Journal of Intelligent Systems and Internet of Things (JISIoT)
Vol. 11, No. 01, PP. 1-11, 2024
DOI: https://doi.org/10.54216/JISIoT.110101

Intelligent Classification of JPEG files by Support Vector Machines with Content-based Feature Extraction

Rabei Raad Ali[1]*, Najwan Zuhair Waisi[1], Yahya Younis Saeed[1], Mohammed S. Noori[1], Eko Hari Rachmawanto[2]

[1]Department of Computer Engineering Technology, Northern Technical University, 41000, Mosul, Iraq, [2]Study Program in Informatics Engineering, University of Dian Nuswantoro, Semarang, Indonesia
Emails: rabei@ntu.edu.iq; najwan.tuhafi@ntu.edu.iq; dr.yahya.albugg@ntu.edu.iq; moh.sami@ntu.edu.iq; eko.hari@dsn.dinus.ac.id

Abstract

Nowadays, multimedia files play a basic role in supporting evidence analysis for making decisions about a crime through looking at files as a digital guide or evidence. Multimedia files such as JPG images are a common format because many documents and memorial images on laptops are valuable. In addition, many JPG images on Laptops are valuable and have fewer structure contents, making recovery possible when their file system is missing. However, intelligent systems for fully recovering corrupted JPG images into their original form is a challenging research issue. In this research, a support vector machine (SVM) as intelligent classifier algorithm is proposed to classify JPG or non-JEG image clusters as part of multimedia files. The SVM classifies the data clusters on three content-based feature extraction (entropy, byte frequency distribution, and rate of change approach to derive cluster features) methods to optimize the identification of JPG image content. The SVM classifier is applied using a radial basis and polynomial kernel functions in MATLAB software. The experimental results show that the accuracy of classification of the SVM classifier with the polynomial function is 96.21%, and the SVM classifier with the radial basis function is 57.58%.

Received: May 11, 2023 Revised: August 17, 2023 Accepted: November 19, 2023

Keywords: Intelligent Systems; JPG images; Support Vector Machine; Intelligent Classification; Machine Learning

1. Introduction

Traditional image recovery from corrupted file systems plays a significant role in digital forensics investigation. Many different pieces of data are preserved for investigation, of which bit-copy images of disk drives are a common way for the process of digital forensics. The images are mainly considered objective court evidence. Since the Joint Photographic Experts Group (JPEG) image format is less structured than other image formats (e.g., Bitmap, Portable Network Graphics, Graphics Interchange Format, and Tagged Image File Format). JPEG is a standard image file format having less structured contents that make its retrieval "recovery" possible when the file system is missing.

Recently, Studies estimated that the global number of Laptop users has increased to 3.2 billion from 2016 to 2023 [1]. Subsequently, multimedia files such as JPG images have become the current trend in retrieving important information from Laptops. A JPG is an international compression standard for continuous-tone still images for both grayscale and color types. The JPG compressed data format has three main portions which are: (i) A start of image marker, (ii) a frame, and (iii) an end of image marker.

There are two classes of segments of a JPG format: Marker segments and entropy-coded segments. Marker segments contain header information, tables, and other information required to interpret and decode the compressed image data, whereas entropy segments contain the entropy coded [2-4]. Marker segments always begin with a marker,

unique two-byte value markers that identify the function of the segment and distinguish various structural parts of the compressed data formats, a full list of these markers, and a full list of markers. Modern operating systems store files contiguous, but a missing file can easily become fragmented due to the operating system fragmentation process. In some situations, digital images are exposed to corruption due to operating file system processes or human errors [2]. Therefore, recovery of the corrupted JPG images is a challenging research issue when their file system is lost. Figure 1 shows a corrupted data file still existing in the clusters given to the original file in a metadata entry.

The main idea behind cluster data is to read each cluster in the dataset and then analyze its contents to find out some relationships connecting the clusters that belong to a particular file. It calculates metadata information like character counts or statistical information over the bytes of the clusters. These clusters are later reassembled to recover the original file. Modern computer device storage is divided into fixed-size storage units called blocks/clusters that carry data of a particular file [3]. Ali *et al.* [4] define a deformed file as a file that has been divided into multiple parts where all parts are stored in non-contiguous locations on a disk drive. A cluster is typically composed of 1–64 sectors with 512 bytes amount of data in each sector. Three techniques exist for identifying files: Signature, structure, and content. The benefits and drawbacks of each strategy are numerous. Consequently, only a few methods can deliver complete solutions. [6]. Storages of computer devices are split into fixed-size storage units called sectors. Figure 2 shows different groups of data storage units.

An essential approach that works on any file type's markers is the signature technique. By categorizing the affected portion of the file, a systematic technique is utilized to identify the corrupted file. However, non-contiguous cluster examples of deformed files are not classified by the signature and structural techniques [7]. Therefore, a content method tries to handle some instances of corrupted data files. In other words, it tries to identify and categorize any data clusters that might be present in a file and be useful for file recovery. There are three categories of file classification approaches: Signature-based, structure-based, and content-based. Each category has several advantages and disadvantages. Therefore, none of the categories are perfect and can provide comprehensive solutions [7].

A signature-based classification is a straightforward approach that has been successfully proven to carve contiguous files. This approach works on the header – footer data of the image, that is, header and footer [4]. However, in many cases, the signature block/cluster is damaged or disconnected due to storage damage or system fragmentation process [6]. Identifying the file type is an essential step in recovering files with missing or damaged file system information (Hammad *et al*., 2022). As a result, it lacks handling fragmented data in both consecutive, contiguous, non-consecutive, and non-contiguous order cases as explained in Section 2.5 (Ali, 2023). A structure-based classification approach has been used to carve the fragmented file, by identifying and deleting the fragmented portion of the file in the scan area [7].

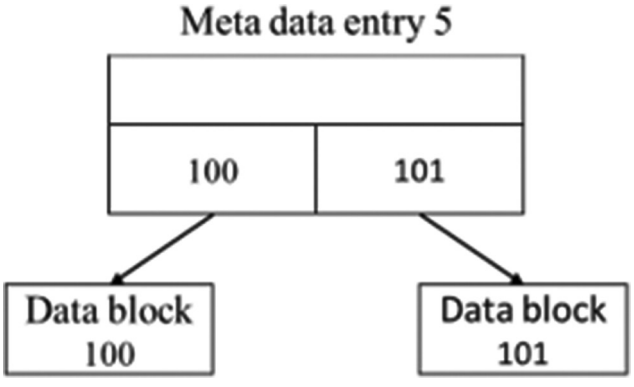

Figure 1: Metadata entry of a deleted file [5].

Figure 2: Data storage units of computer storage [7].

A content-based classification approach attempts to handle some cases of fragmented data files. This approach recovers files by analyzing the contents of the scan area. The content-based approach of file classification is still in its preliminary stages and not fully explored [1-6]. There are only a limited number of carving methods that reconstruct files based on the content of the scan area of images [9]. The study adopts file classification as an advanced approach that identifies files in the absence of file system metadata. Many identification methods have been introduced to identify the image files. Identification methods can be identified into the Lossless approach and Lossy approach. The first approach is more efficient than others to decrease the amount of data. The content approach of file classification is still in its preliminary stages. For instance, Ali *et al.* [4] improved identification using an extreme learning machine (ELM) classifier approach. The ELM applied based on content-based feature extraction to identify the contents of each data cluster. The ELM is applied as a binary machine learning classifier, whether a JPG file or a non-JPG file. There are a few numbers of identification methods that identify corrupted JPG images in the scan area [6-8]. In Ali *et al*. [7], the authors present a comparison between two classifiers which are an ELM method and an SVM method to solve the classification fragmentation problems in the JEG file. This problem is prevalent in forensics challenge datasets such as RABEI-2017. The following points are the contributions:

- Studying and analyzing the effect of the combinations of different feature vectors for improving JPG image cluster classification in a continuous series of JEG data clusters.
- The byte frequency distribution (BFD), entropy, and rate of change (RoC) are used as feature extraction methods. The SVM classifier applied with radial basis, and polynomial functions produce high accuracy and solve classification issues.
- The RABEI-2017 dataset tests and evaluates the SVM classifier, whether it is included as a JPG or non-JPG cluster.

The following sections are organized as follows: The next section presents the literature review. Section 3 discusses the materials and methods of this work. Section 4 shows the performance metrics. Section 5 discusses the implementation and results. Finally, Section 6 presents the conclusion and future work.

2. Related Work

Recently, there have been several image classification algorithms to develop JPEG with privacy protection for JPEG images. This section describes some of these classification methods to solve the identification process challenges using feature extraction methods and classification algorithms. Ali *et al.* [5] introduce a convolution neural network (CNN) which is: SVM, and long short-term memory (LSTM) to evaluate the signal classification for electrocardiography heartbeat. Furthermore, this study compares a CNN classifier to present the best classifier. The compassion shows that the LSTM is the best classifier.

Hammad *et al.* [6] propose a Malware Classification (MC) method to correctly identify malware. The proposed method has been used in five stages. First, it created 2D malware images; in second stage, the visual malware images used visualized malware pre-processing for scaling to appropriate the classification model's input size; in the third stage, the classification methods automatically classify based on Tamura and GoogLeNet. Last stage, the K-Nearest Neighbour (KNN), SVM, and ELM are employed to perform MC. A standard mailing unbalanced dataset is used by the proposed method for testing. The experimental result showed that the accuracy rate was extremely high in the proposed method.

Ali *et al.* [7] introduce a comparison between two classifiers which are an ELM method and an SVM method to solve the classification fragmentation problems in image files. Both methods automatically classify the images based on the RoC, BFD, and entropy features. The ELM and SVM automatically give a class label of fragmented data clusters to classify image files or another file type.

Tiwari and Chahande [9] propose an identification method to identify apple fruit infections using K-Means classification. A shading JEG imaging model is designed to find red-green-blue color JPG images from apples with different diseases such as apple rust, apple rot, apple scab, and apple blotch. The JPG image histogram and limerick bearing ratio of each apple disease test are used to select and study a few JPG image clusters. The outcomes demonstrate that the K-Mean classifier can precisely identify and categorize apple illnesses.

Kavitha *et al.* [10] introduce an automated identification method to assign a class label to malignant (cancerous) and melanoma (non-cancerous) cells using several machine-learning classifiers. The SVM, KNN, Random Forest (RF), and Naive Bayes (NB) automatically classify the images using evaluation measures of three methods median filter (MF), hybrid partial differential equation. Besides, by eliminating noise from the skin lesion, the MF is used to reduce

the resulting picture characteristics. The findings demonstrate that the SVM algorithm is superior to the KNN, RF, and NB classifiers in terms of classification accuracy.

Xia and Xu [11] design and implement a Regression-based Music Emotion Classification (RMEC) approach to retrieve and manage music information. The RMEC approach has a training process and testing process. In the training process, the polynomial regression, SVM, and K-Plane piecewise regression algorithms are used to achieve this approach. The music data are predicted and regressed in the test process to classify the VA value. The results show that the SVM algorithm improves accuracy and better by 84.9%.

Ahammed *et al.* [12] introduce a digital hair removal (DHR) system for the classification of JPG images. The DHR system works based on an SVM, decision tree, and KNN for classification. The black-hat transformation with Gaussian Filtering method is used to denoise the JPG images. Furthermore, to extract underlying input forms from the skin JPG images, the Automatic Grab cut Segmentation technique with statistical features and Gray Level Co-occurrence Matrix algorithm is utilized.

3. Materials and Methods

This section discusses the proposed architecture in this study. A JPG image and no JPG image are both given a class label using the SVM classifier technique as an intelligent classifier. The SVM classifier uses entropy, BFD, and RoC approaches to derive cluster features. The methods used a series of data clusters that have deformed data clusters. Figure 3 shows the main components of intelligent systems in this research.

In Figure 2, the first step of the work is to prepare an RABEI-2017 that serves to test and validate the SVM method. The dataset is used to check the ability of the SVM method to identify JPG clusters from other file types. The feathers extraction includes three content-based feature extraction methods to distinguish multimedia file types.

SNM method is represented as a fingerprint and used in the following component. The class desertion uses an SVM classifier to evaluate feature extraction based on two kernel functions. It classifies the files into JPG file clusters and non-JPG file clusters. The details of each component are explained in the following subsections. The following algorithm shows the classification pseudocode.

Step 1: Start;
Step 2: Read the image file cluster after the scan area;
Step 3: While is not the end of the file, do
Step 4: Classify file clusters by ELM or SVM based on Entropy, BFD, and RoC;
Step 5: Calculate the accuracy of ELM or SVM;
Step 6: End-while;
Step 7: Generate report;
Step 8: End;

In the following sections, we will dissect the various components, strategies, and principles that form the foundation of the study activities outlined in this paper.

3.1. Dataset description

Selecting a dataset is an important part of analyzing research. The RABEI-2017 dataset is prepared to provide an environment for testing identification problems that are addressed in this research. The dataset has eight files that are used to recognize different cases for both non-fragmented and fragmented problems. A RABEI-2017 dataset is used to test and evaluate the intelligent classification algorithm in this research. It has eight multimedia files, which are four JPG image files and four Microsoft Word files, as shown in Figure 4.

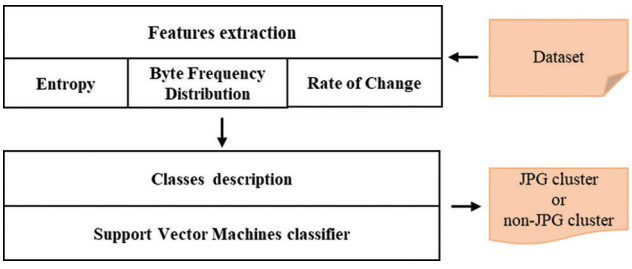

Figure 3: The architecture of intelligent classification.

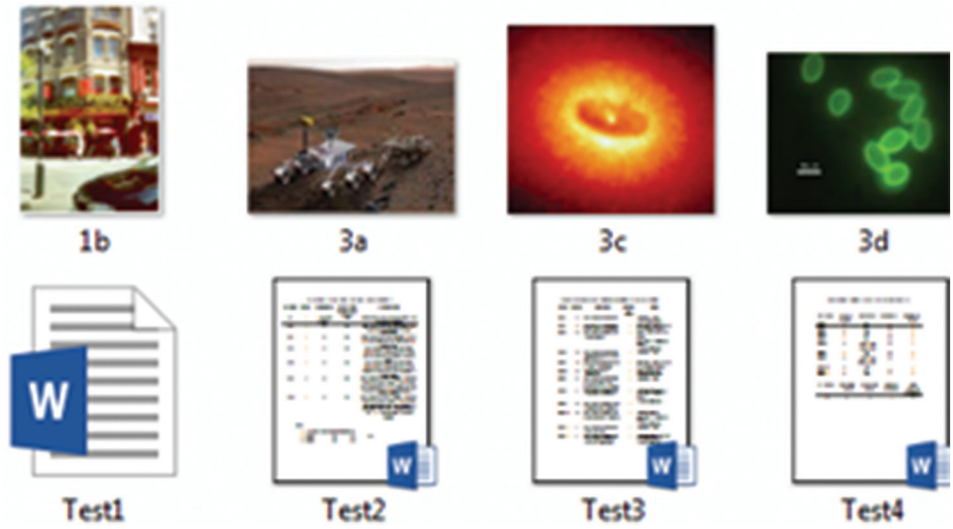

Figure 4: The multimedia files in the dataset [4].

In this dataset, the total number of clusters in the dataset is 400 clusters (200 JPG clusters and 200 non-JPG clusters). The cluster utilized during this study comprises 512 bytes, and 513 features are used to represent it. The dataset is split into 30% for training and 70% for testing.

3.2. Features extraction

Feature extraction is a process in which relevant and meaningful information is extracted from raw data to create a reduced representation that retains the most important aspects of the original data. It is a fundamental step in data preprocessing, pattern recognition, and machine learning tasks. The goal of feature extraction is to transform complex and high-dimensional data into a simplified and more manageable form, which can then be used as input for various algorithms and analyses. Furthermore, the feature extraction works as a fingerprint of the file's content [13]. Entropy, BFD, and RoC as content-based feature extraction methods are used in this research.

- Entropy: It quantifies the amount of uncertainty or randomness in a set of data. It is one of the potential content-based features to distinguish multimedia file clusters [14]. The entropy value is between 0 to 255 bytes and generates a byte with a value between zero and one. Equation (1) finds the possibility of occurrences of features rate.

$$p(i) = N(i) / l \tag{1}$$

where *p(i)* is the possibility of occurrences, *N(i)* is the number of occurrences of a byte value, and *l* is the size of a file defoamed.

Then, the entropy of deformed clusters can be represented by Equation (2)

$$E = -\sum P(i) \log^{P(i)}, 0 < P(i) \leq 1 \tag{2}$$

- BFD: It also known as Byte Histogram is a statistical representation that shows the frequency of occurrence of each byte value within a given set of data. It produces a histogram for file clusters based on byte value rates without the order of the bytes [15]. The features rates contain analyzing several byte values to find a centroid model. The BFD contains 256 features, as defined in Equation (3).

$$P(i) = f(i) \tag{3}$$

- RoC: It refers to the speed at which a quantity or variable changes about another variable. It is a fundamental concept in mathematics and is often used to describe how one variable is changing concerning another variable. The RoC can be expressed as a ratio or a derivative, depending on the context. It rates the difference between every consecutive byte value as represented in Equation (4) [16]. It tracks the appropriate byte value in their corresponding clusters from their orders. It does not specify the direction of the stream change of byte value, where c and c_{i+1} are the value of the consecutive byte.

$$RoC_i = | c_i - c_{i+1} | \quad (4)$$

3.3. Classes description

The major usage of the classes' description is in the identification stage, in which it classifies the file clusters [17-25]. It is the procedure of assigning a class label for every input of a binary classification or multi-classification cases by looking at a set of attributes [4].

A support vector machine (SVM) classifier is a powerful and versatile machine-learning algorithm used for both intelligent classification and regression tasks. It belongs to the category of supervised learning algorithms and is particularly effective for solving binary classification problems, where the goal is to categorize data points into one of two classes. The main key advantages of the SVM method are:

- Effective in high-dimensional spaces: It works well in cases where the number of features is larger than the number of samples. This makes it suitable for tasks involving intelligent classification, JPG image recognition, and other complex data types.
- Robust to overfitting: It aims to maximize the margin between classes, which often leads to a better generalization of new, unseen data. This can help prevent overfitting compared to some other algorithms.
- Flexibility with kernel functions: The use of different kernel functions allows SVM to model complex relationships in the data, even if they are not linear. This makes it capable of handling diverse data distributions.
- Global Optimum: SVM seeks to find the hyperplane that maximizes the margin, and this optimization problem typically has a unique solution, ensuring a global optimum.

In addition, the SVM classifier is a versatile tool that excels at solving binary classification problems by finding an optimal hyperplane that maximizes the margin between classes. Its ability to handle high-dimensional and non-linear data makes it a valuable choice for a wide range of applications. Figure 5 shows the SVM classifier architecture.

This research uses the SVM classifier algorithm to classify JPG file clusters. The SVM is defined as a set of related supervised learning models used for classification problems [6,8]. It has a high-quality generalization ability for solving linear and non-linear classification problems [4]. As shown in the above figure, the SVM classifier has multi-layers of artificial neurons which are input-layer, hidden-layer, and output-layer [19].

It focuses on solving the optimization issue with the training test, as represented in Equation (5).

$$(x_1, y_1), \ldots . (x_\lambda, y_\lambda), x \in R^n, y \in \{1, -1\}) \quad (5)$$

where (x_1, y_1), (x_2, y_2)., (x_l, y_l) as training dataset, where x_1 is a sample data and $x_1 \in R^n$ when y_λ is its label, $i = 1,2\ldots,$ l and $y \in \{-1,1\}$.

The SVMs are used for mapping the input vectors to a feature space to classify the transformed data cluster through a linear function using the following Equation (6).

$$F(x) = \text{sign}\left(\sum_{i=1}^{l} y_i \propto_i^* K(x_i, x) + b\right) \quad (6)$$

where $\propto_i^*$ are the coefficients and $K(x_i, x)$ is a kernel value. The following Equation represents a linear decision surface.

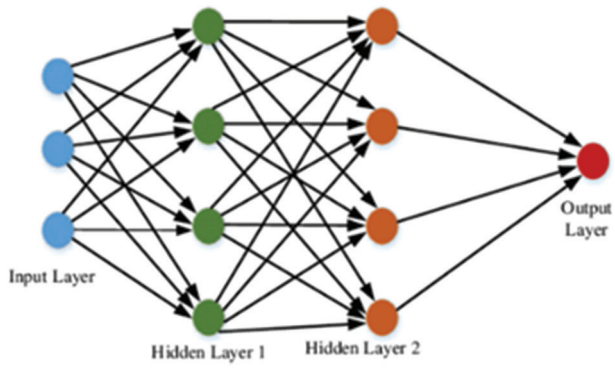

Figure 5: The multimedia files in the dataset [4].

$$(w.x) - b = 0 \quad (7)$$

where the $w.x \in R^{Nn}$ and b is the boundary among + and -.

The kernel functions map the data to build a more accurate classifier algorithm. The polynomial and radial basis functions are selected as the nonlinear kernel in the SVM classifier.

- Polynomial Kernel: It is a type of mathematical function used in machine learning, particularly in the context of SVMs and other kernel-based algorithms. It is essential for transforming data into higher-dimensional spaces, making it easier to find a separating hyperplane or decision boundary between different classes of data that might not be linearly separable in the original feature space. It is the similarity of training samples in a feature space over the original variables' polynomials [20]. The polynomial kernel is defined as Equation (8).

$$K(x_i, x_j) = (\gamma x_i T x_j + r)d \ , \gamma > 0 \quad (8)$$

- Radial Basis Function Kernel: It is a type of kernel used in machine learning algorithms, particularly in the SVM method. This kernel is especially effective for capturing complex and non-linear relationships in data. It is one of the most widely used kernels. It is employed when there is no information about the data [20]. The following equation shows two points x and y, on this kernel.

$$K(x, y) = exp\left[-\frac{x - y^2}{2\sigma^2} \right] \quad (9)$$

where, σ is variance and $x-y^2$ is Euclidean distance between x_1 and x_2.

4. Performance Metrics

In this intelligent system, the classification performance of most classifier algorithms is evaluated using three measures which are: accuracy, precision, recall, and F measure [21]. Figure 4 illustrates an estimation of the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) parameters.

- Accuracy: it is a commonly used metric in the context of classification problems in machine learning algorithms. Accuracy measures the proportion of correctly predicted instances among all instances in a dataset. Furthermore, accuracy indicates how often a model's predictions match the true labels of the data. It is the properly identified samples and the total number of tests as presented in Equation (10).

$$\text{Acc} = \left[\frac{TP + TN}{TP + FN + FP + TP} \right] x100\% \quad (10)$$

- Precision: It is a fundamental metric used in binary classification to assess the quality of a model's positive predictions. It measures the proportion of correctly predicted positive instances between all instances that the model predicted as positive. It is the properly identified positive tests and the total number of predicted positive tests as presented in Equation (11).

$$Pre = \left[\frac{TP}{TP + FP} \right] x100\% \quad (11)$$

- Recall: It is a metric commonly used in binary classification to measure a model's ability to correctly identify all positive instances from the dataset. It provides insight into the model's capability to capture all instances that truly belong to the positive class. It is the total number of properly identified positive tests classified correctly, as presented in Equation (12).

$$Rec = \left[\frac{TP}{TP + FN} \right] x100\% \quad (12)$$

- F1 measure: It is a metric that combines precision and recalls providing a single value that balances both aspects of a model's performance in binary classification tasks. It is particularly useful when there is a trade-off between precision and recall, and you want to evaluate the overall effectiveness of a model. The weighted average returns a single value by combining recall and precision as presented in Equation (13).

$$F\bar{m}easure = 2x\left[\frac{(Pre \times Rec)}{(Pre + Rec)}\right]x100\% \quad (13)$$

where the true positive is TP, the true negative is TN, the false negative is FN and the false negative is FP.

To calculate the FP, the present JEG image cluster may be seen as negative, and the JEG image cluster may be seen as positive. To determine the FN, the existing JEG image cluster may be seen as positive and the expected malware kinds as negative. Figure 6 explains the confusion matrix [22]. It is a table that is often used to describe the performance of a classification model on a set of data for which the true values are known. It helps in understanding how well a classification model is performing by showing the counts of TP, TN, FP, and FN predictions. These parameters represent the confusion matrix (CM) of multi-classification tasks in contrast to the conventional CM of binary classification tasks.

5. Implementation and Results

MATLAB software version 2017 implements the SVM classifier algorithm. The implementation's main challenge was extracting the image features because some of these images have fragmentation problems. To solve this problem, we need to recover these images fully. The SVM has many types of activation functions and settings. Choosing the best SVM model that meets this work needs requires initial tests. In addition, the SVM model design that provides good generalization needs to be configured. The 10-fold cross-validation process is assessed to produce accurate results. Randomly, the data are divided into testing and training.

The training is 70% set and testing is 30% set and can be created. In the SVM model, a 10-fold cross-validation with 600 hidden neurons is adopted to avoid over-fitting during the training procedure and selecting the best-performing prototypes. Table 1 shows the confusion matrix results of the polynomial and radial basis functions.

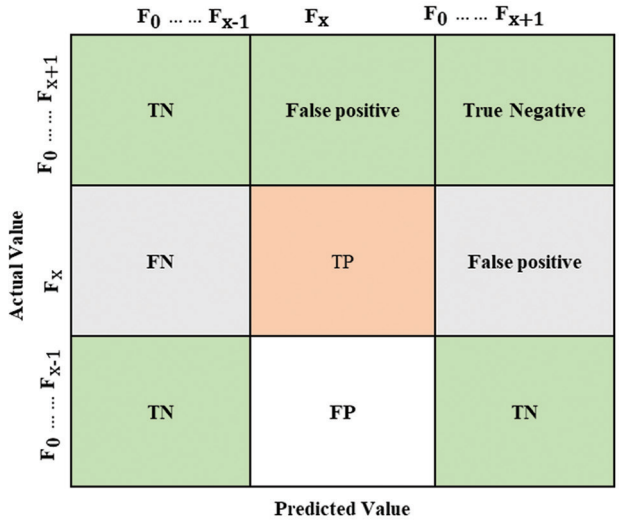

Figure 6: Confusion matrix.

Table 1: The confusion matrix

JPG	Non-JPG
Radial basis function	
1	0
0.8485	0.1515
Polynomial	
0.9545	0.0455
0.0303	0.9697

Table 2: The SVM classification measures

Accuracy	Preciseness	Rendering	F measure
Radial basis function			
0.5758	0.5758	0.5758	0.5758
Polynomial			
0.9621	0.9621	0.9621	0.9621

SVM: Support vector machine

Table 3: The classification analysis

	Radial basis function		Polynomial	
	JPG	Non-JPG	JPG	Non-JPG
Actual	66	56	63	2
Predicted	0	10	3	64
Error classification	66	66	66	66
Total File Cluster	132		132	
Accuracy	57.58%		96.21%	

Table 2 summarizes a comprehensive overview of the performance evaluation metrics, derived from the application of the SVM method. These metrics provide valuable insights into how well the SVM model is performing across various aspects of classification.

The outcomes demonstrate in this intelligent system that the SVM method using Polynomial functions outperforms the radial basis function in terms of accuracy when identifying clusters of JPG and non-JPG data. The categorization analysis of the RABEI-2017 dataset's results is shown in Table 3.

As shown in Table 3, the performance of the SVM method is presented concerning two different kernel functions: The polynomial function and the radial basis function. These kernel functions play a crucial role in transforming the data into higher-dimensional spaces, allowing the SVM method to effectively separate different classes. The reported accuracy values for the SVM algorithm are 96.21% for the polynomial function and 57.58% for the radial basis function. These accuracy percentages provide insight into how well the SVM algorithm is performing with each kernel type.

6. Conclusion

This study focuses on multimedia file cluster classification when its file system is absent. The study's main objective is to recognize the file type of every cluster. Three feature extracts are applied to extract fingerprints for each data cluster dataset to be investigated. The SVM is employed for the RABEI-2017 dataset. This study achieved the intelligent classification performance for JPG and non-JPG cluster classification. The results showed that the accuracy of intelligent classification of the SVM classifier with the polynomial function is 96.21%, and the SVM classifier with the radial basis function is 57.58%. Therefore, this research has demonstrated that SVM can be used efficiently in JPG file recovery. Therefore, effective and intelligent classification methods are needed to improve the accuracy rate of JPG image cluster classification. To overcome some of the limitations of JPG image classification, this work adopts content-based feature extraction in corrupted JPG image cluster classification. In future work, this intelligent classification model can be used for other data recovery fields for classification tasks, such as medical JPG images, drone imaging, and another important field. Furthermore, the new test set prepares multi-JPEG files intertwined and non-linear scenarios that are yet to be addressed by this research.

References

[1] C. A. Sari, I. P. Sari, E. H. Rachmawanto, E. Proborini, R. R. Ali, and I. Rizqa, "Papaya fruit type classification using LBP features extraction and naive bayes classifier", in*: 2020 International Seminar on Application*

for Technology of Information and Communication (iSemantic), IEEE, United States, pp. 28-33, Sep, 2020.

[2] R. R. Ali, K. M. Mohamad, S. A. P. I. E. E. Jamel, and S. K. A. Khalid, "A review of digital forensics methods for JPEG file carving", *Journal of Theoretical and Applied Information Technology*, vol. 96, no. 17, pp. 5841-5856, 2018.

[3] D. Kröger, J. Peper, and C. Rehtanz, "Electricity market modeling considering a high penetration of flexible heating systems and electric vehicles", *Applied Energy*, vol. 331, pp. 120406, 2023.

[4] R. R., Ali, W. S. Al-Dayyeni, S. S. Gunasekaran, S. A. Mostafa, A. H. Abdulkader, and E. H. Rachmawanto, "Content-Based Feature Extraction and Extreme Learning Machine for Optimizing File Cluster Types Identification". in: *Future of Information and Communication Conference*, Springer, Cham, pp. 314-325, Mar. 2022.

[5] O. M. A. Ali, S. W. Kareem, and A. S. Mohammed, "Evaluation of electrocardiogram signals classification using CNN, SVM, and LSTM algorithm: A review", in: *2022 8th International Engineering Conference on Sustainable Technology and Development* (*IEC*), pp. 185-191, Feb. 2022.

[6] N. Hammad, N. Jamil, I. T. Ahmed, Z. M. Zain, and S. Basheer, "Robust malware family classification using effective features and classifiers", *Applied Sciences*, vol. 12, no. 15, p. 877, 2022.

[7] R. R., Ali, L. N. Dawd, S. A. Mostafa, E. H. Rachmawanto, and M. A. Jubair, "Content-based Feature Extraction and Extreme Learning Machine for Optimizing File Cluster Types Identification", in: *International Conference on Innovative Computing and Communications*. Springer, Germany, pp. 1-12, 2023.

[8] L. Zhang, D. Zhang, and F. Tian, "SVM and ELM: Who Wins? Object Recognition with Deep Convolutional Features from ImageNet". in*: Proceedings of ELM-2015*, Springer, Cham, vol. 1, pp. 249-263, 2016.

[9] R. Tiwari, and M. Chahande, "Apple Fruit Disease Detection and Classification Using K-Means Clustering Method". in: *Advances in Intelligent Computing and Communication*, Springer, Singapore, pp. 71-84, 2021.

[10] P. Kavitha, V. Jayalakshmi, and S. Kamalakkannan, "Classification of Skin Cancer Segmentation using Hybrid Partial Differential Equation with Fuzzy Clustering based on Machine Learning Techniques". in: *2022 International Conference on Edge Computing and Applications*, IEEE, United States, pp. 1-8, Oct. 2022.

[11] Y. Xia, and F. Xu, "Study on music emotion recognition based on the machine learning model clustering algorithm", *Mathematical Problems in Engineering*, vol. 2022, p. 9256586.

[12] M. Ahammed, M. Al Mamun, and M. S. Uddin, "A machine learning approach for skin disease detection and classification using image segmentation", *Healthcare Analytics*, vol. 2, p. 100122, 2022.

[13] X. Liu, J. Zhang, and Z. Pei, "Machine learning for high-entropy alloys: progress, challenges and opportunities", *Progress in Materials Science*, vol. 7, p. 101018, 2022, doi: 10.48550/arXiv.2209.03173

[14] W. Qiu, R. Zhu, J. Guo, X. Tang, B. Liu, and Z. Huang, "A new approach to multimedia files carving". in: *Bioinformatics and Bioengineering (BIBE). 2014 IEEE International Conference on*, IEEE, United States, pp. 105-110, Nov. 2014.

[15] R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal, "Cancer statistics", *CA: A cancer Journal for Clinicians*, vol. 73, no. 1, pp. 17-48, 2023, doi: 10.3322/caac.21763

[16] J. Fan, J. Lee, and Y. Lee, "A transfer learning architecture based on a support vector machine for histopathology image classification", *Applied Sciences*, vol. 11, no. 14, p. 6380, 2021, doi: 10.3390/app11146380

[17] D. K. Choubey, S. Tripathi, P. Kumar, V. Shukla, and V. K. Dhandhania, "Classification of Diabetes by Kernel based SVM with PSO". in: *Recent Advances in Computer Science and Communications* (*Formerly: Recent Patents on Computer Science*), vol. 14, Bentham Science Publishers, Sharjah, pp. 1242-1255, 2021.

[18] S. Ghosh, A. Dasgupta, and A. Swetapadma, "A study on support vector machine based linear and nonlinear pattern classification". in: *2019 International Conference on Intelligent Sustainable Systems* (*ICISS*), IEEE, United States, pp. 24-28, Feb. 2019.

[19] É. Bonnet, E. J. Kim, A. Reinald, S. Thomassé, and R. Watrigant, "Twin-width and polynomial kernels", *Algorithmica*, vol. 84, no. 11, pp. 3300-3337, 2022.

[20] M. Ahammed, M. Al Mamun, and M. S. Uddin, "A machine learning approach for skin disease detection and classification using image segmentation", *Healthcare Analytics*, vol. 2, p. 100122, 2022, doi: 10.1016/j.health.2022.100122

[21] G. P. Burrai, A. Gabrieli, M. Polinas, C. Murgia, M. P. Becchere, P. Demontis, and E. Antuofermo, "Canine mammary tumor histopathological image classification via computer-aided pathology: An available dataset for imaging analysis", *Animals* (*Basel*), vol. 13, no, 9, p. 1563, 2023, doi: 10.3390/ani13091563

[22] J. Wu, "Small Sample Datasets Build Powerful Image Classification Models". in: *International Conference on Cloud Computing, Performance Computing, and Deep Learning (CCPCDL 2022)*", vol. 12287. SPIE, Bellingham, pp. 464-471. Oct. 2022.

[23] P. Elangovan, and M. K. Nath, "En-ConvNet: A novel approach for glaucoma detection from color fundus images using ensemble of deep convolutional neural networks", *International Journal of Imaging Systems and Technology*, vol. 32, no. 6. pp. 2034-2048, 2022. doi: 10.1002/ima.22761

[24] L. B. Handoko, D. R. I. M. E. H. Setiadi, C. A. Rachmawanto, R. R. Sari, and R. R. Ali, "An analysis of imperceptibility and robustness performance in CRT image watermarking based on color space theory", *Journal of Physics: Conference Series*, vol. 1501, no. 1, p. 012015, Mar. 2020.