Divergence (KLD)and other measures of distance used in core digital signal processing (DSP).Predictors and features which are used in teleradiology and machine-based auto diagnosis in medicine are not usually put into consideration and ,therefore, often do not convince medical practitioners about data hiding security techniques.

In [11] made a case for why we should not rule out MIW. They described how it could be used to bring evidence in the case of telemedicine litigation. In this technique, the watermark should not be removed from the medical image, and even many should not be aware of its existence in the first place. They quickly added that it remains a challenge to convince both the legal and medical practitioners that watermarking and steganography can provide trusted evidence without compromising diagnosis. The difficulty arises from the stringent constraints generally placed on the integrity of data used for medical, military and legal applications [12]. This justified further research by [12] on digital watermarking techniques that may not violate this stringent constraint. The shortcoming of the research in [12] is that it is subjective, and only a few experts were involved. The criteria used for evaluation do not apply to machine learning algorithms [13]. Performed an objective evaluation of the impact of watermarking on computer-aided diagnosis in medical imaging. They used about 500Breast Ultrasound as their dataset. Two watermarking algorithms were used on half of the samples. The evaluation parameters included PSNR, Watson Parameter and bits-per-pixel. They tried to establish the effect of spread Spectrum DCT (SS-DCT) and High Capacity Data Hiding (HCDH) watermarking on the segmentation and classification accuracy of the lesions in the image. They found that with an appropriate choice of parameters, both water marking systems can perform well without any adverse effect on segmentation and classification accuracy. However, SS-DCT could alter the accuracy if high embedding strength is used. Their approach is most related to our work but not in the area of pneumonia disease classification using X-ray scans.

In a recent study by [14], some watermarked Fundus eye scans were tested against some models [15-18] used to classify Healthily, Macular Edema and Central Serous Chorio Retinopathy (CSCR) eyes diseases. The original accuracy of their models ranged from 95% to 100%. Their results show that there was no difference in classification accuracy for the original and watermarked test set. However, few test data were used (15 to 45). Also, it was not clear if the original model was trained and the watermarked training set. Again, this study failed to recognize that if watermarking is adopted for integrity checks, future training sets would contain watermarked data and not just the test data.

In recent times, the interpretation of medical scans and computation of biomarkers used for diagnosis is done by machines. This use of automated diagnosis would help one to perform a repeatable and objective evaluation for different steganographic algorithms. Specifically, we have explored a larger dataset (2400 train set and 202 test set) for Pneumonia to study the effect of watermarked medical image on medical auto diagnosis. The goal is to bring out these effects for medical decisions and establish parameters that should be directly used to accept or reject data hiding algorithms for image security through data hiding. In other to put our work and choice of biomarkers in context, we will go further to review other works that used textural biomarkers that are like the ones chosen for this research.

Texture analysis using textural biomarkers have been used in several image analysis to study the health of patients. The study by Kim [19] used heterogeneity (or homogeneity) and entropy texture analysis to study survival outcomes for patients with breast cancer. It was established that there is a relationship between the level of entropy and heterogeneity of T2-weighted images and the recurrence-free survival (RFS) rate of patients.

Among other image features (biomarkers), textural image biomarkers were used in [20] to differentiate between three types of mouse xenograft tumour models automatically. Up to thirty textural image biomarkers obtained from contrast-enhanced Ultrasound scans were used. Similarly, Xu [21] employed 96 features (52 features of the gray-level co-occurrence matrix (GLCM) and 44 features of the gray-level run-length matrix (GLRLM)) extracted from the regions of interest (ROIs) of ultrasound images to classify two liver diseases of hepatocellular carcinoma and liver abscess. They used the extracted features to train a Support-Vector Machine (SVM), which had 88.88% classification accuracy.

In this paper, we limit our study to four textural biomarkers that are commonly used in literature [19-21] for medical image classification for computer-aided diagnosis. They include Energy, Contrast, Homogeneity and Entropy. These features are a subset of the fourteen [22] texture feature operators from the Gray-Level Co-Occurrence Matrix (GLCM) of the ROI image [23-24] They can be employed for both content and diagnostic information integrity checks. We have selected these features for various reasons: (i) they are considered closer to the medical profession than the PSNR, and SSIM image quality check parameters, (ii) they represent good examples of the robust(Energy and Entropy) and the non-robust feature types used in machine learning algorithms.

doi: 10.37532/jabcb.2020.9(4).176

(iii) the curse of dimensionality problem [25] where the performance of classification systems set in due to the use of too many features, could be solved with few relevant features, and (iv) they are unique features that can characterize the disease and thus their validity as Pneumonia biomarkers can be established. This fourth reason is why the deep machine learning method was not adopted in this research as it masks the effect of specific features used in classification. In the next section, we will present the details of our evaluation method.

Proposed Method

The conceptual framework for the proposed evaluation system is shown in [Figure.1]. Only one model known as the *baseline model* (Model-1) will be trained using the original dataset. Various models (Models 2 to n)were trained using the watermarked versions of the original data at various levels of data pay load as well as different embedding strengths, α.

The baseline model was developed for medical image classification using the original dataset without any watermark inserted in the ROI and without any form of pre-processing. After this, different levels of water marking (leading to different embedding strengths) were applied to the original image. For each of the water marking levels, a new model is created using the corresponding watermarked image as training data. Each of the models was used to classify a test set, which is not part of the training set. Were corded the magnitude of changes in any of the performance parameters due to the models created from watermarked images compared to the baseline model.

In the next sub- sections, we will provide the details of the feature extraction and SVM training process.

Feature Extraction

GLCM provides information about the positions of pixel pairs having gray level values, (i, j) at a distance, d measured in one of the directions, θ=0^0,45^0,90^0,135^0 about the reference pixel. This is depicted in Figure 2 with d=1. Equation 1 is then used to generate the co-occurrence matrix, from which three out of the four texture features are computed.

$$GLM = G_{d,\theta}[i,j] = C_{i,j} \quad (1)$$

.The features are computed from the ROI of the image after proper image segmentation such as the one shown in [Figure.3] The

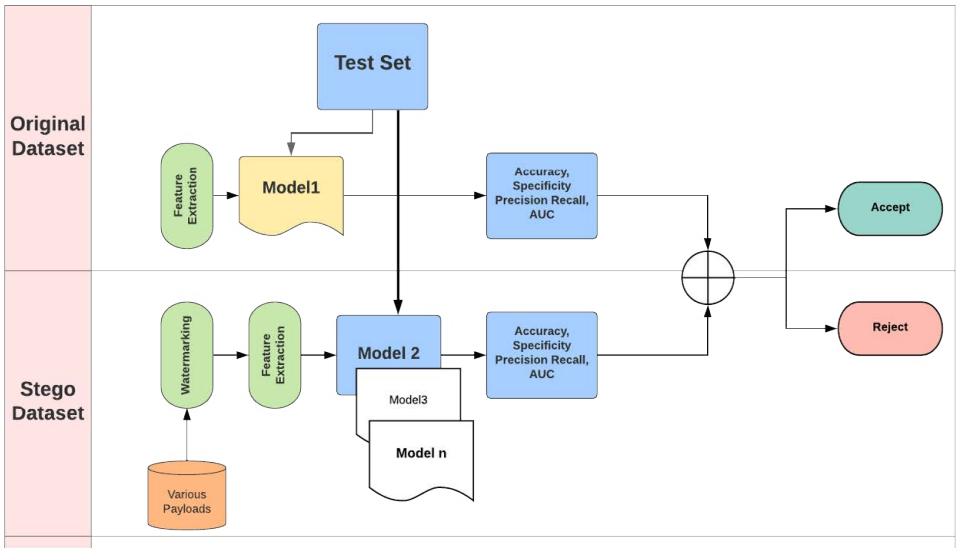


Figure 1: Evaluation Framework. The result from each model is compared with the baseline model before the algorithm and its parameters are accepted or rejected.

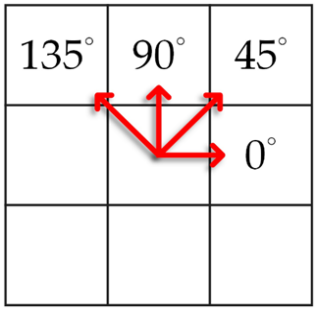
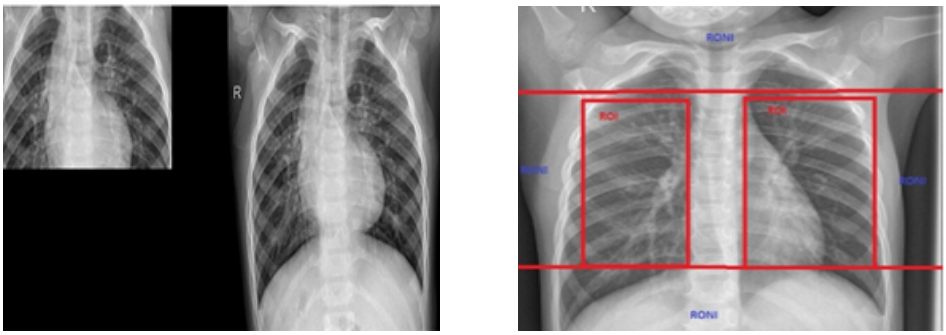


Figure 2: GLCM Computation for Textural features. This is taken from the ROI of the medical image. The ROIs were determined based on the recommendations of the doctors who first worked on this dataset.

doi: 10.37532/jabcb.2020.9(4).176



(3a) ROI (Left) of Chest X-ray (Right) for auto-Pneumonia diagnosis.

(3b) Determination of ROI and RONI

Figure 3: The rectangular three-quarter portion of each half of the X-ray scan is automatically extracted as ROI.

textural changes in the ribs of the patient are the indicators of having or not having pneumonia.

The features that were extracted from the above Co-occurrence matrix are (1-3):

ROI Contrast, C_t: Contrast is a measure of the local variations present in an image.

$$C_t = \sum_i \sum_j (i-j)^2 G_{d,\theta}[i,j] \quad (2)$$

ROI Energy, C_e: Energy is computed from the Angular Second Moment(ASM):

$$C_e = \sqrt{ASM} \quad (3)$$

Where: $ASM = \sum_i \sum_j (G_{d,\theta}[i,j])^2$ (4)

ROI Homogeneity, C_h: Homogeneity refers to how continuous a pixel value spans along a given direction without meeting corners or edges. A homogeneous surface has all the pixel having the same gray value.

$$C_h = \sum_i \sum_j \frac{G_{d,\theta}[i,j]}{1+|i-j|} \quad (5)$$

ROI Entropy H(X): T is the entropy of the ROI region. This is computed as a feature as:

$$H(X) = \sum_i p(i) \log p(i) \quad (6)$$

Where p (.) is the probability density function (pdf) or probability mass (pmf) of image, X. This function is already implemente in MATLAB as entropy ( ) function.

The existence of well-established libraries for extracting them from a medical image also influenced our choice of these biomarkers. The features extracted in this way are used individually to train the SVM Classifier as discussed in the next section.

7 Support Vector Machine (SVM) Classification

SVM supports binary (only two classes) classification where each subject is either in the ***positive*** or ***negative*** class. For this study, the positive class is pneumonia while the negative class is normal.

This classification by an SVM is performed through a *kernel f unction*, Φ, that can map the training examples, x_i in to a higher dimensional space. For a training example x_i with a corresponding label x_j, the general form of a Kernal function is given as:

$$K(xi,xj)=\Phi(xi)T\Phi(xj) \quad (7)$$

Four types of kernel are in common use:

LinearKernel: $K(xi,xj)=x_i^T x_j$

PolynomiaKernel: $K(xi,xj)=(xi,xj)d$. Where d is the degree of the polynomial.

Gaussian Kernel or Radial basis function (RBF): $K(x_i, x_j) = e^{\left(\frac{\|x_i - x_j\|}{2\sigma^2}\right)}$ **Where σ stands for a window width.**

Sigmoidkernel: $K(xi,xj)=tanh(k(xixj)+\varphi)$.Where k and φ are some kernel parameters.

So, the SVM uses a chosen kernel equation to test if a patients X-ray scan is positive for pneumonia or not. This leads to the four possible outcomes of a classification or prediction algorithm: True Positive (TP), False Positive (FP),Truenegative(TN) and Falsenegative(FN). This form what is called a Confusion Matrix, C

$$C = \begin{bmatrix} TP & FN \\ FP & TN \end{bmatrix}$$

TP is a correctly predicted positive class; FP is a false positive prediction where a subject in negative class is predicted as being in a positive class. FN is the opposite of FP, where a subject in a positive class is predicted as being in the negative class, while TN is when a negative class is predicted as negative class. With the confusion matrix, other performance parameters for a machine learning algorithm, such as accuracy, specificity, recall (sensitivity) and precision can be defined. The equations that follow define the performance parameters used in this study.

Accuracy is a measure of the ratio of all correct predictions, whether positive or negative, to the entire test set. It is not a good parameter if the number of subjects in each class is not the same.

$$accuracy = \frac{TP+TN}{TP+TN+FN+FP} \quad (8)$$

In this study, accuracy is the ratio of the sum of patients correctly diagnosed (predicted) as normal and those correctly diagnosed as having pneumonia to the number of patients that arrived for diagnosis (subjects used for test set). We have limited our study to equal training and test sets to avoid the complexity and bias introduced by unequal datasets.

Specificity is the True Negative Rate (TNR) as it is the ratio of the number of correct negative predictions divided by the total number of the negative class (N).

$$specificity = \frac{TN}{TN+FP} \quad (9)$$

This is the ratio of of the number of people correctly predicted as normal to the total number of normal people. It monitors the supposed assurance that patients are not predicted as having a disease that they do not have.

Recall or Sensitivity or True Positive Rate (TPR) is the ratio of correct positive predictions to the total number of the positive class (P).

$$recall = \frac{TP}{TP + FN} \quad (10)$$

This is the ratio of the number of patients correctly predicted as having pneumonia to the total number of people who really have pneumonia.

Precision or Positive Prediction Value (PPV) is the ratio of correct positive predictions to the total positive predictions.

$$Precision = \frac{TP}{TP+FP} \quad (11)$$

In our experiment, this is the ratio of the number of patients who have pneumonia to the number of those predicted as having pneumonia. Case Study: The Pneumonia Data set.

Pneumonia diagnosis using a Chest X-ray scan and an SVM trained model is used in this research. We also used a recently developed Spread Spectrum watermarked algorithm for adding watermark to the X-ray images.

Description of data set

Chest Xray: The Pneumonia Chest X-ray data set for the Kaggle Competition was used. It is made of various (large) sized X-ray images of patients that are either normal or have been diagnosed with Pneumonia disease. The same data set was used in [26] for deep learning image classification of bacterial and viral pneumonia. Some images are so large that they produced ROI of size 1285x1241. We utilised equal number of positive (Pneumonia) and negative (normal patient) classes, each with 1200 images and then 202 images for the test set. That is 101 images for each class.

Steganographic or Watermarking algorithm

Once a disease prediction is to be made via a machine learning model, and there is the likelihood of watermarking either the images used for training the model or the images that will be classified in the future. Our framework serves as a black-box evaluation mechanism to ascertain the potency of the algorithm for future predictions. The specific steganographic or watermarking algorithm applied to the image is immaterial to this framework. However, as a means of evaluating a recently developed algorithm, the Constant Correlation Compression Code Scheme ($C4S$) Steganographic algorithm in [27] was used as a case study in this research. A summary of the $C4S$ insertion strategy is shown in Figure 4.

$C4S$ is an additive spread spectrum watermarking technique that can be easily used for either fragile, semi-fragile, or robust watermarking by adjusting the parameters ρ and ϵ (epsilon). The parameter, ρ is a real-valuednumberagreedbetweenthesenderandthe receiver.Itisembedded in such a way that the correlation value at the receiver between an image sub-block, Xi and a secret-key-generated sequence, W equals $\pm\rho \pm\epsilon$. Hence, ϵ is a control parameter for determining the level of fragility (or robustness) of the water marking algorithm. In general, $\epsilon<p.\epsilon$ is a tolerance parameter. G is a gap required between insertion zones to detect tampering.

Experimental set up

This section is set up for SVM model building and evaluation. All models and analyses are based on the existing 40dB benchmark [28] for image distortion. Initially, the number of bits per sample (capacity) increased from cr= 1,2,3,...until visible degradation is noticed, or computed PSNR between original and watermarked image becomes 40dB the average among all the samples in each data set class. Then, the embedding strength, α and base correlation values,ρ, was controlled and capacity observed. The dynamic embedding strength α is the major parameter that determines both accuracy and distortion.

The specific SVM used in this experiment is the fitcsvm provided in MATLAB2017b. It is a fast algorithm for training low to medium-dimensional data sets for binary classification. The default hyper-parameters were used. Hence, not tuning was performed. A supervised learning method is used. The extracted feature vectors are stored in a variable X, while the corresponding labels are stored in a variable, y. The training starts with a call to:

$$SVM\ Model = fitcsvm\ (X,y) \quad (12)$$

where SVM Model is the trained model with a feature vector, X. This process is repeated using various payloads and embedding strength, α.To test the trained model, the same type of feature or biomarker used for training is extracted from the test set and used in this form.

$$[score, label]=predict\ (CVSVM\ Model\ ,\ X\ Test) \quad (13)$$

Results

SVM models were trained using a clean original image (Cr=0), the watermarked image at one bit per sample (Cr=1),and then watermarked image at two bits per sample(Cr=2).The performance results for the X-ray data set are presented in Table 1.The models where Cr=0 is the base line model because no watermark was added

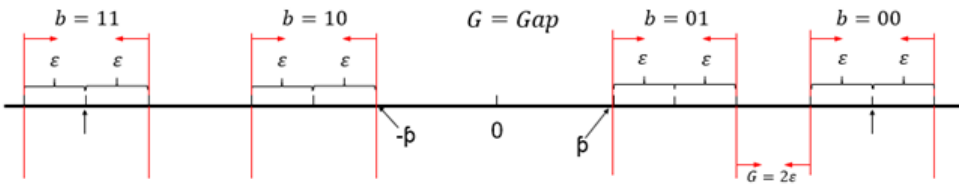


Figure 4: C_4S Insertion Strategy: Pre defined correlation channels within the image represents groups of bits. Each channel is separated by a channel gap, **G**. The width of each channel is 2ϵ.

to the images before the extraction of the features or biomarkers used to train this model.

The performance of the models as a function of the payload is presented next. The embedding strength was not controlled in any way in these experiments. These are shown in Figures 5 and 6. Figure 5 shows that Contrast and Homogeneity are good classification features for pneumonia, but they are not stable to watermarking. There were significant changes due to the high rate of distortion. This calls for adequate control of both watermarking.

Figure 6 shows that energy and entropy features are least affected by watermarking. Entropy is not affected at all. The value of accuracy, recall, specificity and precision remained the same for all payloads. There was a slight change in these parameters in terms of energy. These are the robust features.

The negative result of Figure 5 called for further controlled experiments. In response to this, the independent variable was changed to maximum embedding strength (to control distortion) and not the payload. This produced more models. The results for models based on the Contrast feature are shown in Table 2.

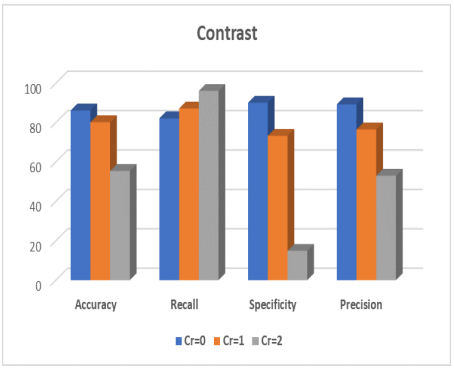
In order to visualise the effects on accuracy, recall, specificity and precision, Figures 7 and 8 are presented. These figures simultaneously present results about Contrast, Homogeneity, Energy and Entropy features for the purpose of comparison.

Figure 7a shows a decline in accuracy for Contrast and Homogeneity, an increase in accuracy for Energy, and constant accuracy for Entropy. For Homogeneity, the accuracy at a maximum α of 0.5 is the same as the accuracy of the baseline model. This result shows that some biomarkers are more stable to image changes than others.

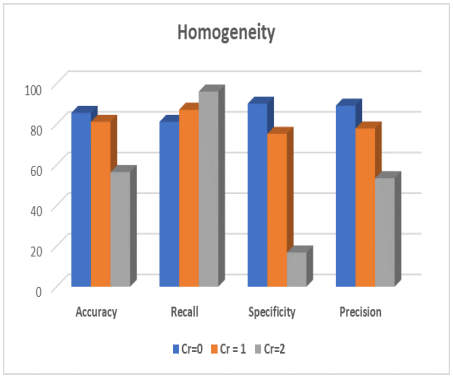
The result of Figure 7b for the recall is an interesting one. Recall can be interpreted as the accuracy of the positive class. For Contrast, Homogeneity and Energy, recall increases as maxi- mum embedding

Table 1: SVM performance with contrast a straining feature at various embedding capacity, r is the number of bits embedded in a sample by the $C4S$ method.

Cr	Accuracy (%)	Specificity (%)	Recall (%)	Precision (%)
0	86.14	82.18	90.1	89.25
1	80.2	87.13	73.27	76.52
2	55.45	96.04	14.85	53.01

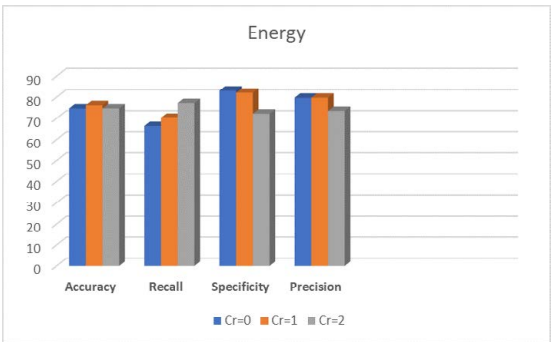


(5a) Contrast

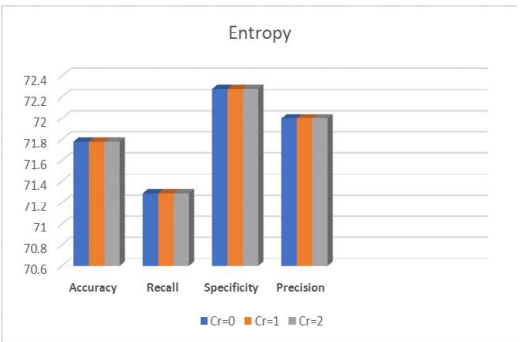


(5b) Homogeneity

Figure 5: Effects of Steganography on Contrast and Homogeneity Models. Cr= 0 indicates that no watermark was added to the image while Cr= 1 and Cr= 2 means that 1 and 2 bits, respectively, of watermark has been added in each 8x8 block subject to some distortion rates.



(6a) Energy

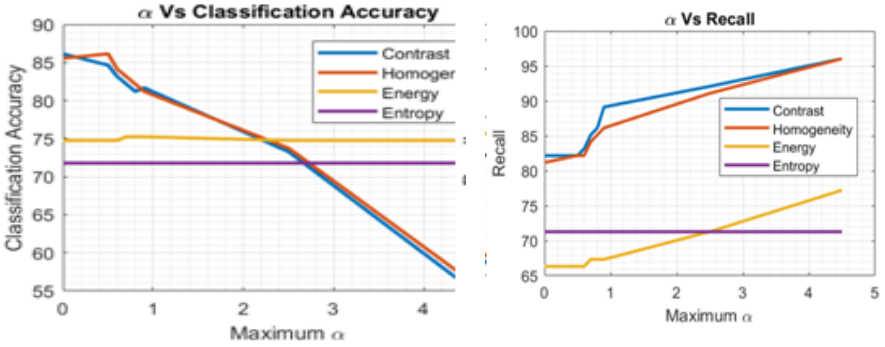


(6b) Entropy

Figure 6: Effects of watermarking on Energy and Entropy Models. Entropy is generally not affected as there is no significant change in the distribution of pixel elements in the watermarked version.

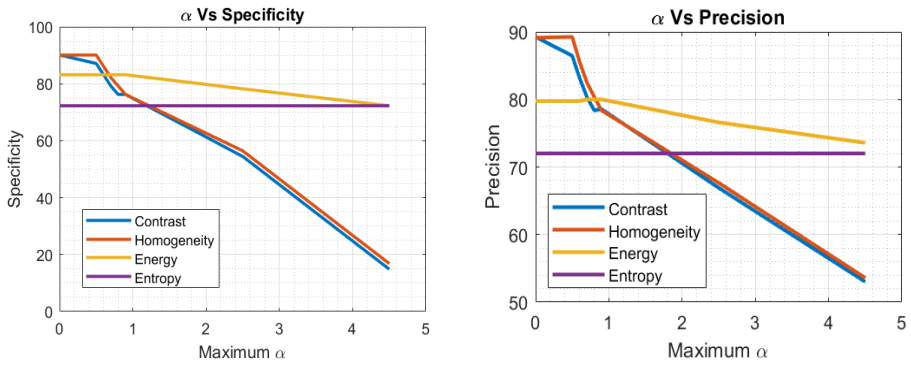
Table 2: SVM performance with contrast as training feature at various watermark embedding strengths, α.

α	Accuracy (%)	Recall (%)	Specificity (%)	Precision (%)
0	86.14	82.18	90.1	89.25
0.5	84.65	82.18	87.13	86.46
0.6	83.17	83.17	83.17	83.17
0.7	82.18	85.15	79.21	80.37
0.8	81.19	86.14	76.24	78.38
0.9	81.68	89.13	76.24	78.57
2.5	73.27	92.08	54.46	66.91
4.5	55.45	96.04	14.85	53.01



(7a) Accuracy (7b) Recall

Figure 7: Effect of Steganographic embedding strength, α on the Accuracy and Recall for Pneumonia disease classification.



(8a) Specificity (8b) Precision

Figure 8: Effects of watermarking on Specificity and Precision.

strength increases. This means that more accurate predictions of people having pneumonia are being made.

Specificity is the accuracy of the negative class. Figure 8a indicates that as the embedding strength increases, the specificity decreases. This means that the false alarm (Type I error) rate also increases. The implication is that normal patients could be referred for further investigations. However, for homogeneity, this does not occur until α= 0.6; for energy, it is until α= 0.9; while for Entropy, it never occurred for the range of αused in the experiment. The Contrast had the most rapid negative response.

Precision had a similar trend with specificity. This is shown by 8b.If more type I error occurs, then the precision of a prediction would decrease. This does not stop recall from increasing though if some more accurate positive predictions are also being made. Because current machine learning methods use other forms of image pre-processing before feature extraction, a future study will investigate how the pre-processing and data argumentation compares with mere watermarking and whether they can override both the positive (in terms of energy) and negative (in terms of Contrast) effects on classification accuracy, recall (sensitivity), specificity and precision [29-30].

Discussion and Comparison

Concerning the research questions posed in the introduction, and based on the results above, it can be inferred that different biomarkers are affected differently by data hiding security techniques. With low embedding strength and smaller payload, the effect is generally low.

However, adaptive embedding algorithms affect different parts of the image differently, creating variation in Contrast and homogeneity. This change tends to move the value of these features from one class to another. Hence, some of the predicted values using the watermarked models gradually misclassify the test data as the embedding strength and payload increases to change the Contrast and homogeneity of the train data.

On the other hand, it should be noted that some biomarkers or features are more resistant to watermarking than others, such as entropy and energy. Entropy is related to the information content of an image (or any data). The non-variation in the model trained with entropy feature shows that the C_4Salgorithm used as the case study in this research does not change the content of an image even for reasonably large embedding strength (α= 4).

Further interpretation of this behaviour is found in a recent report by Edwards in a recent Communications of the ACM. SVMs and deep neural networks (DNN) have two features that they can recognise: robust and non-robust. Frequently, the non-robust features are subtle, but representatives and serve good features to predict a class. However, they are easy to subvert as little pixel changes tend to shift them to another class. On the other hand, robust features continue to deliver correct results even when the pixels are changed by small amounts.

Interpreting why Recall improved as the embedding strength increases prompted for more observations. We noticed that the negative class and the positive class responded differently to watermarking. In general, the decision boundary shifted to favors correct classification of positive class (higher recall). Correspondingly, the reduction in accuracy came from the misclassification of the negative class (lower specificity). However, as noted by Kermany in false-negative (Type II error) result is *far more serious* in disease classification. Hence, an increase in recall reduces type II error and ensures that the maximum number of people having pneumonia is correctly referred for further investigations.

From the above, Steganographic methods affect biomarkers based on the type of image feature modified by the embedding process. Secondly, for the most affected biomarkers, there is a decrease in type II error. This, however, increases false alarm (Type I error). Whereas all errors are important to be removed from a medical system, over watermarking implies that some normal patients are likely to be referred for second-level examination when they are not supposed to. This was considered better than the opposite, where a sick person is declared normal. At an embedding strength of more than 2.5, the accuracy of the watermarked model would reduce beyond 70% benchmark and could be considered adversarial. Therefore, it is important to establish the parameter boundaries for any new steganographic or watermarking system for which the algorithm becomes adversarial regarding the biomarkers of interest. This is generally not an issue for human-based diagnosis as the human eye is robust to the subtle changes in the non-robust features.

Our focus is to create empirical data in the troubled waters of medical image ROI. This concern is justified by the fact that image comparisons that occur in teleradiology in terms of the type of machine used or parameter settings before image capturing is also related to ROI. Also, it is a well-known fact that apart from watermarking, medical images undergo other pre-processing and post-processing operation that does not affect only the RONI neither are they reversible. This has given few researchers the courage to evaluate the effect of such modification on a diagnosis. Following this method, similar experiments can be carried out for any medical image, biomarker, and corresponding steganographic algorithm.

The advantages and superiority of our method lie in the following:

We can incorporate the opinions of more experts and historical patient scans in one study. Only three experts were involved, whereas, in about five experts were involved. A recent study managed to get seven expert evaluations. These are opposed to the hundreds of experts and exactly 2602 patients whose data were used in this study.

Testing several specific criteria as mentioned by Ludewig will be very tedious for human experts. With our method, automated models could be used to extract the specific features that could be used to test each criterion. Both the individual and combined effect could be evaluated repeatedly.

For remote, developing rural areas where experts are not accessible, our method becomes an efficient way to transfer and utilize the knowledge of experts in image evaluation. Existing models can always be retrained as more expert information becomes available without requiring the presence of the experts.

Conclusion

In this work, we have introduced the Machine-learning evaluation methods for automated applications that will be used in environments where medical experts are not readily available. As the dataset used has already incorporated the subjective knowledge of experts, the initial human input has been taken care of as currently being used in medical evaluation methods. Different diseases have different medical image biomarkers that could be objectively measured. In this paper, the chosen biomarkers were limited to well-known image biomarkers that apply to a wide range of diseases. Because medical image watermarking is deemed domain-specific due to human health's critical nature, we chose five textural image biomarkers to evaluate the effect of our watermarking algorithm. It is evident in terms of these biomarkers that there is a modification effect caused by the watermarking algorithm. However, the effect is very low on certain image biomarkers such as energy and entropy. Some effects are positive as they increase the accuracy of predicting the positive class. For the negative effect, further improvement is required in this respect and only watermarking at very low embedding capacity, and the payload is recommended. As a last resort, reversible or RONI-only watermarking is applicable for such biomarkers.

Acknowledgement

Melbourne Research Scholarship and CSIRO Data61 for scholarships that aided this study.

References

1. Liang S, Shen G (2011) Biomarkers of Glioma. Molecular Targets of CNS Tumors, Dr. Miklos Garami (ed.)
2. Deng J, Wang Y (2017) Quantitative magnetic resonance imaging biomarkers in oncological clinical trials: Current techniques and standardization challenges. Chronic Diseases and Translational Medicine 3: 8–20.
3. Rocek A, Zatloukal V (2015) Comparative Study of Negative Aspects Elimination of Medical Image Watermarking Methods. International Journal of Information and Electronics Engineering 5: 6–9.
4. Nyeem H, Boles W, Boyd C (2013) A Review of Medical Image Watermarking Requirements for Teleradiology. Journal of Digital Imaging 26: 326–343.
5. Shereen MA, Khan S, Kazmi A, Bashir N, Siddique R (2020) COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses. Journal of Advanced Research 24 :91–98.

6. Eze PU, Parampalli U, Evans RJ , Liu. D (2020) A New Evaluation Method for Medical Image Information Hiding Techniques. In the Proceedings of 42nd IEEE Engineering in Medicine and Biology 6119 – 6122.
7. Al-Quershi OM, Khoo BE (2011) Authentication and Data Hiding using a hybrid ROI-based watermarking schemes for DICOM images. Journal of Digital Imaging 24: 114–125.
8. Maity HK , Maity SP (2012) Intelligent Modified Difference Expansion for Reversible Watermarking. IJMA 4: 83–85.
9. Memon NA , Gilani SAM (2010) Watermarking of Chest CT scan Medical Images Authentication. Journal of Computer Mathematics. 88: 265–280.
10. Zain JM, Clarke M (2007) Reversible Region of Non-interest (RONI) watermarking for Authentication of DICOM images. IJCSNS 7: 19–28.
11. Coatrieux G, Quantin C, Allaert F, Auverlot B, Roux C (2011)Watermarking: a new way to bring evidence in case of telemedicine litigation. European Federation for Medical Informatics 611–615.
12. Zain JM, Fauzi ARM, Aziza AA (2006) Clinical Evaluation of Watermarked Medical Images. In International Conference of the IEEE Engineering in Medicine and Biology Society, New York, NY. 5459–5462.
13. Garcia-Hernandez JJ, Gomez-Flores W, Rubio-Loyola J (2016) Analysis of the impact of digital watermarking on computer-aided diagnosis in medical imaging. Computers in Biology and Medicine 68 (2016), 37–48.
14. Hassan B, Ahmed R, Hassan O (2019) An Imperceptible Medical Image Watermarking Framework for Automated Diagnosis of Retinal Pathologies in an e Health Arrangement 7:69758 – 69775.
15. Hassan B, Ahmed R, Li B (2018) Computer aided diagnosis of idiopathic central serous chorioretinopathy. In Proceedings of 2nd IEEE 824 – 828.
16. Hassan B, Ahmed R, Li B, Hassan O, Hassan. T (2019) Automated retinal edema detection from fundus and optical coherence tomography scans. In Proceedings of 5th ICCAR, Beijing, China 1 – 6.
17. Hassan B, Raja G (2016) Fully automated assessment of macular edema using optical coherence tomography (OCT) images. In Proceedings of International Conference of Intelligent System Engineering 5 – 9.
18. Hassan B, Raja G, Hassan T, Akram MU (2016)Structure tensor based automated detection of macular edema and central seous retinopathy using optical coherence tomography images. 33: 455 – 465.
19. Kim J, Ko ES, Lim Y, Lee KS (2016) Breast Cancer Heterogeneity: MR Imaging Texture Analysis and Survival Outcomes. Free Access Breast Imaging 282: 665–675.
20. Theek B, Opacic T, Magnuska Z, Lammers T, Kiessling F (2018) Radiomic analysis of contrast-enhanced ultrasound data. Scientific Reports 8. 11359: 1 – 9.
21. Sendren Sheng-Dong Xu, Chun-Chao Chang, Chien-Tien Su, Pham Quoc Phu (2019) Classification of Liver Diseases Based on Ultrasound Image Texture Features. Applied Sciences 9.
22. Haralick RM, Shanmugan k, Dinstein I (1973) Textural Features for Image Classification. IEEE Transactions On Systems, Man and Cybernetics 3: 610–621.
23. Clausi DA, Jernigan MD (1998) A Fast Method to Determine Co-Occurence Texture Features. IEEE Transactions on Geoscience and Remote Sensing 36: 298–300.
24. Nanni L, Brahman S, Ghidoni S, Menegatti E (2013) Different Approaches for Extracting Information from the Co-Occurrence Matrix. PLOSOne 8:1–9.
25. Huber MB, Bunte k, Nagarajan MB, Biehl M, L et.al (2012) Texture feature ranking with relevance learning to classify interstitial lung disease patterns. Artificial intelligence in medicine 56: 91–97.
26. Kermany DS (2018) Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell Press, Elsevier Inc. 172: 1122 – 1131.
27. Eze PU, Udaya P, Evans R (2018) Medical Image Watermark and Tamper Detection Using Constant Correlation Spread Spectrum Watermarking. International Journal of Computer, Electrical, Automation, Control and Information Engineering 12: 107–114.
28. Eswaraiah R, Reddy ES (2015) Robust medical image watermarking technique for accurate detection of tampers inside region of interest and recovering original region of interest. IET image Process 9: 615–625.
29. Edwards C (2019) Malevolent Machine Learning. Commun. ACM 62:13 – 15.
30. Ludewig E, Richter A, Frame M (2010) Diagnostic imaging evaluating image quality using visual grading characteristic (VGC) analysis. Vertinary Research Communications 34: 473–479.

Author Affiliation

Top

School of Computing and Information Systems, University of Melbourne, Australia