Bioinformatics-Inspired Analysis for Watermarked Images with Multiple Print and Scan

By

Abhimanyu Singh Garhwal

A thesis submitted to Auckland University of Technology in fulfilment of the requirements for the degree of Doctor of Philosophy

September 2017

Acronyms Used in This Thesis

BIIA - Bioinformatics-Inspired Image Analysis

BIIIA - Bioinformatics-Inspired Image Identification Approach

BIIIG - Bioinformatics-Inspired Image Grouping Approach

DNA – Deoxyribonucleic Acid

MPS – Multiple Print and Scan

MSA – Multiple Sequence Alignment

NW - Non-Watermarked

NWA – Needleman Wunch Algorithm

NWD – Non-Watermarked and Degraded

NWND – Non-Watermarked and Non-degraded

PSA – Pairwise Sequence Alignment

SWA – Smith Waterman Algorithm

W – Watermarked

WD – Watermarked and Degraded

WND – Watermarked and Non-Degraded

Abstract

Image identification and grouping through pattern analysis are the core problems in image analysis. In this thesis, the gap between bioinformatics and image analysis is bridged by using biologically-encoding and sequence-alignment algorithms in bioinformatics. In this thesis, the novel idea is to exploit the whole image which is encoded biologically in DNA without extracting its features.

This thesis proposed novel methods for identifying and grouping images no matter whether having or not having watermarks. Three novel methods are proposed. The first is to evaluate degraded/non-degraded and watermarked/non-watermarked images by using image metrics. The bioinformatics-inspired image identification approach (BIIIA) is the second contribution, where two DNA-encoded images are aligned by using SWA algorithm or NWA algorithm to derive substrings, which are exploited for pattern matching so as to identify the images having a watermark or degradation generated from MPS. The outcomes of identification affirm the capability of BIIIA algorithm. Furthermore, it asserts that DNA-based encoding is the best way for digital images as well as SWA algorithm is the best one for the sequence alignment.

The last one is the bioinformatics-inspired image grouping approach (BIIGA), where the DNA-encoded images are aligned by using multiple sequence alignment (MSA), which is exploited by using the phylogenetic tree to group the watermarked / non-watermarked and degraded / non-degraded images; the resultant analysis confirms the potential of BIIGA algorithm. All three methods are empirically verified and validated by using real datasets.

Keywords: Multiple print and scan, multiple sequence alignment, local pairwise and global alignment, image quality metrics, image analysis, pattern matching, phylogenetic tree, bioinformatics tool.

Table of Contents

List of Figures

List of Tables

Attestation of Authorship

“I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another author (expect where explicitly defined in the acknowledgements), nor material which to a substantial extension has been submitted to the award of any other degree or diploma of a university or other institute of higher learning.”

Signature: Date: 01 September 2017

Acknowledgement

I am very indebted to my primary supervisor and my secondary supervisor who have inspired, supported, and educated me over the past three and half years. I extend my gratitude to both of them in helping me to complete my PhD in image analysis. Their constant support, encouragement and guidance have helped me to develop a deep understanding of the subject. Apart from the decision making, communications and professionalism, the most important thing, I have learnt from them is to be an academic person. Their remarkable suggestions to approach research problem and excellent sense of strategy have guided me all through my research.

I am very grateful to Bumjun Kim, Senior Technician, SCMS for all his kindness and tremendous support; Ramon Lewis, the technician of SCMS. Vijay Naidu and Sreenivas Sremath Tirumala deserve special acknowledgement for their guidance, friendship and a lot of thoughtful discussions because they deserve it. Thanks to Ruchi Vishnoi, Ekanki Vishnoi and Shobhraj Meena for their support and motivation.

On a personal level, I would like to express my love and gratitude to my parents, brothers and sisters for their continuous understanding, support and encouragement during my PhD study. I appreciate Falgun Shah for his immense support, who has offered me since I started my PhD study. I cannot end without mentioning the name of Mr. Vikram Golcha, whose support helped me to start my PhD in New Zealand.

Lastly, I offer my regards and blessings to all of those who have supported me in any way in the completion of this study.

Abhimanyu Singh Garhwal
Auckland, New Zealand
September 2017

Chapter 1

Introduction

This chapter starts with the background and motivation of the research, a brief introduction to sequence analysis with sequence alignment and sequence visualization in bioinformatics, image analysis, watermarking and multiple print-and-scan image degradation. The scope of this thesis includes bioinformatics tools and approaches to resolve the pattern matching in image analysis. The contribution of this thesis is explained in the last section.

1.1 Motivation

The nature of research methods can be understood by “Neo-Darwinism,” i.e., modern evolutionary analysis introduced the connection between two important discoveries: the unit of evolution (genes) with the mechanism of evolution (natural selection). This biological evolution inspired and led to the foundation for the emergence of a relatively late field of computing known as “evolutionary computing” for answering complex issues in computer science. For example, the particle swarm optimisation (PSO) (Nahdliyah, Fitriyani, & Biyanto, 2017; Zhang & Rahmat-Samii, 2016), ant colony optimisation (ACO) (Gajalakshmi & Srikanth, 2016; Dorigo, Birattari, & T. Stutzle, 2006) and genetic algorithm (Yousef, et al., 2016) in a variety of fields are now well developed and recognised.

Evolutionary approaches focus at the level of an individual entity or in populations of entities like bioinformatics evolutionary computing and the nature-inspired computing approach. From the last 40 years, researchers have been using pattern matching tools for resolving bioinformatics problems on an enormous scale, which has become a highly profitable industry. It seems that there is a unidirectional flow of interactions, i.e., from pattern matching to bioinformatics. A relatively less investigated and new area of research is to reverse the interaction, i.e., from bioinformatics to pattern matching. That is how the latest development in our understanding of advanced bioinformatics tools can be used for pattern-matching tasks; in particular, how alignment algorithms and biological image representation can be used to develop a novel algorithmic solution for image analysis by using pattern matching. To the best of our knowledge, this unconventional way of thinking is relatively new and very little work is reported in literature so far, e.g., for 2D shape classification (Bicego & Lovato, 2016; Lovato, Milanese, Centomo, Giorgetti, & Bicego, 2014; Lovato & Bicego, 2012), 2D shape recognition (Bicego & Lovato, 2012) , image matching (Bicego, Danese, Melzi, & Castellani, 2015; Kim, Chang, Liu, Lee, & Lee, 2009; Kim, Chang, Lee, & Lee, 2010) and video genome project (Bronstein, Bronstein, & Kimmel, 2010).

The motivation behind this research project on identification and grouping images explained in this thesis can be abridged in two situations: one for identification and the other for grouping degraded (copies made by multiple scan and print) / non-degraded images and watermarked/non-watermarked images.

- *A large number of digital images are a big problem in World Wide Web*

  Basically, there are two main reasons for this. The first is human beings boosted easily to access all forms of digital content when they need it, which leads to the explosion of digital images in quantity. Secondly, the Internet becomes super cheap that enables a person to easy access for uploading and downloading digital images. Consequently, piracy and ownership as well as counterfeiting have turned out to be more pervasive, which causes a big threat to digital images ownership. A typical solution to the ownership identification is watermarking. However, there is huge development in the area of watermarking for digital right management but this option becomes less favourable when a hacker removes or damages the watermarks by using malicious software. Additionally, increasing watermark types and watermarking algorithms will add more complexities during identification.

  From a bioinformatics perspective, a solution of this problem is to develop a biologically-based pattern analysis algorithm where this pattern provides the DNA-like information. This pattern may be unique for a particular category of watermarked images. Additionally, this pattern analysis algorithm will be independent on the process of watermark embedding, thus allowing one universal approach for image identification having all types of watermarks.

  However, even if we addressed the solution that will come out as a good idea, how do we guarantee that these patterns cannot be counterfeited and duplicate patterns be created?

- *The excessive technical advancement in making high-quality copies of images by using scanners and printer is a big problem for genuine use of digital images. Grouping the original, copied or degraded images becomes an issue.*

Comprehensively, there is three primary reasons for this. The initial one is that the high-quality scanners and printers have turned out to be no more luxury. Second, all forms of content (image, documents, audio, and video) are transferred and shared digitally, and have turned out to be the more pervasive, which increased the threat of counterfeiting, piracy and security. The third reason is that our human eyes have limited range of vision because it is hard to discriminate and group the counterfeited and original one by using our naked eyes.

If we want to filter and reorganise images according to their categories by using bioinformatics tools, which will emphasize on the requirement of a biologically-based pattern analysis algorithm. There are a number of components to the both cases, in spite of the differences in the applications of this approach. The first is the biologically-based pattern analysis. The second is for grouping images on the basis of patterns extracted. The extracted patterns will be used to determine the image groups and then placed in the expected ones. The third is that images are identified in both cases; the second scenario is a natural extension of image identification.

To implement the biologically-based pattern analysis, the images must be converted into biological format, i.e., four letters of DNA: "A", "T", "C", and "G". In the bioinformatics, the idea of patterns extraction between two DNA sequences usually is implemented by using pairwise sequence alignment. For more than two DNA sequences, multiple sequence alignment is taken into consideration.

The goal of this thesis is to inspect the bioinformatics tools and approaches at DNA level that can be employed for developing new bioinformatics-inspired models for pattern matching in image analysis. Our focus will be image analysis for evaluation and discrimination (i.e., identification and grouping) of watermarked and degraded images by using MPS as well as non-degraded / watermarked images by using bioinformatics, i.e., bioinformatics-inspired image analysis (BIIA) to identify the methods and strategies that achieve acceptably accurate identification and grouping. That leads to divide BIIA further into two parts: bioinformatics-inspired image identification approach (BIIIA) and bioinformatics-inspired image grouping approach (BIIGA).

An evidence for relative lack of research in "Bioinformatics Inspired Image Analysis of Watermarked Multiple Print and Scan Images" comes from a straightforward series of searches using Google Scholar.

- There are about 3.23 million hits for "bioinformatics" and 2.15 million hits for "image analysis."
- There are more than 65 thousands hits for "bioinformatics" + "image analysis," and around 21 hits for "bioinformatics inspired" + "image analysis."

- There are around thirty-two thousands and eight hundred hits for “watermarked” and 56 for “Bioinformatics” + “image analysis” + “watermarked.”
- There are two hits for “multiple print and scan” and zero hits for “bioinformatics” + “image analysis” + “watermarked” + “multiple print and scan.”
- There are more than 6.5 millions hits for “images” and zero hits for "bioinformatics" + "image analysis" + "watermarked" + "multiple print and scan " +"images".

These are certainly peripheral results. Nevertheless, the principal reason behind the searches persists and can be expressed as follows: while bioinformatics techniques are well developed and adopted for distinguishing species that are under observation for identification and grouping (typically, for conserved-region identification). There has not been any endeavor to draw inspiration from bioinformatics on how to use DNA for discriminating non-watermarked / watermarked image degraded by using MPS. Notably, even though the massive number of cases for conserved regions to identify the biological evolutionary relationship between the DNA biological sequences of humans and mammals, alteration in them will cause potential anatomical and behavioral difference. The degradation of watermarked / non-watermarked images (by multiple print and scan) studies in literature has neglected them. The inspiration of driving this thesis is to investigate what seems to be a gap in watermarked / non-watermarked image degradation by using multiple print and scan: modelling the different aspects of degraded and watermarked / non-watermarked images by using MPS for evaluation and discrimination (i.e., for identification and grouping) through motivation from bioinformatics. Further affirmation for this argument is supported by the literature reviews in Chapter 2.

The idea of the introducing bioinformatics alignment algorithms and biological image representation for image analysis was needed to understand the unusual relationship between them. That leaves many open questions. How can we use the bioinformatics algorithms in image analysis? How can we represent images biologically? Is biological representation used for images, does it deserve or not? If so, then how we can identify and group the images using bioinformatics tools, in particular, focusing on the identification and grouping of degraded and watermarked

(W) as well as non-watermarked images, non-degraded and watermarked / non-watermarked images?

Which specific characteristics of bioinformatics have been driven the background of this research project, particularly discussed in later chapters? Coming back to the series of Google search, it is also understandable that there is much unpredictability, what and how bioinformatics tools for pattern matching can be used for evaluation and discrimination (i.e., identifying and grouping) of the watermarked / non-watermarked images degraded by using MPS-based image analysis. Furthermore, it is required to explain what is meant by using "evaluation," "discrimination (i.e. for identification and grouping)" and "bioinformatics-inspired image analysis." So, before investigated the proposed approach further, it is necessary to elaborate understanding the concepts to lay a clear-cut base for the proposed algorithms and results.

1.2 Scope

1.2.1 Concepts of Bioinformatics

Bioinformatics is a combination of computer science and biology that can be explained as the intersection of biology and information technology (I.T.), a tool for data mining in biological databases or biological information management. The term bioinformatics was first used by Paulien Hogeweg and Ben Hesper in the beginning of the 1970s, who addressed the work "The study of informatics processes in biotic systems" as bioinformatics (Hogeweg, 2011; Hogeweg & Hesper, 1978; Hogeweg, 1978). This description leads towards a new field that is parallel to the physical process (biophysics) and a chemical process (biochemistry) of biological systems (Hogeweg, 2011). Pattern analysis was introduced in bioinformatics around the 1960s (Hagen, 2000) and Paulien Hogeweg developed an integrated set of non-supervised and supervised pattern analysis BIOPAT systems for biotic systems (Hogeweg, 2011). The proposed research uses pattern analysis to investigate data derived from biologically-represented images. Our questions around finding some conserved patterns that are remaining after degradation by using multiple rounds of print and scan in an image.

At the beginning of this century, computing researchers' paradigm was shifted from macro to micro level for developing novel methods and algorithms (Libeskind-Hadas & Bush, 2013). The rise of quantum computing (Vandersypen & Leeuwenhoek, 2017; Khan, Saha, & Pal, 2017), DNA computing (Rondelez & Woods, 2016), and bioinformatics computing (Lizhen, Zhong, Weikuan, & Meng, 2017; Alex, Oswaldo, Antonio, Cornejo, & Perkins, 2016) are a few instances of this paradigm shift. Most of the hypothesis was mining deeper into natural systems, resulting in the development of better algorithms. Here, the betterment represents the improved effectiveness regarding novelty for contributing to literature and the quality of output. In this thesis, we will investigate bioinformatics alignment algorithms, biological image representation, watermarked / non-watermarked and degraded images by using MPS to develop novel methods in the field of image analysis and bioinformatics. The mission will be to go “deeper” into the bioinformatics for image analysis than the achieved so far with bioinformatics alignment algorithms and biological image representation to see what betterments are possible if any.

The bioinformatics field is an interdisciplinary area that deals with the development of software tools and methods with the help of biological data. It is interdisciplinary because it includes mathematics, computer science, statistics and engineering for understanding and analysing biological data. Bioinformatics is mainly used to understand genetic diseases, adaptations, desirable effects in agriculture crops, etc. A deeper understanding of the biological process is the primary goal of bioinformatics by developing and applying computation techniques rigorously, for example, data-mining, machine learning, pattern recognition, etc. Sequence alignment, gene expression and protein structure predictions, design and discovery, gene finding, protein-protein interactions and evolution modelling are different processes where extensive research work is going on. Three important subdisciplines of bioinformatics are:

- Efficient development and implementation of computer program that allow us to manage and use multiple kinds of information.
- Development of novel algorithms and statistical measures for computing relationships among participants in a large dataset.
- Protein domain, nucleotide and amino acid sequences, protein structure analysis and interpretation.

This thesis covers first two subdisciplines of bioinformatics, developing novel algorithms and computer programs for image analysis using images represented as DNA by applying sequence alignment to find relationships among NWND, NWD, WND and WD images.

Bioinformatics span

The crux for analyzing, organizing and understanding biological data using computers leads to expand biological analysis by using bioinformatics methods in two dimensions: breadth and depth (Luscombe N, Greenbaum, & Gerstein, 2001). The vertical axis (i.e. depth) represents the rational drug design process that aims towards analyzing a single protein to maximize understanding about the protein it encodes. The analysis starts with the gene sequence; from there, we calculate the protein sequence. The prediction algorithm (Escobar, et al., 2016) determines the structure; geometrical calculations can interpret the protein surface shape. Force fields around the protein molecule were discovered by using simulation (Lopes, Guvench, & MacKerell, 2015; Freddolino, Park, Roux, & Schulten, 2009). Lastly, the docking algorithms (Sliwoski, Kothiwale, Meiler, & Lowe, 2014) will determine the ligands design for binding to the proteins that leads to the drug designing. Through this, the protein functions could be changed.

Breadth is the second dimension of bioinformatics that aims towards comparison of one gene with another for biological analysis. In other words, sequence analysis is performed. That indicates a DNA, RNA or peptide (protein) is subjected to a vast range of analytical processes to find out structure, features, functions or evolution. Search in the biological database through sequence alignment and other methods are included in sequence analysis (Durbin, Eddy, Krogh, & Mitchison, 1998). A pair of relevant proteins is compared to their sequences and structures by using simple algorithms (pairwise sequence alignment).

As the number of proteins is increased (i.e. from 3 to 100), the algorithms are further improved for performing multiple sequence alignments, and sequence patterns or structural templates are extracted that determine a group of proteins. We can use this data to generate phylogenetic trees that can be used to track the evolutionary journey of proteins. Lastly, with even more proteins or data (i.e. more than 100), large-scale databases are required for storage. More complexity arises for comparisons; multiple scoring schemes are needed by which we can perform genomic scale censuses that generate exhaustive statistical accounts of protein features. Like abundance of particular functions or structures in different genomes. A phylogenetic tree is also generated by this data that will track the evolutionary journey of the whole organism.

In this thesis, our research focus is on the second dimension of bioinformatics (i.e., the breadth) for developing novel methods in images analysis (i.e., watermarked and non-degraded (WND), non-watermarked and non-degraded (NWND), watermarked and degraded (WD), non-watermarked and degraded (NWD) by using MPS) in the domains of bioinformatics and image analysis by incorporating inspirations from sequence analysis.

1.2.2 Sequence Analysis

The primary rationale for utilizing sequence analysis is an inspiration for image analysis; the sequences have crucial information that determines the habits, personality and inheritance properties of species. For example, biological sequences are made up of C (Cytosine), A (Adenine), G (Guanine) and T (Thymine) nucleotide base pairs. The structure and position of these pairs define the habits, personality and inheritance properties of the species (Mathkour & Ahmad, 2009). The main focus of biologists is on distinguishing species by using functional properties to get optimal results. Extraction of meaningful information from massive repositories of sequence data gives us very compelling outputs related to functional characteristics of genes. This inspires this thesis to extract meaningful information from a biological sequence (DNA) of images (i.e. NWND, NWD, WND, and WD) with the help of sequence analysis.

Sequences may have a variety of natures and types that include multilanguage sequence, genetic sequence, RNA sequence, DNA sequence, protein sequence, etc. It is hard to study all kinds of sequences in short time; thus, the scope of this thesis deals with only DNA and protein sequences.

According to the scope, we explained the sequence hierarchy for genome sequence analysis. Genome sequence analysis is divided into two parts: sequence alignment and visualization. For understanding more diversity in the functional properties of species, sequence alignment can be used. In sequence alignment, DNA, RNA or protein sequences are arranged in a way to identify similar regions that result in structural, evolutionary and functional relationships between sequences. Further sequence alignment is divided into local alignment (Polyanovsky, Roytberg, & Tumanyan, 2011), global alignment (Needleman & Wunsch, 1970; Smith & Waterman, 1981), multiple alignments (Rani & Ramyachitra, 2016) and duplication. Information extracted from sequence alignment is utilized to track any possible evolutionary relationships among species, the degree of relevancy, diversity in the species and genetic relationship among the species.

The second part of this genome sequence analysis is sequence visualization. These methods are very powerful that will exploit human vision for analyzing and organizing the sequence data. It includes pattern recognition (Ridder, Ridder, & Reinders, 2013), motif concentration (Rampasek, Jimenez, Luptak, Vinar, & Brejova, 2016), structure prediction in two dimensions (2D) or in three dimensions (3D) (Soding, 2017), phylogenetic tree (Wilgenbusch, Huang, & Gallivan, 2017) and microarray (Kauffmann, Gentleman, & Huber, 2009). These are very useful in recognizing and visualizing duplicating patterns in genomes (Tao, Liu, Friedman, & Lussier, 2004). Pattern recognition transforms and classifies entities from the patterns extracted from accumulated raw data. Raw data accumulation has obscure patterns; so, by using pattern recognition, we can obtain a pattern which is more meaningful. Pattern recognition is divided into two parts: supervised pattern recognition (unrelated data to the measurement methods, like labels, are accessible) and unsupervised pattern recognition (labels are not present). Pattern recognition can be implemented in two fundamental ways: statistical (i.e. having statistical decision theory as a foundation) and syntactic or structural (i.e. having human perception and cognition as a foundation). Statistical pattern recognition cannot discriminate morphological patterns because of its quantitative nature that motivates us to use a syntactic or structural pattern recognition approach in this thesis for image analysis where pattern matching is checking specified patterns in a given token sequence (Soroushnia, Daneshtalab, Plosila, Pahikkala, & Liljeberg, 2014). In pattern recognition, matching sequences must be exact, but that is not compulsory in pattern matching. This proposed work uses both of them: pattern matching for analyzing and identifying the WD, WND, NWND and NWD images; and pattern recognition for grouping of the WD, WND, NWND and NWD images in their group by using a phylogenetic tree. This thesis examines sequence analysis to derive inspiration and also inspect sequence alignment (i.e. local, global & multiple alignments) and sequence visualization (i.e. pattern recognition and phylogenetic tree) to develop a novel bioinformatics-inspired image analysis for watermarked images after MPS.

1.2.3 Image Analysis

The main goal of this thesis is to apply bioinformatics concepts (sequence alignment and sequence visualization) in the domain of image analysis. One compelling question can be requested: Why should we use the bioinformatics concepts sequence alignment and sequence visualization as an inspiration for image analysis, when hundreds and hundreds of methods and algorithms are already present in the literature to solve the issue? No Free Lunch Theorem gives the answer (Wolpert & MacReady, 1996). This theorem explains that an algorithm cannot be found that, on average, will outperform compared to any other algorithm. Another instance is image segmentation algorithms that contemplate cell size and shape to perform segmentation on closely packed cells in tissues (Lin, et al., 2003; Dufour, et al., 2005). Cells are well separated, consistent intensity can be isolated by watershed algorithms, but this will not work for tightly packed cells in tissue (MacAulay & Palcic, 1988). Therefore, some algorithms work best for certain conditions but are not suitable for all environments with different conditions. It indicates that some trade-offs exist among all persisting algorithms in different conditions. None of the algorithms satisfies all the required conditions concurrently. Finally, we can argue that there is always a requirement for novel algorithms that can examine the data in a new style. In this thesis, we will attempt to generate novel image analysis methods (pattern matching and pattern recognition methods) using inspiration from bioinformatics sequence alignment and sequence visualization. The following topics cover image analysis and watermarked images as well as multiple print and scan.

Image Analysis

Digital image analysis (Michler, 2008) is a conversion process where the input image is altered to get an output, i.e., it gives some information presenting an explanation or judgement where digital image processing can be considered as revision of an image into another image, i.e., the input image is processed to get a transformed (Michler, 2008) image or image attributes as output.

Image analysis deals with mining of significant data from images, predominantly from digital images by applying digital image processing methods. The ultimate goal of digital image processing that enhances the quality of the work to identify patterns and objects, is to extract helpful data from images, improve and reconstruct the images. For achieving this worthwhile aim, tools and methods have been developed for efficient processing of images. Algorithms are a key part of the toolset for the development process. In spite of that, images are very complicated with probably large databases, resulting in a very time-consuming job for image processing. Many other challenging problems also need to be explored. There are many applications of image analysis in industry and all fields of science: in astronomy for determining planet size (Kremer, Stensbo-Smidt, Gieseke, Pedersen, & Igel, 2017) in defence, remote sensing (Wilschut, et al., 2013), medicine (Toennies, 2017) etc. The scope of this thesis will cover only three aspects of image analysis: evaluation of degraded images after MPS, identification of images from a group of images, grouping images having a particular set of properties. In other words, the proposed BIIA for image analysis divides further into two parts: bioinformatics-inspired image identification approach (BIIIA) and bioinformatics-inspired image grouping approach (BIIGA), where sequence alignment used for DNA represented NW and W images.

Watermarked Images

Digital watermarking is described as the method of modifying a digital image to embed secure information digitally (Cox I. , Miller, Bloom, & Honsinger, 2002). It is based on steganography (Marvel, Boncelet, & Retter, 1999) or data hiding, the word steganography deriving from the Greek word "covered writing". A plethora of algorithms for the watermark (the generic term used for any image, text, audio, video and any other information having user secret or identification data) embedding and extraction have been developed and applied. Current digital watermarking schemes to preserve the authenticity and integrity of digital data (photos, videos, documents) have focused on problems of preserving watermarks in the context of degradation, compression and decompression techniques. In this thesis, watermarked images refer to the ones that have a digital watermark in the form of a shape, text or other image; non-watermarked images (NW) contain no watermarks. A proposed evaluation, identification and grouping approach was tested on both original versions of watermarked / non-watermarked images, and a degraded version with a multiple print-and-scan processes.

Multiple Print and Scan

Image degradation means that visual information on the source image is diminished (Ye & Doermann, 2013). The research work in this thesis investigates image degradation from a digital viewpoint into consideration. Multiple print-scan operations are iteratively committed to a hardcopy or a softcopy from the same source, i.e., taking a hardcopy of the original document or photo, the document is scanned and converted to a softcopy; then, the softcopy is printed, the printout is treated as the printed hardcopy. The operation is iteratively in use of the same photocopy machine for many rounds. There are a variety of reasons for image degradation in real image print and scan. As a single round of print and scan, the degradation is generated as a result of physical characteristics of the print and scan, such as scanner defocusing and image binarization (e.g. fixed and adaptive threshold), poor flaking and poor toner adhesion of ink to papers, low print contrast, etc. (Baird & Chaudhuri, 2007). Moreover, other factors like lossy image compression, not enough sampling, etc. also can contribute to the artifacts of image degradation. If the same print-and-scan process is repeated, then all degrading factors add more and more degradation to the images under test. As the number of print-and-scan rounds increases, the more degraded the image is. Therefore, it is critically significant to deeply investigate this problem for the sake of visual alleviation and understanding. To the best of our knowledge, this is the first time such an evaluation, identification and grouping algorithm for degradations of non-watermarked / watermarked image from MPS has been taken into account.

Elementary research work on multiple print-and-scan (MPS) and degraded images focused on developing a robust watermarking approach against degradation. The less addressed effort is a bioinformatics-inspired approach to identify watermarked / nn-watermarked images or degraded copies by using MPS. The aim of the thesis is to investigate DNA biosequences for biological image representation and Smith-Waterman algorithm (SWA) (Smith & Waterman, 1981) and Needleman-Wunsch algorithm (NWA) (Needleman & Wunsch, 1970). These algorithms were used for sequence alignment to extract common substrings for detecting patterns or conserved regions, and develop an automatic signature extraction method for identifying degraded and watermarked / non-watermarked images by using MPS or original variants of watermarked / non-watermarked images. Experimental results reveal the feasibility of a proposed method of identification of original or degraded variants of watermarked / non-watermarked images by using common substrings acquired from DNA image representation, alignment by using SWA and NWA. In this thesis, the contribution is in the domain of bioinformatics and image analysis, i.e., a novel evaluation, identification (BIIIA) and grouping method (BIIGA) developed based on bioinformatics sequence alignment, biological image representation and pattern matching that examines and identifies degraded after MPS and original copies of watermarked / non-watermarked images. This approach may overturn our understanding of identification of degraded / non-degraded and watermarked / non-watermarked images and may lead to a new era of syntactic-based watermarked / non-watermarked and degraded / original images for developing the next generation of identification of those degraded and watermarked / non-watermarked images.

1.3 Thesis Structure

Acknowledging the fundamental inspiration and motivation behind this research work, the primary aim of this thesis is to go no less than one level further past BIIA methods (i.e., BIIIA & BIIGA) into biologically-represented images that support sequence alignment and visualisation to inform the proposed BIIA methods (i.e., BIIIA & BIIGA) with data and approaches found at this deeper level. In the last section of this thesis, we will assess whether going further has given any advantage. The subsequent subsections will explain the contribution and outline of this thesis.

1.3.1 Contribution of Thesis

The contribution of this thesis is to apply the existing knowledge in bioinformatics; the image analysis is explained as follows:

- Comprehensive literature review of the existing BIIIA and many gaps in current image analysis methods are outlined in Chapter 2.
- A detailed analysis of BIIIA method is explained in Chapter 3 based on the latest bioinformatics knowledge.
- Evaluation of image degradation by using MPS along with eight metrics, i.e., CC, Bias, ERGAS, RMSE, RASE, Q, SSIM, and DSSIM is presented in Chapter 4.
- Successfully detection of watermarked / non-watermarked images (original or degraded variants from MPS) by using biologically-based representations of the image, sequence alignment algorithm and pattern matching in biometrics is proposed in Chapter 5.
- For image identification, SWA is better than NWA for biologically-represented images, and suitability to DNA representation is presented in Chapter 5.
- Successful use of the phylogenetic tree for grouping original / degraded copies of the watermarked / non-watermarked images is proposed in Chapter 6.

1.3.2 Organisation of Thesis

The structure of the thesis is as follows:

Chapter 2 includes an introduction to bioinformatics and image analysis. As we know, bioinformatics and image analysis are very broad areas; so, we only focus on bioinformatics concepts related to the scope of this thesis. A comprehensive literature review in image analysis is also presented with special attention on BIIA using image evaluation metrics, biological image representation with bioinformatics sequence alignment and visualization. This chapter concludes with outlining gaps in the literature, providing a base for proposed novel BIIA methods.

Chapter 3, on the basis of the previous chapters, outlines the literature gaps. This chapter presents the research methodology used with open questions that are addressed in this thesis and the proposed three methods. First for evaluation of image degradation by using MPS for watermarked / non-watermarked images. Second, two novel BIIA methods, which we call as bioinformatics-inspired image identification approach (BIIIA) and bioinformatics-inspired image grouping approach (BIIGA).

Chapter 4 proposes a novel model for evaluating degraded and non-watermarked / watermarked images after MPS by using different metrics. This chapter uses various metrics with several degraded images by using MPS. It attempts to develop a model for analyzing watermarked / non-watermarked and degraded images after MPS. The rest of this thesis will deal with discriminating (i.e., for identification and grouping) non-watermarked and degraded / non-degraded images, watermarked and degraded / non-degraded images, watermarked / non-watermarked images based on conserved region identification by using DNA biosequences.

Chapter 5 proposes a bioinformatics-inspired image identification system (BIIIA) algorithm to identify non-watermarked / watermarked and degraded images by using MPS, which were evaluated for MPS degradations in the previous chapter. This chapter explains the idea of conserved region application in the discrimination of images. The experiments are conducted and analyzed on standard images in digital image processing.

Chapter 6 attempts to develop a bioinformatics-inspired image grouping approach (BIIGA) algorithm for grouping non-watermarked / watermarked and degraded images after MPS, which are identified after MPS degradations in the previous chapter.

Chapter 7 concludes that this thesis with a summary of future work.

1.4 Summary

This chapter introduced the motivation and scope of bioinformatics in image analysis. Bioinformatics in depth and breadth was discussed for knowledge discovery and sequence alignment. Sequence analysis regarding sequence alignment and sequence visualization was discussed in detail. The contribution of this thesis is in the bioinformatics and image analysis by using sequence alignment of biologically-represented images in DNA to group and identify watermarked / non-watermarked and non-degraded / degraded images from MPS.

Chapter 2

Literature Survey

This chapter inspects bioinformatics tools for image analysis. Section 2.1 includes - the image analysis and its major methodologies. In Section 2.2, *bioinformatics-based image analysis is reviewed, including biological image encoding, sequence alignment, phylogenetic trees. The chapter continues with a background of the image watermarking system and MPS degradation evaluations on details in Section* 2.3 *and Section* 2.4, *respectively. In Section* 2.5, *emergent research problems for bioinformatics image analysis are addressed. Finally, we summarise the literature survey chapter.*

Bioinformatics-inspired image analysis (BIIA) has grown relatively recently in contrast to other nature-motivated computing methods, for example, particle swarm optimization, genetic algorithms, etc. The main focus of this chapter is on BIIA and the motivations it supplies for generating BIIIA and BIIGA methods in the domain of image analysis. Keeping in mind the end goal of framing the reason for building up a novel BIIIA and BIIGA calculation, it is important to give a foundation of BIIA. Therefore, this chapter explains standards, functionalities and main theories related to the development of BIIA in recent years.

2.1 Image Analysis

Digital images are used extensively in every field of science (like mining, cells, genes mapping, etc.) and social media (like Facebook, LinkedIn, Instagram, etc.). This motivates us to develop new methods in image analysis to meet the current demand for better analysis. Image analysis can be used from sewer pipe deformation evaluation (Kun, Luxmoore, & Davies, 1998) to astronomical and archaeological image multispectral image analysis (Roberto & Hofer, 2009). It is a field of science that studies image details.

2.1.1 Image Analysis

Three different categories are assigned to different types of image analysis tasks: low level, medium level and high level (Pour, 2015). For low-level image analysis methods, input and output are both digital images, for instance, processes like noise reduction and contrast enhancement, whereas an image acting as input and output is some kind of information extracted from the image, i.e., the output is not an image for the medium-level image analysis. That includes object detection like face detection and image segmentation. High-level image analysis is most difficult where the input is an image, but the output is “knowledge.” For example, a person’s image may act as an input and the output will determine whether the person is sad or happy. The focus of our thesis is to develop a novel high-level image analysis by using bioinformatics tools and techniques that applied to the input images so as to get the knowledge about the identification and grouping of W/NW and WD / NWD images.

2.1.2 Major Methodologies in Image Analysis

It’s hard to explain large number of methods used in the image analysis. However, an attempt is performed to discuss the major areas of image analysis that include shape analysis, matching, segmentation and description (Mantas, 1987). More details about these methods are described below.

- *Shape analysis.* It is comprises of a group of methods like spatial techniques, transform techniques and global shape analysis, etc.
- *Matching:* It denotes the class of operations that include comparing of pixels with each other. A subclass of matching is to detect the alterations or motion (i.e. motion detection) in a scene by supplying images at different times. This method is known as time-varying imagery (Nagel, 1978; Nagel, 1983). Other category methods are applied to remote image sensing systems.
- *Segmentation:* Pixel clustering and classification processes for different regions of an image is referred as segmentation (Peng, Zhang, & Zhang, 2013). It is an intermediate process in image analysis where from background, the object of interest is extracted to identify which part of the data array makes up an object in the real world. Depending on the specific needs, different tasks are supported by using segmentation like measurement, visualisation, registration, reconstruction and content-based search. For instance, higher accuracy is required in measurement than in visualisation, while for large datasets, efficiency is more important for searching than for surgery simulation and planning (Olabarriaga & Smeulders, 2001). Different approaches are used for segmentation, the most common being edge detection, thresholding, and advanced segmentation techniques (Gayathri & Raajan, 2016).
- *Description:* Image analysis algorithms are used to represent or describe the test image into its main features. These features are represented by a parsable string of numbers or characters. This parsable string acts as input for suitable recognisers which can be either structural (syntactic) or statistical.

In this thesis, image analysis processes like matching, segmentation and shape analysis are not discussed further. This thesis aims to explore the last method (i.e. description) where the image is represented as a parsable string of numbers or characters without extracting the features from the test image. This indicates that the whole test image is encoded in a DNA string of characters inspired from bioinformatics, then sequence alignment is used to extract the common substring; in the end, this common substring is put in a syntactic recogniser. How bioinformatics has been using image analysis tools and techniques are discussed in the next section so as to build a theoretical foundation for BIIA.

2.2 Bioinformatics Image Analysis

In bioengineering, the fastest and furiously developing field is bioinformatics image processing (Zhou Z. , 2016). It is a branch of the bioengineering field that mainly focuses on biological image processing and analysis as well as biological information analysis. The main focus of bioinformatics image analysis is to excavate digital information from biological image sequences or biological images. In literature, how the bioinformatics image analyst addresses bioinformatics concepts like sequence alignment and phylogenetic trees for images other than bio images, is explained in the next subsections.

2.2.1 Existing Techniques of Bioinformatics Image Analysis

In bioinformatics, there are different technologies and tools available for biological image and other image analysis. Descriptions of all the tools are beyond the scope of this thesis, only the categories of these tools covering research work on bioinformatics image analysis are introduced in this thesis. Eliceri et al. divided image analysis tools into two categories, i.e., generalist image analysis tools and niche image analysis tools on the basis of the problems addressed by the tools under development (Eliceiri, et al., 2012). A detailed explanation about these methods is given below.

- *Generalist image analysis tools:* These tools for image analysis focus on more general issues, for instance, contrast enhancement, cropping, image segmentation, etc. Normally, these tools are modular, leading to more flexibility for many applications, for example, Image Pro Plus from Media Cybernetics, Imaris by Bitplane Scientific Software*,* etc. Some open source tools are developed for specific problems, but after some time, they are extended with other functionalities for other purposes. These tools include Cell Profiler (Carpenter, et al., 2006), BioImageXD (Kankaanpaa, et al., 2012), Icy (Chaumont, et al., 2012), Fiji (Schindelin, et al., 2012), IMOD (Kremer, Mastronarde, & McIntosh, 1996) , etc.
- *Niche image-analysis tool:* In academia, most tools for image analysis are developed for resolving very specific tasks. These tools are designed for particular types of cells (particularly neurons), imaging modalities, organisms,

etc. Among the many tools, some are still in use, and some are outdated (Eliceiri, et al., 2012), for instance, an image analysis tool for neurosciences.

These tools, designed to perform image analysis, are most relevant to requirements and needs. For instance, Icy is an excellent choice for cell segmentation, behavioural analysis, and cell tracking; Fiji provides a unique process in the analysis of electron microscopy data; meanwhile, BioimageXD is the tool of choice for 3D visualisation.

In the perspective of open source platforms, for image analysis tools, ImageJ (initially called as NIH) has a remarkable position. The notable place is due to its free availability, having extensive multipurpose image analysis capability and having been used for a long time (Abramoff, Magalhaes, & Ram, 2004; Collins, 2007; Schneider, Rasband, & Eliceiri, 2012). Another reason for its success is that researchers can extend ImageJ foundations while developing the algorithm specific to the research. This resilience helps to make ImageJ a pioneer in image analysis and popular among both users and developers. It is a Java-based tool for general purposes. For the next generation of multidimensional image data analysis, ImageJ2 is under development (Schindelin, Rueden, Hiner, & Eliceiri, 2015) (Rueden, et al., 2017).

All these tools in one way or other way use segmentation, shape analysis, matching or description methodology in their process of image analysis. In this thesis, niche tools for image analysis will be proposed that focus on the description methodology of image analysis. In other words, specific functionalities like classification and identification are developed using bioinformatics tools and encoding images into biological DNA. That raises many questions. Is DNA as a substrate used for image analysis or other purposes? If yes, then how it is used? These questions motivate us to study further about DNA for representing different media, including biological image representation using DNA, which is discussed in the next subsection.

2.2.2 Biologically-Based Image Representation

Images could be represented cognitively, as “pixel-region-object-scene” hierarchies and having relationships among them (Xie, Gao, Wu, & Zhang, 2011). Semantically, images could be represented by using and-or graphs in four levels, known as “scene-object-parts-primitives” (Xie, Gao, Wu, & Zhang, 2011). The above two ways of image representation are used to analyse images for a particular set of applications. Another interesting way to represent an image is to encode it in biological DNA. DNA has a long history of being used it as a substrate, in different areas like Hamiltonian paths (Adleman, 1994), Boolean circuits (Takahashi, Yaegashi, Kameda, & Hagiya, 2005), neural networks (Qian, Winfree, & Bruck, 2011), chemical reaction networks (Chen, et al., 2013), steganography (Clelland, Risca, & Bancroft, 1999), cryptography (Leier, Richter, Banzhaf, & Rauhe, 2000) and watermarking (Gibson, et al., 2010). In the literature, the biological image representation is used for storage and retrieving images from the dataset. This is performed by using the four basic blocks of DNA (Deoxyribonucleic acid) sequences – adenine (A), thymine (T), guanine (G), and cytosine (C) (Bornholt, et al., 2016). The main motivation behind images represented in DNA is to revolutionise computer storage. The basic idea of storage and retrieval of digital data (image or text) as DNA, is firstly to store images or data (i.e. 0 and1 bits) as DNA, then encode the DNA sequences like street addresses and zip codes for easy retrieval and then use sequencing techniques to read and transform it to the original arrangements utilizing street addresses to arrange the data in original sequences (Bornholt, et al., 2016).

How to convert ones and zeros to four basic blocks of DNA i.e. A, C, T and G, is crucial. If the approach focuses on making it very dense, then the retrieval error rate is very much less (Bornholt, et al., 2016). A decent arrangement of work has been done to store the digital data (i.e. from images, text etc.) on the DNA, i.e. converting digital information to DNA (Davis, 1996; Cox J. P., 2001; Wong, Wong, & Foote, 2003; Church, Gao, & Kosuri, 2012). An error correction code was used by Goldman et al. for developing a new approach (by four-fold redundancy) using DNA data storage, but redundancy raised the length of DNA and made the techniques expensive (Goldman, et al., 2013). Low-density Parity-Check codes (LDPC) were introduced in 2014, by Aldrin Kay-Yuen Yim et al. to encode data in big blocks of DNA (Yim, et al., 2014). Other works, DNA synthesis (write) and sequencing (read and access) are focused on improvement in the cost of DNA synthesis and error of free storage as well as retrieval of digital data. These improvements include Reed-Solomon codes-based DNA storage (Grass, Heckel, Puddu, Paunescu, & Stark, 2015), DNA channels for data storage (Kiah, Puleo, & Milenkovic, 2015), rewritable random access DNA-based storage (Yazdi, Yuan, Ma, Zhao, & Milenkovic, 2015), error correction scheme (Limbachiya, Dhameliya, Khakhar, & Gupta, 2015) and altering the Goldman process (Goldman, et al., 2013). All the above approaches focus on successful error-free encoding and decoding of digital images, text, video, etc. for storage and retrieval of digital data.

The encoding scheme used in this research work for converting images to DNA, is inspired by Naidu and Narayanan's approach (used for biological representation of computer viruses) (Naidu & Narayanan, 2016) by adding the required steps because, for pattern matching a hexadecimal code is a must in our case (i.e. Naidu and Narayanan's approach is followed) that is achievable by adding a few steps. Secondly, our focus is on grouping and identification of degraded / non-degraded images, i.e., if some data is lost during the process of conversion of an image to DNA and DNA to hex, the proposed approach is still able to identify and group the degraded images. That automatically improves the robustness of the method for identification and grouping of degraded images. To be on the safe side, we checked first by converting the image into DNA by the proposed approach and retrieved successfully with a little bit of degradation. None of the above techniques addressed the issue of *how the DNA representing digital images can be used for identification and grouping of images*. This thesis addresses this issue by including required steps for biology-based encoding of the image by using sequence alignment between the two DNAs of biologically-encoded images, which is explained in detail in the next subsection.

2.2.3 Biological Sequence Alignment

Sequences used in this thesis are defined as a string of residues of Deoxyribonucleic acid (DNA). It is explained below (Kaur & Chand, 2016):

> *Deoxyribonucleic acid (DNA):* DNA sequences are made up of nucleotides having genetic information. Residues: Adenine (A), Cytosine (C), Guanine (G) and Thymine (T)

Biological sequence alignment is the cornerstone of bioinformatics for establishing functional, evolutionary or structural relationships between DNA sequences. Sequence alignment will achieve this by identifying similar regions between two sequences of DNA by using different algorithms (Kaur & Chand, 2016). The motivation behind using sequence alignment is described below:

a. *Evolution determination.* To discover, the part susceptible to mutation or the part that preserves structures of the given two sequences. This allows us to define the evolution process by predicting future gene mutations and the probable origin of diseases.

b. *Structural similarity recovering.* In proteins, sometimes some sequence parts are missing that lead to a particular disease. If we find that and develop a similar sequence, then diseases related to the missing part will be cured. Generally, this technique is used in medicine design.

c. *Pattern detection.* For grouping and identification of sequences into classes and families, biologists use pattern detection. That is, they extract patterns in the form of subsequences, which can be one or more for a particular class. Then, they compare subsequences (patterns) to other sequences to find the same subsequence. If a similar one is found, then that entity belongs to that class.

d. *Comparing a new sequence versus a dataset.* If the biologists discover a new sequence for finding out that sequence function and other properties, it is compared with the existing datasets of sequences. If a match is found, then it expresses that they have the same structure, and the new sequence has the same function and properties, i.e., similar to the sequence features existing in the dataset.

The pattern detection inspires the proposed thesis by comparing a new sequence versus a database from above motivation of sequence alignment. From now onwards, in this thesis, we will discuss only (c) and (d) points. Pattern detection is performed in two ways: pattern matching (for image identification i.e. BIIIA) and pattern recognition (for grouping of images i.e. BIIGA), where patterns are extracted by using sequence alignment of two DNA or protein sequences of images called as signatures (common subsequences) these signatures are tested by using (d), i.e., comparing a new sequence versus a database. Here, new sequences are denoted as signatures and databases are different image datasets that are created for this research work. By using the pattern matching technique, these signatures are matched with different image datasets to identify whether the images belonging to this signature exist or not, i.e., BIIIA. For pattern recognition, a new sequence represents multiple aligned sequences. These multiple aligned sequences are used to group all images to the respective category, i.e., by using BIIGA.

Sequence alignment algorithms in biology are divided into two parts by using the number of sequences in the alignment; pairwise sequence alignment (PSA) will use two sequences, multiple sequence alignments (MSA) will utilise three or more sequences (Chen & Wang, 2016). During sequence alignment, if both sequences positions have the same letter, then this position has been conserved during evolution. If they have a different letter, then the sequence position is not conserved, the gaps (represented as '–') are inserted wisely in the sequences to maximise the match. Initially, a gap penalty score is given, which equals to the regular gaps entered, but keeps the sequence together, an affine gap penalty is used as an alternative for inserting a large number of gaps. For continuous gaps, Gap Open (headmost gap) is allocated with a regular gap penalty score, and Gap Extended (posterior gaps) is allotted with a lower penalty score (Kaur & Chand, 2016). More details about PSA and MSA are given below.

- *Pairwise sequence alignment (PSA):* Pairwise sequence alignment can be implemented in two ways: global (i.e., from starting to the end of sequence alignment) and local (i.e., detecting local regions of high similarity) alignment (Chowdhury & Garai, 2014). Historically, the most prominent global algorithm for sequence alignment is the Needleman-Wunch algorithm (NWA) (Needleman & Wunsch, 1970); a typical pairwise local alignment is shown using Smith-Waterman algorithm (SWA) (Smith & Waterman, 1981) which has been enormously used since the introduction in the 1970s/1980s. SWA, originally developed for finding common molecular subsequences, has been applied for object classification (Roth & Ommer, 2006), spatial activity recognition (Riedel, Venkatesh, & Liu, 2006), partial shape matching (Chen, Feris, & Turk, 2008), malware identification (Naidu & Narayanan, 2016) and more. Similarly, NWA has been used for malware sequence alignment (Dinh, Brill, Li, & He, 2016) and other processes. Processing speed issues of these algorithms was solved by Gotoh, who proposed (Gotoh, 1982) an enhanced version of SWA and NWA, Dynamic Programming (DP) in SWA and NWA which is exhaustive in nature (Chowdhury & Garai, 2014). Mathematically, it was proven that DP gives optimal alignment for pairwise sequence alignment (Shyu, Sheneman, & Foster, 2004). A substitution matrix is used to represent the match / mismatch score value to determine the extent to which two sequences are aligned. The most parsimonious method is the identity (ID) scoring matrix because no assumptions is made for finding a connection between one character and another in the string. Because of that benefit, the ID matrix will be used for our experiments rather than popular biological scoring matrices, such as BLOSUM (Block Substitution Matrix) and PAM (Point Accepted Mutation) (Cohen, 2004).

Homola et al. had used sequence alignment for retrieving images from datasets that hold a query image (Homola, Dohnal, & Zezula, 2011). Hung-Sik et al. (Kim, Chang, Lee, & Lee, 2010; Kim, Chang, Liu, Lee, & Lee, 2009) had proposed an initial work that utilises sequence alignment for image matching in 2009 and 2010, respectively. They extracted, image spatial relations, colour and textural features; then, these features are converted to DNA or protein sequences. Image matching was done by performing sequence alignment between two image sequences by using Basic Local Alignment Search Tool (BLAST) (Lobo, 2008; Altschul, Warren, Miller, Myers, & Lipman, 1990) based on SWA.

The approach for finding similarity in images has two limitations. The first one is that we had to map image features to the 4 DNA or 23 protein alphabets. They developed a “Composite Conversion Table,” which is not easy to generate a *“universal mapping for various features.”* The second limitation is with their substitution matrix, which decides the similarity (when two alphabets are similar) and penalty (when one is replaced by the other). A simple uniform matrix was used, where 1 is in the diagonal and -1 is for all other elements. That indicates the letters are only similar to themselves. Gaussian distributed matrix attempts to convey similarities between mapped features, for instance, yellow is more similar to gold rather than blue. These two limitations were improved by Pawel et al. (Drozda, Gorecki, Sopyla, & Artiemjew, 2013) by using Bag of Visual Words (BoVW) in 2013. They proposed that BoVW was more appropriate to the sequence alignment framework and NWA for sequence alignment. Firstly, their approach is not limited to particular alphabet lengths. Secondly, it is easier to develop a substitution matrix by calculating the distance between each pair of vocabulary centroids (Drozda, Gorecki, Sopyla, & Artiemjew, 2013; Drozda, Sopyla, & Gorecki, 2014).

NWA and SWA, with a reduced gap penalty, were used for 2D shape recognition by Bicego and Lovato (Bicego & Lovato, 2012) in 2012. They represented a shape with eight directional chain codes and then mapped each chain code into eight amino acids in a one-to-one manner: D, C, Q, E, G, A, R and N using Matlab to ensure no information loss. Then, NWA was used for alignment to get the alignment score and the nearest neighbour classifier for classification. They could enhance the results in many ways, i.e., by increasing the number of aminos used for mapping, by specifying shape-scored matrices and so on. Further, improvement in this work was made by Lovato and Bicego using BLAST for sequence alignment in 2012, with three different sets of experiments (Lovato & Bicego, 2012). The first experiment was with the BLAST-default setting, for sequence alignment by removing filters, which is used for eliminating areas of low complexity (like same symbol repetitions). Of course, they are informative in biology as well as in shapes which should not be eliminated. The second experiment was with a BLAST-reduced gap penalty, i.e., they changed the default gap opening penalty from 11 to 6 and the gap extending penalty from 1 to 2. They changed this parameter because biological assumptions that do not hold for 2D shape classifications must be relaxed. In biology, a big gap penalty is usually used but that does not hold for 2D shapes. These changes are supported by the improved results for classifications of chicken and vehicle datasets, i.e., from 78.92% to 82.06% and 82.02% to 84.37%, respectively. Finally, the last experiment was done on a substitution matrix. They chose a substitution matrix that penalises highly if there is a change in the sequence. BLOSUM90 matrix was used, instead of default BLOSUM62 (block substitution matrix 62) (Henikoff & Henikoff, 1992) (the greater the number at the end of word BLOSUM, the more conservative the substitution matrix is) specifically by compelling algorithms for best alignment. In the case of 2D shapes, an exact matching can be favoured, whereas, in biology, somehow equivalent amino acids are likely to be exchanged. Results for classification of chicken and vehicle datasets improved from 78.92% to 83.41% and 82.02% to 85.42% respectively. In 2014, a novel substitution matrix, Shape-BLOSUM (S-BLOSUM), was introduced for 2D shapes classification (Lovato, Milanese, Centomo, Giorgetti, & Bicego, 2014) where S-BLOSUM was designed to have prior knowledge about conserved regions of 2D shapes by learning the match / mismatch rate of conserved regions of shapes affiliated to the same class. This matrix has been incorporated into the selected representation of the 2D shape. Further experiments were done in 2016. For deeper analysis of their approach, previous BLAST and BLOSUM configurations used by Lovato and Bicego in 2012 were combined with three different types of coding (i.e. ‘single’ using amino acid, ‘triplet frequency’ using DNA and ‘triplet distance’ using DNA) (Bicego & Lovato, 2016). The results were promising and better than previously. Further analysis was fullfilled by using MSA, but unsatisfactory results were obtained in the retrieval case.

Another promising and exciting work for reversing the interaction between bioinformatics tools and pattern recognition is 3D shape matching by using bioinformatics (Bicego, Danese, Melzi, & Castellani, 2015). SWA and NWA with two encoding schemes, i.e., using DNA and amino acids, was used for 3D shape matching (Bicego, Danese, Melzi, & Castellani, 2015). In the approach, the Fiedler vector was used for extracting authentic mesh vertices, local geometric features were collected by using a shape index at each vertex. Then, these local geometric features were mapped in DNA or amino acid sequences. After that, NWA and SWA were applied to get the alignment and similarity measures which were used in the nearest neighbour classification scenario. The overall results were very promising.

Naidu and Narayanan developed an approach for identification of polymorphic malware variants (Naidu & Narayanan, 2016; Naidu & Narayanan, 2014; Naidu & Narayanan, 2016) by using biological representation and bioinformatics sequence alignment for malware. Their approach did not address the issue “how can we convert an image to biological representation, i.e., DNA”, without that, subsequent analysis cannot be done. We added steps required for W / NW image identification, after that, we followed the same approach as Naidu and Narayanan.

- Multiple sequence alignment has the same purpose as PSA; the only difference is that it is performed for three or more than three sequences with different algorithms (Bacon & Anderson, 1986). Different heuristic approaches have been developed for solving NP-complete optimisation problems of MSA. In the literature, three main categories for MSA are found: consistency-based methods, progressive methods, and iterative refinement methods (Rubio-Largo, Vega-Rodríguez, & Gonzalez-Alvarez, 2015). Details about these categories explained below.

  a. *Consistency-based.* Consistency-based group will develop local and global alignment datasets between every set of sequences that are used for precise MSA among all provided sequences. Some consistency-based tools are: PROBabilistic CONSistency-based multiple sequence alignment (ProbCons) (Do, Mahabhashyam, Brudno, & Batzoglou, 2005), Tree-based and Consistency

Objective Function For alignment Evaluation (T-Coffee) (Notredame, Higgins, & Heringa, 2000).

b. *Progressive methods.* These methods will calculate distance matrices from each set of provided sequences; then, any hierarchical clustering algorithm (like Unweighted Pair Group Method with Arithmetic Mean (UPGMA)) is used to build a guide tree which will provide us the alignment. A major drawback of this method is the incorrect gap in the starting that will pass on to final alignment. The most important progressive methods are Clustal omega (Sievers, et al., 2011), Kalign (Lassmann, Frings, & Sonnhammer, 2009), ClustalW (Thompson, Higgins, & Gibson, 1994) and WebPRANK (Loytynoja & Goldman, 2010).

c. *Iterative refinement method.* The most representative and iterative refinement methods are: Multiple Alignment using Fast Fourier Transform (MAFT) (Katoh & Standley, 2013) and MUltiple Sequence Comparison by Log-Expectation (MUSCLE) (Edgar, 2004). Iterative refinement tools start with a progressive alignment; then, they iterate to correct inaccurate gaps, possibly inserted during the progressive development phases.

In this thesis, we are exploring less and investigated ways of image analysis, which consist of applying sequence alignment, i.e., PSA (global and local) and MSA for pattern detection (i.e. pattern recognition and pattern matching).

For deeper analysis and further understanding of image analysis, establishing evolutionary relationships among images, we created phylogenetic trees for W / NW degraded / non-degraded images. This is described in the next section.

2.2.4 Phylogenetic Tree for Sequence Visualisation

Evolution happens with alteration in organism genes from one generation to another that creates the relationships between organisms by unbroken genetic lines. Phylogenetics seek to determine these genetic relationships. Darwin created the first tree of life (i.e. phylogenetic tree) found in his notebook *(Darwin, 1859)*. That indicates the importance of establishing relationships among organisms in the past, present, and future existed from long ago.

Phylogeny answers the question: “how did genetically connected sets of organisms evolved with time? In another word, it tells us relationships between collections of things (genes, organs, proteins, etc.) that advanced from common forefathers. Phylogenetic trees are used to represent the evolutionary connection among biological species and other entities. Before the origin of sequencing technologies for DNA, they were used for describing connections among species in taxonomy (used for description, classification, and naming of species) and systematics (used to classify species) *(Yang & Rannala, 2012)*. Nowadays, phylogenies are used extensively in nearly every section of biology. Furthermore, phylogeny is used not only for explaining the relationships among species, but also for gene family inconsistency *(Maser, et al., 2001),* population histories *(Edwards, 2009),* language evolution *(Gray, Drummond, & Greenhill, 2009),* etc. As of the latter, molecular phylogenetics has turned into a key instrument for genome differentiations, for example, to decode ancient and modern genomes *(Heng & Durbin, 2011),* to restore ancestral genomes *(Jian, 2011)* and gene frequency data in 1967 *(Cavalli-Sforza & Edwards, 1967),* etc. This “tree thinking” is used for various phylogenetic structures and media types (text, audio, videos, language, etc.).

Multimedia phylogeny will develop evolutionary structures to find the history of alterations for a group of digital entities. In this motive, grouping is useful to cluster entities originated from a similar source while placing irrelevant entities in different groups. Multimedia phylogeny is used to develop phylogenetic trees of images, videos and audios. For instance, Image Phylogeny Trees (IPTs) (Dias, Rocha, & Goldenstein, 2012; Dias, Rocha, & Goldenstein, 2010; Dias, Goldenstein, & Rocha, 2013), audio phylogeny (Nucci, Tagliasacchi, & Tubaro, 2013), video phylogeny (Dias, Rocha, & Goldenstein, 2011), image phylogeny forests (Dias, Goldenstein, & Rocha, 2013; Costa, Oikawa, Dias, Goldenstein, & Rocha, 2014), large scale scenarios (Dias, Goldenstein, & Rocha, 2013) and multiple parenting relationships (Oliveira, et al., 2014). Image phylogeny explains how we can trace parent-child relationships among near duplicate images. Duplicated images are transformed versions of an image that conserve its semantics. According to Alexis Joly, a document G1 is a near duplicate of a document G (i.e. original document), if G1= $t(G)$, $t \in \mathcal{T}$, where $\mathcal{T}$ is a group of tolerated transformations (Joly, Buisson, & Frelicot, 2007). A group of different transformations can be applied to a document as well. In that case, $\mathcal{T}$ is a combination of $t1 \in \mathcal{T}$, $t2 \in \mathcal{T}$, $t3 \in \mathcal{T}$. The resulting document $G3 = t3 \circ t2 \circ t1(G)$, $t1 \in \mathcal{T}$, $t2 \in \mathcal{T}$, $t3 \in \mathcal{T}$ will be a near duplicate (Joly, Buisson, & Frelicot, 2007). For example, when uploading, an image passes through transformations (e.g., by applying resize, colour correction, etc.). These images are called as near duplicates. The formal definition given by Alexis Joly is not accompanied in video and image near duplicate literature (Dias, Goldenstein, & Rocha, 2013). In our case, we use the term "near duplicate images" for images obtained by MPS transformation. MPS is applied to watermarked(W) / non-watermarked (NW) images, watermarked and non-degraded (WND) images, non-watermarked and non-degraded (NWND) images to get near duplicate images, called non-watermarked and degraded (NWD), watermarked and degraded (WD) images.

In this thesis, the main goal is to remodel the image phylogeny tree, considering the images degraded from MPS in different scanning modes (i.e. grayscale, color, black and white) by using bioinformatics concepts, i.e., MSA and tools. The core idea behind this approach is that images can mutate as living beings (animals, plants, etc.) evolving in biology. All the image phylogeny approaches are either based on the idea of dimensionality reduction, manifold and spectral clustering, viewpoint localisation, heuristic-based solution-oriented Kruskal algorithms, optimum branching or automatic optimum branching, etc. None of the above approaches deals with bioinformatics like MSA to develop an image phylogenetic tree. That inspires us to extend the work of the image phylogenetic tree by using bioinformatics. Understanding of these tools and techniques of bioinformatics employed in this thesis is explained in the next section.

There are three different thoughts to develop the phylogenetic tree. The first is the evolutionary approach for traditional phylogenetic tree development. After the advent of molecular data, the other two are evolved, i.e., phenetic and cladistic approaches. These thoughts are explained below:

- *Evolutionary systematics*. This is a goal to find relationships among organisms according to how natural selection made them most of the time, only classifications with little attempt show relationships as trees (phylogenies), they depend on the experts.

- *Phenetics*. Introduced in 1957 (Michener & Sokal, 1957) by Michener and Sokal. It defines relationships among a group of organisms by the grouping and classifying organism. The similarity may be phenotypic, anatomical, or molecular. Phenogram expresses the phenetic relationships in the form of a tree-like network. In other words, if a phylogenetic tree is drawn by grouping and classifying species, this is referred to a phenetic approach. For example, a maximum likelihood approach, etc.
- *Cladistics.* This is the study of the pathways of evolution (Hennig, 1966). In other words, cladists are interested in answering questions like: "how many branches there are among a group of organisms?"; "which branch connects to which another branch?"; and "what is the branching sequence?" An ancestor-descendant relationships tree-like network is called a cladogram. Thus, a cladogram refers to the topology of a rooted phylogenetic tree, for example, UPMGA, etc.

The BIIGA was developed for near duplicate images of WND and NWND images from MPS till five rounds. We generate a phylogenetic tree by selecting a phenetics approach because this method focuses on the altogether resemblance of phenotypes in grouping and classifying taxa (tips of a phylogenetic tree representing individual organisms) with no philosophical bias. This meets our objective of grouping near duplicate images.

Molecular phylogenetic trees are mathematical statistical ways for understanding grouping and classifying images during the evolution process. Different ways to generate a phylogenetic tree, each has its strengths and weaknesses. Before proceeding further, we need to know the steps involved in the tree reconstruction method. Four steps are used to reconstruct a phylogenetic tree using DNA sequences as explained below:

a. DNA-based multiple sequence alignment of images under test. It helps to get the input data to the bioinformatics tool for tree reconstruction.
b. Transform the aligned data into the reconstructed tree using suitable approach.
c. Accuracy of tree is accessed
d. The molecular clock is used to allot dates and branch points within the tree.

Step (a) is explained in the previous Section 2.2.3 in detail; we are not discussing it again.

Step (b) is needed to explain in detail that is explained in the following text.

Transform the aligned data into the reconstructed tree using a suitable approach.

Once we get the MSA of DNA data, an effort can be made to reconstruct the phylogenetic tree. The way, in which MSA is transformed to numerical data, will develop the differentiation among tree building approaches because this numerical data is explored mathematically for tree reconstruction. These approaches are further classified into two parts: The first part is directly based on character, sequence phylogenetic tree methods; the second one is indirectly based on character, sequence phylogenetic tree methods are explained below:

a. Indirectly based on character or sequence phylogenetic tree methods are explained below:

   - *Distance-based methods :* Between a pair of sequences, evolutionary distance is computed by using the difference in the number of nucleotides. That value is utilised to determine the lengths of the branches joining these sequences during tree reconstruction. A collection of all evolutionary distances is put in a matrix called a distance matrix (Brown T. , 2002).

Table 2.1 Advantage and limitation of distance based methods

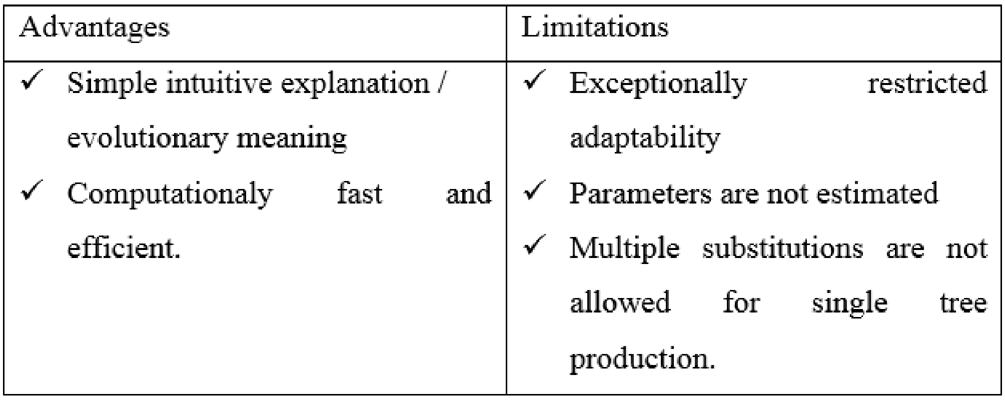
Advantages	Limitatons
✓ Simple and flexible (many algorithms are available). ✓ Computationaly fast and efficient.	✓ Sensitive to gaps in sequence alignment. ✓ Parameter estimation not done. ✓ Only one tree produced. ✓ Simplistic (multiple substitutions are not accounted for).

Examples using this approach are Neighbour Joining (NJ), Unweighted pair group method with arithmetic means (UPMGA) and Minimum Evolution approach. Advantages and limitations of this approach are explained in Table 2.1.

b. Directly based on character, sequence phylogenetic tree methods are described by Table 2.1:

   (a) Maximum Parsimony. Discover a phylogenetic tree that clarifies the information, with a couple of transformative or evolutionary changes as could reasonably be expected or with the most modest number of substitutions. Advantages and limitations of this approach are explained in Table 2.2.

Table 2.2 Advantage and limitation of Maximum Parsimony method



Advantages	Limitations
✓ Simple intuitive explanation / evolutionary meaning ✓ Computationaly fast and efficient.	✓ Exceptionally restricted adaptability ✓ Parameters are not estimated ✓ Multiple substitutions are not allowed for single tree production.

   (b) Maximum likelihood. The Maximum likelihood is a statistical methodology generated by R.A. Fisher in 1920 to determine unknown parameters in a model (Yang & Rannala, 2012). Parameter values are maximised by the maximum likelihood estimates (MLEs) and frequently calculated by using iterative optimisation algorithms. Desirable asymptotic (large sample) properties that MLE have, are that they are consistent (approximate the correct values), efficient (least variance among unbiased approximates) and unbiased.

   Felsentein developed a maximum likelihood analysis algorithm the first time for DNA sequence data (Felsenstein, 1981). A maximum likelihood approach is a phenetic approach that is statistically well established. It tries to find a tree that magnifies the probability of the genetic data during tree reconstruction. It often has less variance than other approaches (i.e., the approximation method is affected minimally by sampling error). Furthermore, for very short sequences of tree reconstruction, the maximum likelihood is inclined to perform better than other methods such as distance or parsimony methods. It averages all possible ancestral states, but parsimony uses only optimal states. A significant limitation of using this topology is that it is very CPU intensive leading to a very time-consuming process not suitable for bigger datasets. Advantages and limitations of this approach are explained in Table 2.3.

To date, none of the tree-building methods makes sure to reflect correctly the evolutionary relationship of a sequence set (Felsenstein, 1988). Advantages and disadvantages of the different approaches from Table 2.1 to 2.3 and the text, indicate and inspire us to select, the phenetic based maximum likelihood approach for developing BIIGA in this thesis.

Table 2.3 Advantages and limitations of maximum likelihood method

Advantage	Limitations
✓ Theoriticaly justified very well. ✓ Assumptions are explicit resulting in that they can be improved and evaluated. ✓ ML approaches are generally consistent. ✓ In most cases, by applying sequence simulation experiments, it had shown that this approach outperforms than all others.	✓ CPU intensive and need long calculation time to reconstruct a tree. ✓ If the model is not specified properly or miss-specified then it has potentially poor statistical properties.

This is because of its flexibility, consistency with a model of evolution and its statistical consistency for model comparison and parameter estimation. Now, after tree construction, we need to check tree accuracy; we perform in two ways as discussed below in the third step of the tree reconstruction.

Step (C) Accuracy of tree is accessed

Due to some limitations in the development of a phylogenetic tree automatically question arises about the correctness of the tree. The accuracy of the tree is tested by using statistical tests developed by (Hillis, 1997; Whelan, Lio, & Goldman, 2001). These tests are complex because the tree is geometric rather than numeric, the correctness of one part may be lesser or greater than the correctness of another part of the tree.

The standard method used for testing the tree is bootstrap analysis, this method applies confidence limits (degree of confidence) at different branch points. This test is based on random sampling with replacement. In the case of phylogenetic tree, it will develop a random alignment for new phylogenetic tree creation. Usually, 1000 new alignments are created resulting in 1000 phylogenetic trees. If its value is greater than 700/1000, then we can allocate a degree of confidence to that particular node. One problem is the precisions are represented. For example, if two groups give a bootstrap value of 90% and they are independent (assumed), the probability of both groups being accurate is $(90/100)^2$= 81%. As the number of groups increases, the overall confidence values become meaningless very rapidly, this test is not suitable for testing large numbers of branches, but it is appropriate for one or two main branches or a small part of a phylogenetic tree (Whelan, Lio, & Goldman, 2001). We performed bootstrap analysis by using a MEGA7 tool for phylogenetic tree creation. Additionally, some statistical tests were also carried out for further verification of correct grouping of the images in their respective categories. These statistical measures are sensitivity (true positive rate), specificity (true negative rate), precision, and negative predictive value (Parikh, Mathai, Parikh, Sekhar, & Thomas, 2008). These are explained below in detail.

- *Sensitivity (True Positive Rate).* This estimates the performance of a binary classification. Other names for this test are *recall or probability of detection.* It is stated as the ratio of the positives that are accurately discovered. For example, in the thesis after phylogenetic tree creation, the sensitivity will tell us the percentage of test images that are accurately grouped in their respective categories.

$$Sensitivity = \frac{Number\ of\ true\ positives}{Total\ numberof\ positive\ samples} \tag{2.4.1}$$

- *Specificity (True Negative Rate):* This is also used for measuring the performance of binary classifications. It estimates the ratio of negatives that are identified. In this thesis, specificity will tell us the percentages of images correctly rejected for grouping and classifying.

$$Specificity = \frac{Number\ of\ true\ negatives}{Total\ numberof\ negative\ samples}$$

(2.4.2a)

$$Specificity = 1 - false\ positive\ rate$$

(2.4.2b)

- *Precision (Positive Predictive Value).* This calculates the proportions of the positive results in other words describes the performance of a test or measure. This thesis uses positive predictive value (precision) to determine the probability of the correct grouping when it is done correctly. If the precision is higher, i.e., as close to 100 as much possible, then this indicates that phylogenetic tree grouping is correct and good.

$$Precision = \frac{Number\ of\ true\ positives}{Numberof\ true\ positive + Number\ of\ false\ positive}$$

(2.4.3)

- *Negative Predictive Value (NPV).* This calculates the proportion of negatives that are not grouped incorrectly. NPV is employed in this thesis to determine the proportion of images belonging to one group and are not grouped in the wrong clade.

$$NPV = \frac{Number\ of\ true\ negatives}{Total\ numberof\ true\ negative +\ Total\ number\ of\ false\ negatives}$$

(2.4.4)

Step (D) Molecular clock used to allot dates to branch points within the tree.

This is the last step in the phylogenetic tree creation where dates are assigned to the branch points in a phylogenetic tree, a molecular clock must be used for this process. This step is out of the scope of this thesis, so it is not discussed further.

Another fascinating work of such unconventionality is the Video Genome Project, in which a video is mapped or encoded in DNA sequences and examined with phylogenetically related tools (Bronstein, Bronstein, & Kimmel, 2010) for searching, matching and comparing two videos in large datasets. In this thesis, we will continue and extend this unusual way of logic by generating a phylogenetic tree for watermarked / non-watermarked images with their near duplicate images generated due to MPS. Watermarking concepts are explained in the next section for understanding better about images under the analysis process in this research project to develop BIIIA and BIIGA.

2.3 Image Watermarking Systems

Digital image watermarking comes under the high level of image processing (Pour, 2015). A digital watermark is secure data (e.g. a pattern of bits, a sequence of characters) embedded to digital data (e.g. photos, videos, scanned documents) that can be later extracted to make an assertion regarding copyright of the data. Secure digital watermarks (typically invisible identification codes embedded in the image data) should be robust in three ways at least: they should be difficult to erase or forge; they should not affect the quality or accuracy of the image; they should not be affected by any compression techniques prior to transmission. Identification of watermarked images from the rest of the images is normally done by applying a watermark extraction algorithm. Watermarking approaches (Panah, Schyndel, Sellis, & Bertino, 2016) will use a unique and different ways to embed and extract watermarks inside an image. Full knowledge of the watermarking system is a must to extract the watermarks; if the watermarked image is degraded, then it is very hard to identify the original watermark. This motivates us to study the identification of degraded and watermarked images and leads to a question: "Is there any universal approach that will discriminate the watermarked images from the non-watermarked images or irrespective methods for embedding and extracting the watermarks from the degraded images?" Finding suitable approaches in image analysis that can fulfill this purposes is currently a major research challenge that inspired us to develop BIIIA for watermarked / non-watermarked images. Further understanding needs exploration about the components and properties of the watermarking system.

2.3.1 Background and Components of Digital watermarking

Without referencing the system and security requirements, there are three main components of a watermarking approach: a watermark generator which generates the desired watermark, desired watermark selection, an embedder which embeds the watermark, detector, and extractor as shown in Figure 2.1. The design of all these components will be potentially influenced by the inputs and outputs, use of keys and relevant constraints derived from the application specific requirements. These components are explained in detail for better understanding.

Media to be watermarked

Different media are watermarked for copyright protection or other security purposes. The media used for watermarking can be a document, a clip of audio, a video or an image. In this thesis, the core focus is on image analysis that leads to use images for watermarking and image analysis.

Watermark Generator or Selection of Data used for Watermarking

Watermark selection is a very critical task for watermarking. Figure 2.2 will show the taxonomy of data that can be used as a watermark. Media like text, audio, video, and images, etc. is used as a watermark from the beginning of the era of watermarking. The aim of media watermarking is to sustain the standard media operations as shown in Figure 2.4, like geometric transformation, compression, rotation, etc.

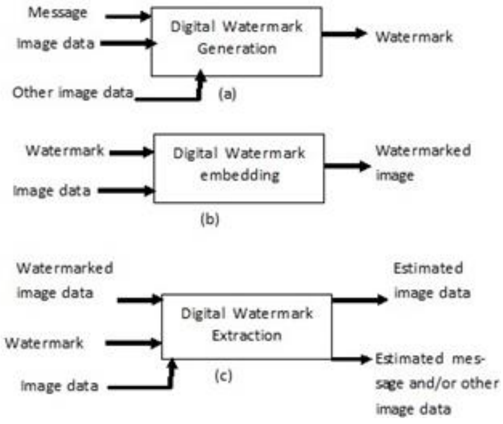


Figure 2.1 Fundamental components of digital image watermarking (a) watermark generation (b) watermark embedding (c) watermark extraction

On the other hand, watermark, such as biological sequence DNA, etc., have different common operations like clustering, sampling, summarisation and alternative data or dimensional reduction techniques (Panah, Schyndel, Sellis, & Bertino, 2016). Further details about media / non-media watermarks are presented below:

- *Media data as a watermark:* There are four types of media data considered for watermarking as explained below:
  a. Text: alphabets and numbers in any combination
  b. Audio: small pieces of audio songs or sound
  c. Video: small pieces of video
  d. Image: any digital image

- *Non-media data as a watermark:* these non-media data typically come under a complex data type (Jiawei Han, Kamber, & Pei, 2011) that can be categorised as follows:

a. *Sequence data.* This includes an ordered list of items, further classified by the characteristics of the events in three groups as follows:

Time series. Numeric data long sequences documented per minute, per hour, or per day (at equal time intervals). For example, stock markets, medical, scientific or natural investigations.

Symbolic sequences. Events or nominal data long sequences, not recorded at equal time intervals. Gaps (time intervals between observed events) for several sequences do not matter. Event sequences are for natural / social developments and science / engineering as well as web click streams / customer shopping sequences.

Biological sequences. Very long and complicated, carrying necessary information DNA, protein sequences having hidden semantic meaning. Gaps between the sequences are crucial.

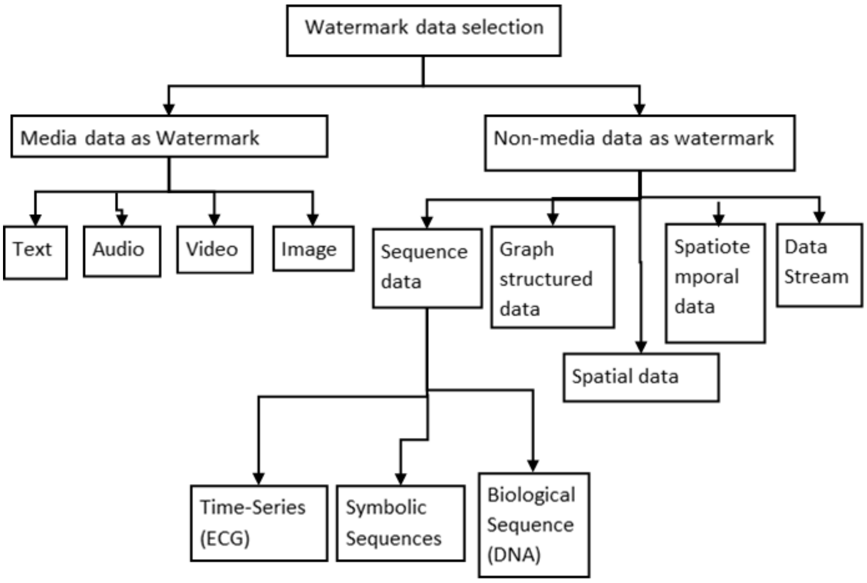


Figure 2.2 Taxonomy of data used for watermark

b. *Graph structured data.* Data values are presented by using vertices and relationships between them and dispense with edges in the graph, e.g., social networks, biological networks, bioinformatics and chemical informatics, etc.

c. *Spatial data.* This kind of data recognises spatial information about objects for example maps.

d. *Spatiotemporal data.* A particular form of spatial data that can be used to denote space and time of moving object trajectories.

e. *Data Stream.* This relates to the immense volume of data that flows into the system that can change dynamically, having multidimensional features. Stream data are present in power-grid flows, computer network traffic, etc. Applying stream data mining applications, we can detect abnormalities in the above data streams

In this thesis, we will explore basic potentialities of the idea to use bioinformatics tools for pattern matching in image analysis, i.e., grouping and identifying. It's beyond the scope of this thesis to test all watermark data types; some of the media data, i.e., text, image, and shape are used for watermarking.

Embedder

For embedding the watermark, the first task is to choose the domain for watermarking and then the algorithm specific for that domain. Based on the domain for embedding the information, current watermarking schemes can be classified into spatial domains and transform domains (Cox, Miller, & Bloom, 2002). This categorisation originated from media representation that can be expressed in a transform domain by using its spectral frequency coefficients and in a spatial domain by using its pixel values. In the same manner, watermarks are embedded by changing the transform domain coefficients like discrete wavelet transform (DWT) or discrete cosine transform (DCT) or in a spatial domain by altering the pixel values, for instance, Least Significant Bits (LSBs) of the pixel. Another way to embed the watermark is by using a combination of both spatial and transform domains to enhance the security of the media.

Problems addressing domain selection for identification of the watermark are very complex and challenging for embedding the watermark. Every domain has its advantages and disadvantages.

The spatial domain has the benefit of maintaining location information. Efficient translation variant of Fourier domain magnitudes and DWT have gained much interest as a multi-resolution representation to conquer the format conversion, mainly JPEG. Anu et al. proposed a watermarking approach by mixing multiple domains and by exploiting individual domain strength. A circular template watermark is embedded in the magnitude of the Fourier transform to invert rotation and scale after the print-and

-scan process; another template watermark is embedded in the spatial domain to invert translations. The message watermark is inserted in the DWT domain (Pramila, Keskinarkaus, & Sepp, 2008). But comprehensive analysis about this approach is missing and they did not compare their approach with any other methods which can robustly identify the degraded watermarked image from MPS. This research work will enhance the work by comprehensive analysis of degraded watermarked image identification by using bioinformatics tools and approaches in pattern matching. Initially, DWT in the frequency domain was used for testing the capability of ideas. Further analysis and checking of the proposed approach were done by using professional watermarking software uMark without having the domain knowledge used for watermarking.

Detector

This is the last component for watermarking that is used for extracting the embedded watermark. Any watermarking algorithm is not considered complete until they develop an extracting algorithm. Detector or watermark extracting algorithm development depends on the algorithm used for the watermarking.

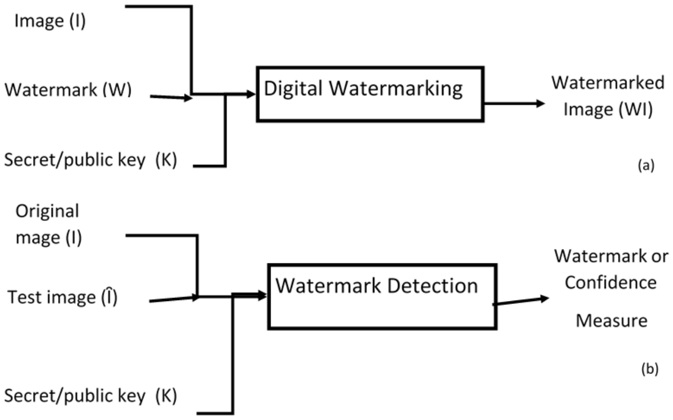


Figure 2.3 Generic digital watermark approach (a) embedding approach (b) recovery approach

The goal of this thesis is to identify the watermarked image, independent on the embedding approach without extracting it, with no idea about the extractor algorithm using bioinformatics tools and approaches.

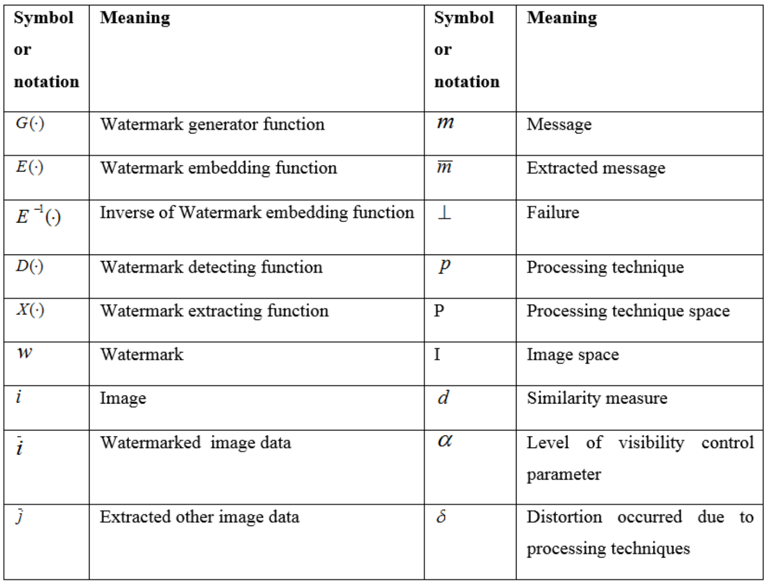
The generic model for embedding and extraction of watermarks is shown in Figure 2.3 for a better understanding of watermarking approaches. Where, Figure 2.3 (a) explains the watermark embedding into the stego image with a secret/public key using any of the digital watermarking approaches to get the watermarked image. In the same way, Figure 2.3 (b) explains the watermark detection using a watermark and/or original image with a test image and the secret/public key to check whether the watermark exists or not in the test image.

For better and further understanding of watermarking, some fundamental properties related to this research are explained in the next subsection.

2.3.2 Fundamental Properties of Watermarking

All fundamental properties defined by Nyeem et al. (Nyeem, Boles, & Boyd, 2014) were used because these are the most recent and suitable definitions for our research work. Notation and symbols used during the literature survey are assembled in Table 2.4. These properties are discussed because they play a major role in the selection of metrics used for evaluations of degraded images.

Table 2.4. Notation and symbols used in watermarking



Symbol or notation	**Meaning**	**Symbol or notation**	**Meaning**
$G(\cdot)$	Watermark generator function	m	Message
$E(\cdot)$	Watermark embedding function	$\overline{m}$	Extracted message
$E^{-1}(\cdot)$	Inverse of Watermark embedding function	$\perp$	Failure
$D(\cdot)$	Watermark detecting function	p	Processing technique
$X(\cdot)$	Watermark extracting function	P	Processing technique space
w	Watermark	I	Image space
i	Image	d	Similarity measure
$\bar{i}$	Watermarked image data	α	Level of visibility control parameter
$\bar{j}$	Extracted other image data	δ	Distortion occurred due to processing techniques

Perceptual similarity (imperceptibility): This suggests the perceptual content of the two images should be sufficiently similar to each other. It is mainly studied for invisible watermarking approaches. Quantitative definitions of perceptual similarity are given below.

Definition 1. (Perceptual Similarity). *Any two images* i_1 *and* i_2 *are said to be* (s,t) *perceptually similar* $s_1(i_1,i_2) \leq t_j$ *for all similarity measures* $s_j \in \{s_1, s_2, ..., s_n\}$ *and threshold* $t_j \in \{t_1, t_2, ...t_n\}$. *For example,* $s_1 = PSNR$ $s_2 = MSSIM$ *that* *represents; there are two similarity measure, given thresholds for these measures,* $t_1 = 60(dB)$ *and* $t_2 = .995$. *Two images are said to be perceptually similar if both* $s_1(i_1, i_2) \geq 60$ *and* $s_1(i_1, i_2) \geq .995$ *are satisfied.*

For evaluation or quantification of perceptual similarity, there are various metrics like Correlation Coefficient (CC), Signal to Noise Ratio (SNR), Mean Square Error (MSE), Peak Signal to Noise Ratio (PSNR), etc. Currently, the lack of globally agreed and efficient measures for visual quality exists (Bovik, 2009; Tefas, Nikolaidis, & Pitas, 2009). Also, not all metrics gives similar estimations; we have to choose *n*-suitable metrics for defining the similarity measure (d) which depends on the application specific requirements.

Visibility. A visible watermarking approach is used to show some necessary information such as company logo, icon, or courtesy by deliberately inserting a watermark such that it appears noticeably on the watermarked image. However, a parameter α is used to control the level of visibility so that the watermark does not become so prominent that it starts blurring the main image. Visible watermarks are necessary for recognition and support of possessing a digital image.

Definition 2 (Visibility). *A watermarking approach is called visible or perceptible, if* $E(\cdot)$ *embeds a given watermark* w *into an image* i *such that the* w *appears at least noticeably in* $\bar{i}$. *That is,* $|E_e(i,w) - i| = \alpha w$ *for all* i, w. *Here* α *is the weight factor that controls the degree of visibility.*

A watermarking approach is called invisible or imperceptible, if $E(\cdot)$ *embeds* w *into* i *such that the* $\bar{i}$ *is perceptually similar to the original image* i. *That is* $E_e(i,w) \approx i$ *for all* i, w *;*

Blindness. Cox et al. informally differentiated blind (oblivious) / non-blind (informed) watermark detector on the basis of access to the original image or some information derived from the original image. Blind one does not need any access to the original image, but other one does (Cox I. , Miller, Bloom, Fridrich, & Kalker, 2008). However, these definitions are not sufficient to realize the three different cases associated with the blindness property. So, Nyeem et al. redefined it again as shown below.

Definition 3 (Blindness). *A watermarking approach is called blind (or oblivious) if both* $D(\cdot)$ *and* $X(\cdot)$ *are independent on the original image* i *and watermark* w*. Formally, for all images* i_1, i_2 *and watermarks* w_1, w_2*, hold both*

$$D_d(\bar{i}, i_1, w_1) = D_d(\bar{i}, i_2, w_2) \tag{2.3.1}$$

$$X_x(\bar{i}, i_1, \tilde{w}_\backslash) = X_x(\bar{i}, i_2, \tilde{w}\ ) \tag{2.3.2}$$

A watermarking approach is called semi-blind if either one of $D(\cdot)$ *or* $X(\cdot)$ *is dependent on the original image,* i *and /or watermark,* w*. Thus, for semi-blind watermarking for all images* i_1, i_2 *and watermarks* w_1, w_2*, either*

$$D(\bar{i}, i_1, w_1) = D_d(\bar{i}, i_2, w_2) \text{ and } X_x(\bar{i}, i_1, \tilde{w}_\backslash) \neq X_x(\bar{i}, i_2, \tilde{w}\ ) \tag{2.3.3}$$

or

$$D_d(\bar{i}, i_1, w_1) \neq D_d(\bar{i}, i_2, w_2) \text{ and } X_x(\bar{i}, i_1, \tilde{w}_\backslash) = X_x(\bar{i}, i_2, \tilde{w}\ ) \tag{2.3.4}$$

Otherwise, a watermarking approach is called non-blind (or non-oblivious or informed) if both $D(\cdot)$ *and* $X(\cdot)$ *are dependent on the original image,* i *and/or watermark,* w*. Formally, for all images* i_1, i_2 *and watermarks* w_1, w_2*, hold both*

$$D_d(\bar{i}, i_1, w_1) \neq D_d(\bar{i}, i_2, w_2) \tag{2.3.5}$$

$$X_x(\bar{i}, i_1, \tilde{w}_\backslash) \neq X_x(\bar{i}, i_2, \tilde{w}\ ) \tag{2.3.6}$$

Invertibility (or reversibility or losslessness). This is a computational property of watermarking where any watermarked images expected to restore its original version when no distortion permitted during embedding of a watermark in an original image.

Definition 4 (Invertibility). *A watermarking approach is invertible (or reversible or lossless) if the inverse of E(·) is computationally feasible to compute and used in D(·) to estimate an exact original image,* i *from the watermarked image* $\bar{i}$*. Otherwise, the approach is a non-invertible watermarking approach.*

If $E_e(i, w) = \bar{i}$, then for an invertible watermarking approach E_e^{-1} , the detection must exist and satisfy $E_e^{-1}(\bar{i}) = (i, w)$. Therefore, such a watermarking procedure can be either blind or semi-blind (according to definition 3). In image applications, an invertible watermarking approach is mainly designed to reverse the effect of embedding on the original image; the embedding function is only used to define the invertibility of approach.

Robustness. Several attempts have been made to define the robustness property of watermarking informally. For example, Piper and Safavi-Naini (Piper & Safavi-Naini, 2009) considered a watermarking approach robust, if it could successfully detect the watermark in the processed images. The strength of this definition depends on how the processed image is defined. On the contrary, Cox et al. defined robustness as the ability to detect the watermark after signal processing techniques (Cox I. , Miller, Bloom, Fridrich, & Kalker, 2008). More specifically, robustness can be defined as the degree of a watermarking approach to modify the host image either by common signal processing techniques or operations devised specifically to render the watermark undetectable (Bovik, 2009). In short, robustness for watermarking has to deal with: (i) the detection ability of the processed image (ii) defining a set of processing techniques. From a signal processing perspective, the two basic requirements for an effective watermarking approach, robustness, and transparency, conflict with each other (Podilchuk & Zeng, 1998).

Definition 5 (Processed Image). *A processed image is one that is not essentially perceptually similar to its original, but has a certain amount of distortion,* δ *is incurred by a processing technique,* $\boldsymbol{p} \in \boldsymbol{P}$ *. That is, if any image* $j \in I$ *is processed by* p *, then, for the processed image* $\boldsymbol{p}(j)$ *, the following is true* $p(j) = j + \delta$ *. Here,* P *is the set of applicable processing techniques for an application such that* $p \in P$*, where P is the space for processing operations.*

Robustness is defined by the detection condition. Suppose a processing technique, $p \in P$ causes distortion of a watermarked image $\bar{i}$. If the watermark is detected from the watermarked image, then it will return$(\bar{i}, \tilde{w})$, or else it will return $\perp$.

Definition 6 (Robustness). *A watermarking approach is defined for the following levels of robustness:*

Robust. A watermarking approach is called robust if $D_d(p(\bar{i}), i, w) = (\bar{i}, \tilde{w})$ *for all* $\boldsymbol{p} \in \boldsymbol{P}$

.

Fragile. A watermarking approach is called fragile if $D_d(p(\bar{i}), i, w) = \perp$ *for all* $\boldsymbol{p} \in \boldsymbol{P}$.

Semi fragile. A watermarking approach is called semi-fragile if $D_d(p(\bar{i}), i, w) = (\bar{i}, \tilde{w})$ *for all* $p \in P_1$ *and* $D_d(p(\bar{i}), i, w) = \perp$*, for all* $\boldsymbol{p} \in (\boldsymbol{P} \setminus \boldsymbol{P}_1)$*, where* $P_1 \in P$.

Embedding Capacity: Embedding capacity for watermarking depends on few other properties like robustness, perceptual similarity, etc. rather than steganography detection problem. Therefore, the embedding capacity is defined on the basis of perceptual similarity of $(i, \bar{i})$, for which the approach works without failure.

Definition 7 (Embedding Capacity). *Watermarking embedding capacity for an image* i *is the maximum size of any watermark* $w = G_g(i, m, j)$ *for all* m *and* j *to be embedded in* i *such that* $E_e(i, w) \approx i$, $D_d(E_e(i, w), i, w) = (\bar{i}, \tilde{w})$ *and there exists* $\tilde{\boldsymbol{m}}, \tilde{\boldsymbol{j}} \mid \tilde{\boldsymbol{j}} = \boldsymbol{j}$ *such that* $X_x(E_e(i, w), i, \tilde{w}) = (\tilde{m}, \tilde{j})$.

In image applications, embedding capacity is usually expressed as the ratio or bit-per-pixel (bpp). According to Definition 7, if watermark embedding capacity is *n* bits, the size of the watermark is m bits (i.e. $w = \{1,0\}^{m}$), then the necessary condition for an invisible watermarking approach is $m < n$. This situation suggests that there should be a hidden assumption of recursive embedding by using an invisible approach. If the required capacity is not achievable in the first run of $E(\cdot)$, the remaining bits can be re-embedded recursively. That assumption severely affects the performance of the watermarking approach in practice, and thus needs to be explicitly stated, if applicable.

Security. Security and robustness are two overlapping concepts in watermarking terminology. Several attempts have been made to define the security and robustness of the watermarking approaches (Kalker, 2001; Pramila, Keskinarkaus, & Sepp, 2008; Voloshynovskiy, Pereira, Iquise, & Pun, 2001). In particular, it becomes a common practice to address security in terms of robustness. The irrespective applications can be said that the security of watermarking reflects different types of attacks (ability to resist any hostile attack) and the robustness is concerned with distortions (ability to withstand distortion) (Villn, Voloshynovskiy, Koval, & Pun, 2006).

Definition 8 (Security). *A watermarking approach is called attack secure if the approach retains security against the attack.*

Attacks on digital image watermarking security are shown in Figure 2.4, where active attacks are responsible for unauthorised removal and embedding.

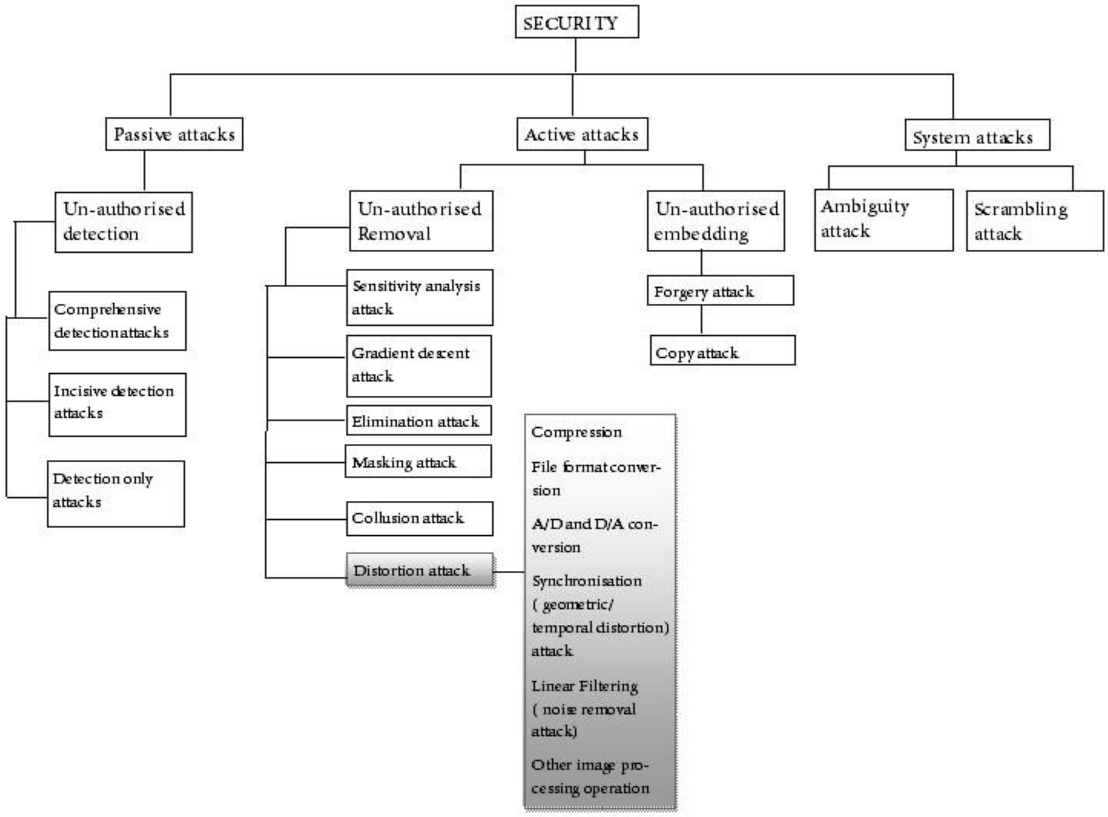


Figure 2.4 The attacks on digital image watermarking

Passive attacks are responsible for unauthorised detection. The proposed research work will focus on the distortions due to MPS on the watermarked image to produce near duplicate images datasets for experiments, highlighted in Figure 2.4.

Definition 9 (Distortion Attack).

Input. A watermarked image $\bar{i} = E_e(i, w_0)$ *and a processing technique,* $q(.) \in Q$ *. where* Q *is the set of applicable processing techniques such that* $Q \subset P$ *.*

Output. A processed image, $q(\bar{i})$

Attack successful condition: $D_d(q(\bar{i}), i, w_0) = \perp$ *but there exists* $w \neq w_0$ *such that* $D_d(q(\bar{i}), i, w) \neq \perp$

Motivation of BIIIA for identification of the degraded watermark

Digital watermarking approaches have received much attention in various digital image applications like medical image watermarking. For the last 30 years general concepts, models, and definitions of digital watermarking have been present but lacking a more comprehensive model that can be used as a basis for discriminating watermarked / non-watermarked images. Extending the problem further, none of the performance analysis tools is available for discriminating degraded images due after MPS, in particular benchmarking tools for instance certimark, optimark (Solachidis, et al., 2001), checkmark (Pereira, Voloshynovskiy, Madueno, Marchand-Maillet, & Pun, 2001), and openwatermark (Michiels & Macq, 2006).

Voloshynovskiy et al. proposed a second generation benchmark (Voloshynovskiy, Pereira, Iquise, & Pun, 2001) which attacks watermarking in a more effective manner than the Fabien-Peticolas-Stirmark tool and concluded that algorithms are robust under Peticolas' attack, performed poorly to the Voloshynovskiy's second generation benchmark tool. This suggests that the claim about degraded watermarked image analysis should persist in the literature and need to be revised by using a standard and comprehensive model; it will provide in-depth analysis about the identification and grouping of WD, NWD, NWND and WD images, which are accurate and universally acceptable.

2.4 Multiple Print and Scan (MPS)

Distortions in the images due to printing are discussed in (Bulan, Mao, & Sharma, 2009; Wu, Kong, You, & Guo, 2009; Ryu, Lee, Im, Choi, & Lee, 2010), and distortions due to scanning are described in (Khanna & Delp, 2010) and for both scanning and printing in (Gaubatz & Simske, 2009; Chiang P. -J., et al., 2010; Chiang P.-J. , et al., 2009). However, all these research works are limited to a single round of printing and scanning distortions. To the best of our knowledge, none of the studies includes discussion of identification and grouping of images after MPS distortion and about the metrics for evaluation of degradation by MPS. To fill this literature gap and to better understand the mechanism of multiple print and scan and its effects, we attempt to identify and group the watermarked / non-watermarked and degraded images after MPS and evaluate image distortion from MPS with the suitable metric selection. That motivates and provides a strong reason for BIIIA and BIIGA development for degraded images after MPS.

During image capture by using a scanner or printer, processes like transmission, filtering, noise addition, halftoning, etc. result in image distortion. For modelling image distortion, various models are proposed. The first systematic study of print and scan distortion was reported by (Lin & Chang, 1999) where a hypothetical model was proposed for the pixel value distortion after print and scan based on their experiments. Although more experiments are needed to verify its validity, they found that this model is appropriate in their experiments using different printers and scanners, as it shows several characteristics of rescanned images.

A document distortion model which simulates four types of noise has been used for validation of distortion models (Kanungo, Haralick, Baird, Stuezle, & Madigan, 2000). Detailed perturbation models of print reflectance modulation resulting from scanner mechanical disturbances have been explained by Loce (Loce & Lama, 1990). In (Moghaddam, 2009) and (Vincent, Nicholas, & Philippe, 2011), a model based on an adaptation of the bleed through restoration method was presented.

Printer and scanner models and methods are introduced (Chiang P. -J., et al., 2010) for an explanation of the scanner architecture as well as embedding an intrinsic signature for the scanner. The same will be explained for laser and inkjet printers where intrinsic (banding and texture based), extrinsic signature and document level signature are embedded for identification of the particular printer or scanner. This thesis also addresses the dearth of research dealing with a lack of end-to-end systems that analyze and ensure the identification and grouping of WD/NWD images from MPS and WND/NWND images.

2.4.1 Image Degradation Due to MPS

The common process, responding for image distortion in print and scan, is the dithering or halftone approach to produce the output. The model for halftone printing by an inkjet printer is introduced (Lee & Allebach, 2005) and the mathematical background related to halftone is discussed (Adler, Kitchens, Martens, Tresser, & Wu, 2003). The reasons for image distortion during the single round of print and scan are the following (Baird & Chaudhuri, 2007):

- Physics of scanner and printer including defocusing, binarization, etc.

- Human error is encapsulating paper position on the scanner, etc.
- Machine error refers to less toner ink and fluctuation in voltage which results in non-uniform illumination, etc.
- Camera positioning, etc.

The reasons for image distortion in multiple rounds of print and scan include the above in each round, plus progressive and incremental compression, decompression, resizing of images due to experimenting needs and transmission (Email, Fax). Volshnovsky, et al. modelled image distortion due to the print and scan operation as a communication channel problem in which distortions are considered as noise; other problems happen during transmission of the message from one place to another location (Voloshynovskiy, Pereira, Iquise, & Pun, 2001).

Degradation of watermarked Images from MPS

A watermarked image will have extra information about the ownership and other data depending on the requirements of these applications. If an image without watermarks is degraded, then only image related data is lost; but if it is watermarked, then the watermarked data may be lost entirely or partially depending on how robust the watermarking approach will be against degradations generated from MPS.

A mathematical characterization of the print-and-scan process divides it into three subprocesses so as to create a straightforward and practical model that can be used to guide the design of data-hiding approaches that survives the single round of print and scan (Solanki, Madhow, Manjunath, & Chandrasekaran, 2005). Lee et al. proposed a robust watermarking method against the print-and-scan process for dithering halftone images (Lee & Chen, 2016). Still, the aspect for MPS is not acknowledged.

Digital watermarking approaches have been shown to be particularly sensitive to image compression techniques commonly used for transmitting images over the Internet with permanent distortion to the watermark (Zheng, Zhao, Tam, & Speranza, 2003). There have been numerous attempts to provide solutions over the past 25 years (Yaghmaee & Jamzad, 2008; Wu & Liu, 1998). Previous work in this area employed watermarks using different approaches like uniform log-polar mapping-based watermarking (Kang, Huang, & Zeng, 2010), a data-hiding approach for printed binary document with tiny and visible dots to recover synchronism and carry the information for semi-fragile authentication of printed documents. Both are robust to general print and scan, but lacking degradations generated from MPS (Kim & Mayer, 2007) w.r.t. the analysis of WD and NWD/NWND images is missing.

The main problem with these techniques is that an image containing a digital watermark needs to be robust against compressions during transmissions and decompressions. A common way to bypass weak watermarking approaches, however, is to print an image (the data file containing the image and watermark) and re-scan it so that a new degraded image is constructed that corrupts and degrades the watermark. This process can be repeated multiple times if necessary to ensure that no trace of the original watermark remains.

In vast literature associated with digital watermarking, different approaches address watermarking differently. For instance, a new theoretical framework for the problem of data hiding in text documents is proposed by R. Villan et al. which explains how this issue can be seen as an instance of the well-known Gel'fand-Pinsker problem. Costa's setup and the family of quantization-based methods were used to show how they can be applied to text data-hiding applications. They considered a text character as a data structure consisting of multiple quantifiable features such as shape, position, orientation, size, colour, etc. They showed that the previous text data-hiding techniques, namely, character feature methods and open space approaches the specific instances of text data-hiding method based on general quantization. Additionally, they proposed a new method of colour quantization for semi-fragile data hiding in printed text documents. The experimental work confirmed that this method has highly perceptual invisibility and a high-information embedding rate, is fully automatable. They also emphasized that this approach is suitable for document identification, authentication, and tamper proofing applications (Villn, Voloshynovskiy, Koval, & Pun, 2006). However, the literature did not address the issue to discriminate the WD/NWD and WND/NWND images by using bioinformatics tools. Before moving further, the evaluation of watermarked / non-watermarked and degraded images is performed for quantification of degradation. This quantification makes a strong theoretical background for converting visible degradation into digital. That gives us a theoretical and numerical proof that images are degraded. Image quality metrics are required for performing degradation evaluation. The next section explores image quality metrics (IQM) used in the thesis for quantifying image degradation from MPS.

2.4.2 Metrics for Measuring Image Degradation

Image quality metrics can be classified on the basis of different criteria. We summarise the methods used in literature for classifying the metrics.

Image Quality Metric Classification Employed

Considerable amounts of the literature have been published on image quality (IQ) assessment. A lot of evaluation metrics are available in past and current publications, The IQ metrics selected here can be used to evaluate perceptual and pixel data. Two basic approaches currently being adopted in the image quality assessment are: objective and subjective metrics.

Objective metrics (Falk, Guo, & Chan, 2007; Gao, Lu, Tao, & Li, 2009; Wang Z. , 2011) and subjective metrics (Chambah, Ouni, Herbin, & Zagrouba, 2009; Falk, Guo, & Chan, 2007; Wang Z. , 2011) are generally used as notions in image quality assessments, yet both concepts are difficult to define precisely. In the literature, objective metrics refer to calculate the image quality using quantified parameters; subjective metrics means how the opinion of a viewer perceives the image.

Objective metrics are further divided into two parts: statistics-based and Human Visual System (HVS)-based. Subjective metrics are divided into two parts: Mean Opinion Score (MOS) (Chetouani, Beghdadi, & Deriche, 2010) and visually best in the group. But both the subjective metrics are very time consuming due to the dependence on human involvement and the process that they are following. Because of that objective image quality metrics are chosen for proposed evaluation. There is a long list of objective IQ metrics for image degradation assessment, for instance, Entropy (Gallager, 2001; Schneider & Fernandes, 2003; Schwartzkopf, Evans, & Bovik, 2002), Variation Entropy for a Unit Boundary length (VEUB), Variation Entropy for a Unit Area (VEUA) (Hase, 2011), etc.

Different conditions and various viewpoints lead to advantages and disadvantages of each metric. Peak Signal to Noise Ratio (PSNR), like well-established metrics, is replaced by others like Structural Similarity (SSIM) Index metric which complements human subjectivity superiorly. Another encouraging occurrence of SSIM is Universal Image Quality Index (UIQI) that is also considered for evaluation. Image distortion evaluation after MPS is a challenging task as it involves finding suitable metrics for a particular type of image distortion under varying conditions of distortion.

In this thesis, the reason why the selection of metrics is primarily rooted in the characteristics of mathematical statistics and human visual systems in image evaluation. Eight metrics are chosen to calculate the degradation: Bias (Ranchin & Wald, 2000), Correlation Coefficient (CC) (Ye & Doermann, 2013) , Erreur Relative Globale Adimensionnelle de Synthèse (ERGAS) (Wald, 2000), Root Mean Square Error (RMSE) (Wald, 2000), Root Average Spectral Error (RASE) (Wald, 2000), Universal Image Quality Index (UIQI), Structural Similarity Index (SSIM) (Nyeem, Boles, & Boyd, 2015), and Distance Structural Similarity Index (DSSIM). More details about them are explained below.

The metrics Bias reveals the differences between the radiance of a degraded image and the original one. The ideal value of Bias is zero. The Bias is defined as below:

$$Bias = 1 - \frac{mean(Degraded)}{mean(Original)} = 1 - \frac{\bar{y}}{\bar{x}} \tag{2.4.1}$$

where $\bar{x}$ is the mean of the original image while $\bar{y}$ is that of a degraded image from print and scan.

CC discovers the similarity between the original and a degraded image from print and scan in which the values range from -1 to 1. CC is shown as equation (2.4.2).

$$CC(x/y) = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}(x_{i,j} - \bar{x})(y_{i,j} - \bar{y})}{\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}(x_{i,j} - \bar{x})^2 \sum_{i=1}^{M}\sum_{j=1}^{N}(y_{i,j} - \bar{y})^2}} \tag{2.4.2}$$

where $x_{i,j}$ the intensity value of the original is image and $y_{i,j}$ is the intensity of the degraded image, $\bar{x}$ is the mean intensity of the original image while $\bar{y}$ is that of the degraded image and M×N represents the size of the original and degraded images respectively.

Root mean square error (RMSE) and peak signal-to-noise-ratio (PSNR) are the most prominent examples of mathematically-based metrics (i.e. they directly measure the difference of pixel intensity) (Cadik, Herzog, Mantiuk, Myszkowski, & Seidel, 2012). PSNR is unsuccessful to quantify structured and localized errors, and not able to discriminate between various kinds of errors; for example, the same PSNR errors have different impacts on human spectators. That motivates the use of other metrics like RMSE. One of the most used methods to quantify the quality of an image is the Mean Square Error (MSE) (Eskicioglu & Fisher, 1995). It calculates pixelwise similarity of two images though the structural information is not considered. MSE values are very large, that is decreased by the square root of the MSE, called root mean square error (RMSE). RMSE is used in this thesis because it is very simple to implement and can be utilized to measure the change in pixel intensity with every round of print and scan.

The metrics RMSE represents the average deviation from the original of a degraded image.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{M}\sum_{j=1}^{N}(x_{i,j} - y_{i,j})^2}{M \times N}} \tag{2.4.3}$$

where $x_{i,j}$ is the intensity value of the original image and $y_{i,j}$ is intensity value of the degraded image; M×N represents the size of the original and the degraded image respectively.

RMSE is further used to calculate the spectral quality of the degraded images in Root Average Spectral Error (RASE) as shown in equation (2.4.4). RASE explains the percentage of relative spectral error that specifies the average performance of the image distortion in the particular spectral bands. As RASE values are calculated in percentage that indicates we have to multiply results with 100 as shown in equation (2.4.4).

$$RASE = \frac{100}{M}\sqrt{\frac{1}{N}\sum_{k=1}^{N} RMSE(B_k)} \tag{2.4.4}$$

where M is the mean radiance of the N spectral bands of the original image. $RMSE^2(B_k)$ represents square of the root mean square error for *k-th* band (B_k) between the degraded and original image.

Additionally, the RMSE value is used for calculating spectral quality during the image reconstruction process by using the metric Erreur Relative Globale Adimensionnelle de Synthese (ERGAS). The metric ERGAS is used for quantifying the synthesis error of the degraded image from print and scan. Spatial quality is calculated by estimating the sharpness of edge but spectral quality estimation performed by using ERGAS. These values are in percentage that implies results are multiplied with 100 as represented in equation (2.4.5).

$$ERGAS = 100\frac{h}{l}\sqrt{\frac{1}{N}\sum_{i=1}^{N}\frac{RMSE^2(B_i)}{M_i^2}} \tag{2.4.5}$$

where h is spatial resolution (pixel size) of the original image while l is the spatial resolution (pixel size) of the degraded image. Here $h=l$ is considered because both images spatial resolution are same that results in$\frac{h}{l} = 1$. $RMSE^2(B_i)$ represents square of the root mean square error for the *i-th* band (*B_i*) between the degraded and original image. M_i is the mean of the *i-th* band of the original image, *N* is the number of spectral bands, and *i* means the index of each band.

This metric measures global radiometric distortion between the original and degraded images. The ERGAS value decreases as quality increases, i.e., if it is closer to zero, spectral quality of the degraded images is good. It gives an accurate prediction of overall spectral closeness between the degraded and original images.

Both RASE and ERGAS calculate the global spectral quality. Only one metric is sufficient to estimate this but for cross verification purpose, we therefore employed these two metrics.

Due to the serious weakness of not representing the distortions perceived by HVS, the standard quality metrics like Peak Signal to Noise Ratio (PSNR) and Mean Squared Error (MSE) are being replaced by new metrics (Wang & Bovik, 2009) like Structural Similarity (SSIM) Index (Wang, Simoncelli, & Bovik, 2004), that matches better with human subjectivity. SSIM similarity measure has proved to be versatile and robust in multiple environments to date. It considers image degradation as a change in structural information.

$$SSIM = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x + \sigma_y + C_2)} \tag{2.4.6}$$

where μ_x and μ_y are (respectively) local sample means of x and y, σ_x and σ_y are the local sample standard deviations of x and y, σ_{xy} is the sample correlation of x and y after removing their means. The items C_1 and C_2 are small positive constants that stabilize each term; that means, variances or correlations of near zero sample do not lead to numerical instability.

Another metric used for image degradation evaluation is derived from SSIM called structural dissimilarity (DSSIM). This metric is a distance metric extended from SSIM, the formula for DSSIM being shown in equation (2.4.7).

$$DSSIM = \frac{1 - SSIM(x, y)}{2} \tag{2.4.7}$$

Further evaluation of image degradation with a more promising case of SSIM is universal image quality index (Wang & Bovik, 2002) which is made up of three parameters structural comparison luminance comparison and contrast comparison. Where $C_1 = C_2 = 0$ and it is explained below in detail.

(a) Structural comparison (Correlation Coefficient (CC)). By considering the loss of correlation between the original and degraded images using equation (2.4.8):

$$CC = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{2.4.8}$$

where x_i and y_i are intensity values of the original and degraded image, $\overline{x}$ and $\overline{y}$ are means of the intensity values of the original and degraded image. μ_x and μ_y are (respectively) local sample means of x and y, σ_x and σ_y are the local sample standard deviations of x and y, and σ_{xy} is the sample correlation of x and y after removing their means. *N* represents the size of the original and degraded image.

$$\sigma_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)^2} \quad , \quad \sigma_y = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \mu_y)^2} \tag{2.4.9a}$$

and $$\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y}) \tag{2.4.9b}$$

for all <u>i</u>=1,2,…,*N*

(b) Luminance comparison (Mean Luminance (L)). By considering luminance distortion between the original and degraded image using equation (2.4.10):

$$L = \frac{2\mu_x \mu_y}{\mu_x^2 + \mu_y^2} \tag{2.4.10}$$

where,

$$\mu_x = \overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \quad , \quad \text{and} \quad \mu_y = \overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \quad , \quad \text{for all } \underline{i}=1, 2,\ldots N \tag{2.4.11}$$

(c) Contrast comparison (D). By considering contrast distortion between the original and degraded image using equation (2.4.12):

$$D = \frac{2\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2} \tag{2.4.12}$$

where,

$$\sigma_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x - \overline{x})^2 \quad , \quad \sigma_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y - \overline{y})^2 \tag{2.4.13}$$

Combining equations (2.4.8), (2.4.10), and (2.4.12) results in the formation of an universal image quality index (Q) for image comparison between two images as shown in equation (2.4.14):

$$Q = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \cdot \frac{2\mu_x \mu_y}{\mu_x^2 + \mu_y^2} \cdot \frac{2\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2} \tag{2.4.14}$$

Literature surveys of image analysis, images analysis in bioinformatics, watermarking approaches and evaluation of MPS problems with their objectives, underlying theory and limitation are studied. Gaps in them motivates us to evaluate MPS degradation and add two new comprehensive approaches BIIIA and BIIGA, in image analysis and bioinformatics literature that will provide us a method of pattern matching using bioinformatics tools. That can be used for identification and grouping of WD / WND, NWND / NWD, and W / NW images. The tools and materials, required to develop BIIIA and BIIGA, are explained in the next section

The research issues related to the evaluation of MPS degradation and development of BIIIA and BIIGA are addressed in the next section.

2.5 Research Problems

Multiple open issues in bioinformatics-inspired image analysis remain. Discussion about these open issues is explained below (Bicego & Lovato, 2016).

- Evaluation of MPS degradation is not addressed in the literature. For providing a base of degradations in terms of values and numbers, MPS degradation is evaluated by using image quality metrics. That provides a base for the research and development of BIIIA and BIIGA. (Chapter 4)
- Universal mapping of images to DNA is lacking: existing techniques of bioinformatics-based image encoding to DNA are varied from the approach to approach motivation. Therefore, it is vital to explore biological image encoding in bioinformatics inspired image analysis. That will help to develop a suitable scheme of image to DNA encoding for BIIIA (Chapter 5).
- Image identification is a popular issue in the image analysis area, but most of the image analysis approaches are based on the feature extraction and then use some classifiers or neural networks or data mining approaches to identify and group the images. A less unexplored area is the unconventional way of

thinking that is: "Is it possible to use bioinformatics tools for image analysis?" (see Chapter 5)

- Watermarked images are analysed either for the development of the robust approach against different attacks or for the extraction of watermark successfully for various verifying purposes like authentication, ownership identification, etc. A key question in the watermarking area is: "Is it possible to identify a watermarked image without developing a watermark identification or extraction algorithm using an approach based on bioinformatics?" Also, the open issue of whether discrimination of watermarked / non-watermarked images is possible by using bioinformatics tools needs to be explored. (see Chapter 5)
- Analysis issues of degraded W / NW images from MPS degradation is another open problem. Less investigated is the identification of W / NW images degraded after MPS as W / NW images as per their originality by using bioinformatics tools. (Chapter 5)
- Image grouping is a crucial task for the image analysis researcher. In the literature, main focus is on 2D and 3D shape recognition and classification by using bioinformatics tools (Bicego & Lovato, 2016; Bicego, Danese, Melzi, & Castellani, 2015; Bicego & Lovato, 2012). How bioinformatics tools can be used for grouping of degraded images from MPS degradation is still an open issue to work out (see Chapter 6).

The above problems show a significant number of complications to bioinformatics -inspired image analysis. There is a real requirement of inspecting the above research issues. However, from the surveyed literature above, there has been no attempt so far towards universally and biologically encoding of images, or identifying and grouping the degraded and watermarked / Non-watermarked images using bioinformatics tools. This 'dearth' revealed in Chapter 1 has been expressed through systematic reviews of bioinformatics tools for image analysis; in the next chapter, how the research methodology will be used to fulfill these literature gaps is addressed.

2.6 Summary

This chapter has reviewed bioinformatics tools and approaches in image analysis for identifying and grouping images including pattern matching as well as bioinformatics tools used in this thesis. Image watermarking components and their fundamental properties were discussed with the degradation from MPS. This chapter also explored and outlined the eight well established metrics (i.e., bias, CC, ERGAS, RASE, RMSE, UIQI, SSIM and DSSIM) for evaluating degradation of non-watermarked / watermarked images after MPS.

Current techniques typically do not identify and group watermarked/non-watermarked and degraded images from MPS and still have limitations of universally accepted biologically encoding of images. Developing BIIIA and BIIGA methods that incorporate practical bioinformatics tools like sequence alignment and phylogenetic tree tools is still an open issue. Integrating bioinformatics tools for image analysis remains untraditional, due to the practical difficulty of integrating these bioinformatics tools to develop BIIIA and BIIGA using pattern matching with accurate identification, grouping watermarked/non-watermarked and degraded images from MPS as well as the real difficulty of developing universally and biologically encoding for a variety of images.

Chapter 3

Research Methodology

The issue of previous work in image analysis and bioinformatucs is to be resolved through the proposed research questions. In Section 3.1, we introduce the general overview of BIIIA. The empirical research methodology used in this thesis is explored and the research problems and questions will be addressed in Section 3.2 and Section 3.3, respectively. In the next section, we describe the proposed methods for evaluation, identification, and grouping images. In Section 3.5, *we explain the design of experimental methods; In Section* 3.6, *we deal with the analysis and result validation. Evaluations of the output and its review are described in section* 3.7. *In Section* 3.8, *we explain about the tools and materials for our research work. Lastly, we summarise this chapter.*

3.1 Introduction

Chapter 1 introduced the motivation for this research project in this thesis, which addressed the gap existing in bioinformatics for pattern matching and image analysis: the use of bioinformatics tools for identification and grouping watermarked/ non-watermarked images and their degraded copies generated from MPS. Chapter 2 focuses on literature review and details this gap with much work concentrating on image degradation generated from a single round of print and scan and algorithm development for robust watermarking. Bioinformatics techniques for pattern matching, like sequence alignment, biology-based encoding, etc., were used for malware and virus identification (Naidu & Narayanan, 2016), 2D shape identification (Bicego & Lovato, 2016), 3D shape matching (Bicego, Danese, Melzi, & Castellani, 2015). Therefore, the following question is proposed: "Is it possible to use bioinformatics tools and techniques in the image analysis of watermarked/non-watermarked images and their degraded copies from MPS?" and if yes, "How?" For answering this question methodically within the extent of this thesis, a research methodology is required to guarantee that the problem can be broken into reasonable subparts. Additionally, the development of sub-question to the fundamental research question is taken into account. Ultimately, such a breakdown will help in the trip of discovery; to the best of our knowledge, this is the first time that bioinformatics tools and techniques were used for pattern matching in identification and grouping watermarked/ non-watermarked images and the relevant degraded copies generated from MPS. In other words, this thesis does not state the full answer to the research questions but provides some fundamental directions for our readers. On the trip of discovery, there will be different ways to test the improvement of hypotheses or advanced hypotheses. This chapter address the research methodology selected to direct this research project. The chapter will come back to the research methodology and will interrogate whether it is feasible or possible.

Critically, any acceptable research methodology, given our research questions, should be covered by the research questions; we therefore need to figure out a hypothesis, extract test conditions, execute programming to test these conditions, run the programme, gather the outcomes, report the results, and after that either reformulate the hypothesis or extract new test conditions, and so on. Iteration of the methodology is done till, in a perfect world, the best programming arrangement is found for the best test condition. At the point when time runs out, the value of what has been accomplished should be assessed both in its particular terms against the probability of continuation. At the end of this thesis, we will return to the merits of this research work.

3.2 Empirical Research Methodology

The empirical method is a collection of techniques for interrogating facts, achieving new understanding or rectifying past understanding. This method is a process to respond scientific questions by examining and performing experiments. The basic steps are as follows:

- Ask questions
- Background research
- Hypothesis construction
- Testing of hypothesis by experiments
- Data analysis and conclusion
- Result communication

A challenging part of this research is the absence of past work on the evaluation of image degradation from MPS, and identifying and grouping the watermarked / non-watermarked and degraded images as well as their original copies (i.e., WND and NWND images) using bioinformatics tools and techniques. Broadly speaking, two steps are followed to fulfill the gap of fully autonomous bioinformatics-inspired identifying and grouping watermarked/non-watermarked images and their degraded copies from MPS. The first step is to encode the research problem biologically which involves the biological image encoding and decoding processes. The second step is to apply bioinformatics tools to sequence alignment and get the common substring or signature to put them in a syntactic recogniser for pattern matching. The planning process comprises a group of steps that initiates with the identification of an issue or a requirement that guides to producing and developing a result that answers the issue or satisfies the requirement. Methodology steps used in this thesis are:

- Define the problem
- Do background research
- Design methods

- Perform experiments
- Analyse and Validate
- Report outputs
- Review and Evaluate
- Re-define the questions

Figure 3.1 shows the methodology adopted in this thesis. The next section deals with the explanation of research problems and the questions.

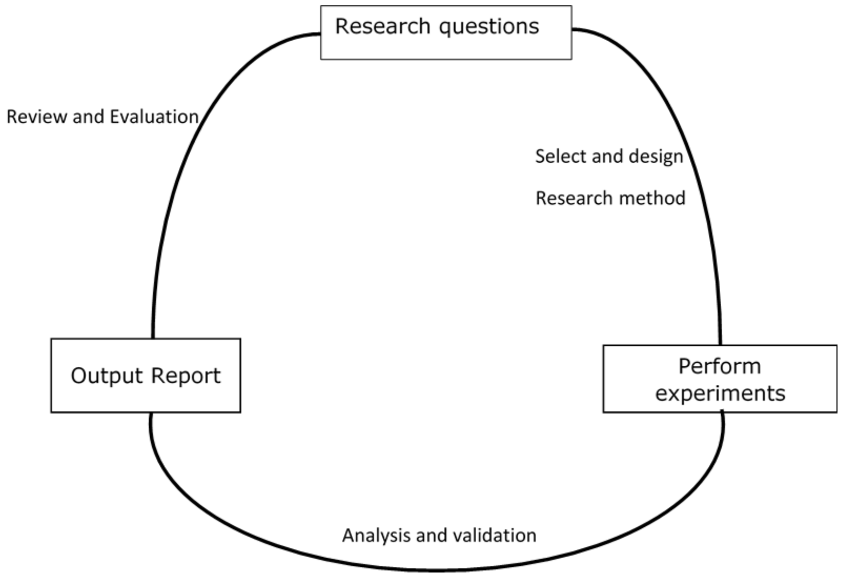


Figure 3.1 Empirical research methodology

3.3 Research Problems and Open Questions

This section explains the problems of using bioinformatics tools and techniques for image analysis using pattern matching, valuations of image degradation from MPS and open questions.

3.3.1 Research Problems

The problem of selecting suitable metrics of image quality for evaluating degradation from MPS on watermarked/non-watermarked images is very difficult because there are no universally agreed metrics of image quality that can perform evaluations of degraded images perfectly. For resolving this problem, the metrics selected for evaluation were based on human visual system and mathematical statistics.

The core problem for this thesis is how to use the bioinformatics tools and techniques in pattern matching for image analysis, especially for watermarked /non-watermarked and degraded images from MPS. This problem is further extended into two parts: the first was how to encode the problem of identifying and grouping watermarked /non-watermarked images biologically. In other words, how to encode an image in the biological DNA. The second problem was how to use the bioinformatics tools and techniques for further processing biologically-transformed images. Alternatively, how bioinformatics tools like JAligner, MAFT, MEGA, etc. (discussed in section 3.8) were applied to the DNA of images for resolving the problem for identifying and grouping watermarked / non-watermarked and degraded images from MPS.

3.3.2 Research Questions

Chapters 1 and Chapter 2 analysed and re-examined the techniques and approaches that were utilised or immensely relevant to the research work conducted in this study. Any effort performed in this thesis is either an advanced achievement of the work investigated in this chapter or an existing approach incorporated in a novel way to accomplish better outcomes.

The literature review in Chapter 2 demonstrated many research questions related to identify and group watermarked/non-watermarked images and their degraded copies from MPS (i.e., WD and NWD images) by using bioinformatics tools and techniques for pattern matching in image analysis. The contribution of this thesis is conveyed by proposing the three research questions as stated below:

Question 1. *Is it possible to investigate and measure degradation of non-watermarked / watermarked images from MPS, if so, what are the suitable metrics? (Chapter 4)*

Question 2. *Is it possible to extract syntactic patterns to identify watermarked (W) / non-watermarked (NW) and the degraded images generated from MPS by using biological representation and bioinformatics alignment algorithms? (Chapter 5)*

Sub question 1: Is DNA biological image representation suitable for identifying watermarked / non-watermarked images?

Sub Question 2: Which algorithm for bioinformatics sequence alignment is the best for identifying watermarked / non-watermarked images?

Sub question 3: Is it possible to identify watermarked / non-watermarked and degraded images from the best biology-based encoding and the best algorithm for sequence alignment from sub-questions 2 (a) and 2(b) ?

Question 3. Is it possible to extract syntactic patterns or signatures for grouping the watermarked (W) / non-watermarked (NW) and degraded images after or before MPS by using biological representation, bioinformatics alignment algorithms and phylogenetic tree? (Chapter 6)

Sub question 1: Is it possible to group non-watermarked and degraded (NWD) images, non-watermarked and non-degraded (NWND) images by using phylogenetic tree analysis?

Sub question 2: Is it possible to group watermarked and degraded (WD) images, watermarked and non-degraded (WND) images by using phylogenetic tree analysis?

Sub question 3: Is it possible to group watermarked / non-watermarked images from a mix of NWD, NWND, WD, and WND images by using phylogenetic tree?

In this study, the first question specifically is a basic one, because a clear interpretation is essential to what it is exactly motivated by studying watermarked/non-watermarked and degraded images from MPS. The second question and its sub-questionss are the foundation questions, for identifying the bioinformatics tools and techniques that are best suitable for image analysis in the proposed research project. As observed from the literature reviews, there is notably insufficient research work on this foundation. Only after answered the second question and the sub-questionss, the third question and its sub-questions could be well addressed. All proposed questions will be answered in the remaining thesis.

3.4 Proposed Method

3.4.1 Limitations of Previous Methods

The drawback of the previous approaches related to image degradation from print and scan was that each method is limited to study a single round of print-and-scan degradation. In addition, insufficient studies have been done for the evaluation of image degradation, identification and grouping methods for watermarked / non-watermarked and degraded from MPS. Furthermore, none of them exploited the bioinformatics tools and techniques, DNA biosequences, PSA (NWA and SWA algorithms), MSA, and phylogenetic trees for identifying conserved regions for discriminating (i.e. identification and grouping) watermarked / non-watermarked and non-degraded images and degraded copies generated from MPS. The findings of this thesis are to employ the bioinformatics-inspired approach that will serve as a base to overcome these drawbacks.

3.4.2 Hypothesis

Hypothesis for evaluating the degradations from MPS

Null hypothesis. The selected eight metrics of image quality were suitable for measuring the watermarked/ non-watermarked and degraded images from MPS.

Alternative hypothesis. None of other measures was appropriate for measuring the watermarked/ non-watermarked and degraded images from MPS.

Hypothesis for BIIIA

The research hypothesis is that, for watermarked / non-watermarked and degraded images, it is possible to identify syntactic structures (patterns) using bioinformatics tools and techniques that help to determine whether a degraded / non-degraded image from MPS contains a type of watermarks or has not a watermark that helps to identify the images of their expected category. If this research hypothesis does not apply to image identification, it is highly unlikely that syntactic structures extracted by using bioinformatics tools will be used for image identification.

Hypothesis for BIIGA

The research hypothesis is that, for watermarked / non-watermarked and degraded images from MPS, it is possible to identify syntactic structures (patterns) using bioinformatics tools and techniques that help to determine whether a degraded / non-degraded image from MPS contains a type of watermark, or has not a watermark or specific degradation properties that help to group images in the expected category. If this research hypothesis does not apply to image grouping, it is highly unlikely that syntactic structures extracted by using bioinformatics tools will be used for grouping images.

3.4.3 Evaluating Image Degradation from MPS

In this research project, the motivation for identifying and grouping watermarked/non-watermarked and degraded images due to MPS comes from the evaluation of image qualities. The idea employed in this thesis is to exploit well-established eight metrics for image quality that are used statistically and mathematically for calculating the image degradation by different means. These eight image quality metrics are used for evaluating image degradation from MPS, for watermarked/non-watermarked images.

3.4.4 The Idea for Developing BIIIA and BIIGA Algorithm

In this thesis, the unconventional idea, using the bioinformatics tools and techniques for pattern matching in image analysis, comes from the Naidu and Narayanan approach of using bioinformatics tools in malware identification (Naidu & Narayanan, 2016). Generally, pattern matching had been used for developing bioinformatics tools and techniques for bio-image analysis. Very few studies was reported in the literature, i.e., bioinformatics tools for pattern matching as the tools for malware identification (Naidu & Narayanan, 2016), 2D shape identification (Bicego & Lovato, 2012), 3D shape matching (Bicego, Danese, Melzi, & Castellani, 2015) and 2D shape classification (Bicego & Lovato, 2016). Two methods are proposed in this thesis: one for image identification, i.e., BIIIA; the second for image grouping, i.e., BIIGA for watermarked/non-watermarked and degraded / non-degraded images after MPS. More details about these methods are given below:

- BIIIA: For the purpose of image identification (i.e., BIIIA) of this thesis, the encoding of images is performed in DNA; then, PSA is employed by using NWA and SWA algorithms; after that, pattern matching is performed by using Clamscan algorithem in Section 3.8.8 for identifying NWND, WND, NWD or WD images.
- BIIGA: To develop BIIGA algorithm, DNA-encoded images were aligned by using MSA algorithms and phylogenetic trees for grouping the NWND, WND, WD and NWD images into their expected categories.

3.5 Design of Experimental Methods

In this research work, the three methods are applied to watermarked/non-watermarked images, and their degraded copies for the purpose of evaluation, identification and grouping. These methods are explained later in this thesis. Each of the methods is the solution to the research issues discussed in Section 2.5. Bioinformatics tools and techniques, like PSA and MSA, etc., were adopted in the methods BIIIA and BIIGA respectively to cope with the problem of identification and grouping of watermarked/non-watermarked and degraded images from MPS and their original copies.

For bioinformatics-inspired image analysis of watermarked/non-watermarked images from MPS and their original copies, a four-step experiment on NWND, NWD, WD and WND images was designed for this research work as shown in Figure 3.2.

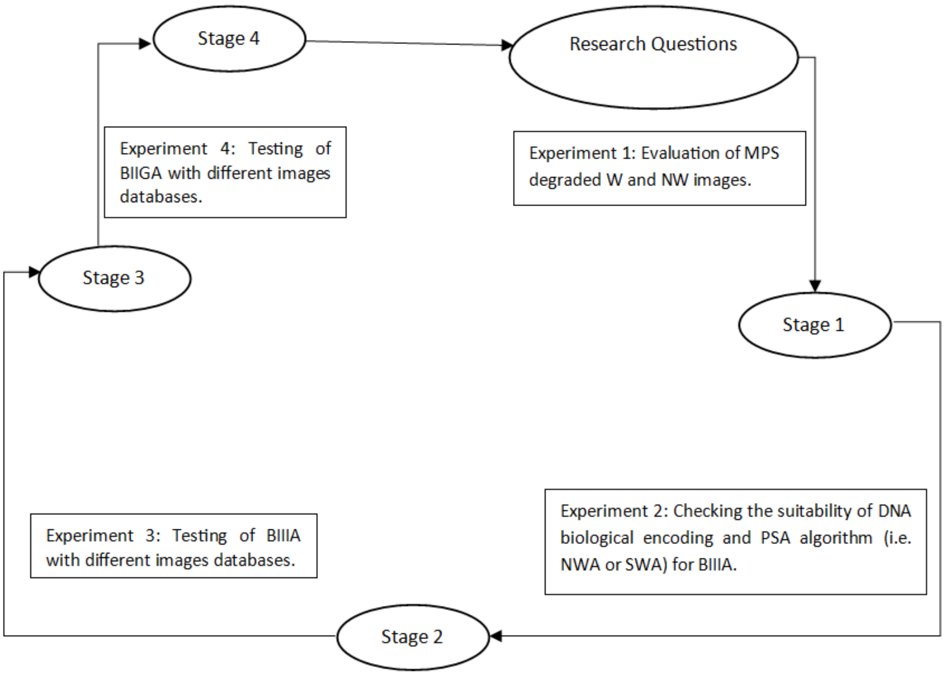

Stage 4
Research Questions
Experiment 4: Testing of BIIGA with different images databases.
Experiment 1: Evaluation of MPS degraded W and NW images.
Stage 3
Stage 1
Experiment 3: Testing of BIIIA with different images databases.
Experiment 2: Checking the suitability of DNA biological encoding and PSA algorithm (i.e. NWA or SWA) for BIIIA.
Stage 2

Figure 3.2 The research roadmap

Stage 1: Evaluation. The watermarked/non-watermarked images and their degraded copies are evaluated for degradation due to the MPS process. Eight metrics were used for evaluation: bias, correlation coefficient, RMSE, RASE, ERGAS, UIQI, SSIM, and DSSIM, which were detailed in Chapter 4.

Stage 2: BIIIA development. The test images are encoded in DNA. PSA, i.e. both NWA and SWA algorithms, employed on the DNA encoded images to identify the suitability for developing the BIIIA algorithm, which will be addressed in Chapter 5.

Stage 3: Test of the BIIIA. After finalisation of the image encoding in biology and the most suitable PSA algorithm for developing BIIIA algorithm is performed based on different datasets to check the robustness of this approach, shown in Chapter 5.

Stage 4: BIIGA. The most suitable biology-based encoding from Stage 2 was used; the MSA was applied to get the aligned sequences. These aligned sequences are used as the input of the tool of phylogenetic tree for grouping the NWND, NWD, WD and WND images, see Chapter 6.

3.6 Analysis of BIIIA and BIIGA

The analysis and evaluation are carried out differently for evaluating BIIIA and BIIGA algorithms based on watermarked /non-watermarked and degraded images after MPS and their non-degraded copies. For image analysis, graphs were drawn to validate the metrics of image quality for our research work. The BIIIA was analysed by using image identification for a particular category of the images (for example correct identification percentage of watermarked images from a group of images). The BIIGA was analysed by using four statistical metrics: sensitivity (true positive rate), specificity (true negative rate), precision (positive predictive value) and negative predictive value for correctly grouping watermarked/non-watermarked and degraded / non-degraded images, for more details, please see Section 2.2.4.

3.7 Evaluation and Review of Output Reports

The proposed methods of evaluation, identification (i.e. BIIIA) and grouping (i.e. BIIGA) will be tested on multiple image sets in Section 3.8.3 so as to answer the following questions: (a) Solvability. Does the proposed evaluation method evaluate degradation successfully? Does the proposed BIIIA succeed in the identification of the watermarked/non-watermarked images and their degraded copies from MPS? Does the proposed BIIGA succeed in the grouping of the watermarked/non-watermarked images and their degraded copies from MPS? (b) Practicability. Is the proposed evaluation method relevant for evaluation of image degradation? Is the proposed BIIIA method relevant to image identification for watermarked/non-watermarked and degraded / non-degraded images after MPS? Is the proposed BIIGA method relevant to a group of images obtained from watermarked/non-watermarked and degraded images and their original copies?

3.8 Tools

Tools and software selection determine the success or failure of the research outcomes. Two steps are performed for providing a strong background for BIIA development.

The first is to evaluate the degradation from MPS with MATLAB R2016a, i.e., numeric values of degradation from MPS provide a theoretical background; the second is the data mining and neural network approaches for identifyhing and grouping degraded images from MPS by using WEKA 3.6 (i.e., very less percentage of identification and grouping by using neural networks and data mining approaches) for dispensing powerful reason to continue with the focus of this thesis.

This thesis focuses on image analysis, inspired by bioinformatics, where the first step is to encode the image biologically (i.e., biological representation of images) and the second is how to use bioinformatics tools for image analysis. For the accomplishment of the first step (to encode the image pattern matching problem biologically), Sections 3.8.1 and 3.8.2 introduced uMark for watermarking and degradation from print and scan that was extended in Section 3.8.3 to explain the process of creation of different images datasets by using the tools described in Section 3.8.1 and Section 3.8.2. These datasets are used for evaluation, identification (i.e., BIIIA) and grouping (i.e., BIIGA). After that, the tools for encoding images biologically are explained in Section 3.8.4, i.e., WUtils and tomeko.net web tools. Then, Section 3.8.5, Section 3.8.6 and Section 3.8.7 describe the second step; i.e., how to use bioinformatics tools in image analysis, i.e., JAligner (local and global sequence alignment tool), MAFT (multiple sequence alignment web tool), MEGA7 (phylogenetic tree creation tool). At last, Section 3.8.8 explains the pattern matching tool Clamscan for image identification.

3.8.1 uMark

Images used for experiments in this thesis are watermarked / non-watermarked. uMark (https://www.uconomix.com/Products/uMark/) is a watermarking software of which free and professional versions are available. For our research, we used the free version. Initially, for testing proposed approach, we used in-house built discreet wavelet transform based watermarking approach; for watermarking bigger datasets, we used the free version of uMark.

3.8.2 Printer and Scanner

Test images were scanned and printed using a Fuji Xerox DocuCentre-V C7775 PCL 6. This machine has the both functionality of scan and print. Different scanning modes are available: scan as a black and white image, a colour image, or a greyscale image.

All images were printed at 300 dots per inch (dpi), scanned at 300 dpi, and saved with .tiff format, as it retains the maximum information of these images. Depending on the type of dataset creation, we chose the scanning mode.

3.8.3 Image Datasets

Image datasets used for experiments were created by using the standard test images. For degradation evaluation, one dataset was created by degrading watermarked/non-watermarked images with MPS. For BIIIA, two different datasets were created. The first dataset was made up of watermarked/non-watermarked images, where 444 non-watermarked images were watermarked with watermarks like text, image and shape using the professional watermarking software: uMark. That resulted in 444 text-watermarked images, 444 shape-watermarked images, and 444 image-watermarked images. For further verifying the BIIIA approach and testing the BIIGA method, a second dataset was created by using the watermarking/non-watermarking images with degradation from MPS. Six standard test images were watermarked using the DWT-based approach; then, these images were degraded by using MPS. More details about the dataset are given below.

A MPS degradation process was used for degrading watermarked/non-watermarked images for creating different datasets. In terms of degrading test images, the same degradation process from MPS was used by replacing the test images with watermarked/non-watermarked images.

Let *O* be the original test image (non-degraded), *P* the printed test image, *S* the scanned test image, and *R* the round number of print and scan. This notation scheme is used for the round of MPS. Test images were degraded at the first time, by printing and scanning, we call it as one round of multiple print and scan (MPS). This is the first cycle; so it is Round 1 (R1). In R1, the original image under test *O* is printed by using a printer and scanned by using a scanner, denoted by using P1 and S1.

R1 (Round 1) = P1→Printed original watermarked image; S1→Scan of P1 image

In short,

$$O \rightarrow \text{<Print>} \rightarrow \text{P1} \rightarrow \text{<Scan>} \rightarrow \text{S1}$$

The same process will be repeated in Round 2, i.e., S1 will be printed to get P2; P2 will be scanned to get S2.

R2: P2→Printed S1 image; S2→Scan of P2 image

In short,

S1 → <Print> →P2→<Scan>→ S2

For Round 3, S2 will be printed to get P3; then, P3 will be scanned to obtain S3.

R3= P3→Printed S2 image; S3→Scan of P3 image

In short,

S3 → <Print> →P3→<Scan>→ S3

Similarly, S3 was printed to get P4, P4 was scanned to get S4; lastly, S4 will be printed to get P5, P5 will be scanned to get S5, so on so forth.

A prominent degradation of images, after five rounds of print and scan, can be seen by using our human naked eye. That forced us to stop further operations of print and scan. In this thesis, because of this reason, all watermarked and non-watermarked images were degraded by MPS till the fifth round of print and scan. Further details about datasets are narrated in the next paragraphs.

a. Dataset for image degradation evaluation

To the best of our knowledge, there is no publicly available watermarked / non-watermarked and degraded image dataset for evaluating the degradation from MPS. That inspired us to create a new dataset for evaluation. Printer and scanner settings used for printing was at 300 dpi and for scanning at 300 dpi, black and white scheme with .tiff extension (as it retains maximum information of image). Dataset details are explained below.

(a) *NWD image dataset.* Six standard NW images, namely, *Baboon, Girl with Blonde hair, Girl with dark hair, Cameraman, Meeting, and Lena* were degraded by using scan and print in Section 3.8.2, till five rounds of print and scan. In total, NWD dataset for evaluation had thirty-six images, namely, six original and thirty degraded images by using five rounds of print and scan.

WD image dataset. Above six standard test images were watermarked, namely, *Baboon, Girl with Blonde hair, Girl with dark hair, Cameraman, Meeting, and Lena.*

A DWT watermarking algorithm was used to embed a watermark inside by using an image having ‘copyright’ as shown in Figure 3.3. All these watermarked images were degraded by using scan and print described in Section 3.8.2, total five rounds. In total, WD dataset for evaluation has thirty-six images; six original and thirty degraded images by using five rounds of print and scan. The WD dataset is shown in Figure 3.4.

Figure 3.3 The image for digital watermarking

b. Image Dataset for BIIIA

The BIIIA approach was tested on two different types of datasets: one was made up of non-watermarked/watermarked and non-degraded images, the second dataset was created, by applying degradation to watermarked/non-watermarked images. The non-degraded image dataset was further divided into four datasets I1, I2, I3, I4. More details about these dataset and its creation are narrated below.

(a) *NWND image dataset (I1):* This dataset consists of 444 TIFF images having multiple resolutions, sizes and other image properties from the famous book in image processing, entitled “Digital Image Processing (3rd edition) ” as original and non-watermarked images.

(b) *WND image datasets (I2, I3, and I4):* The discussed 444 images were watermarked by using three different types of watermarks: text-based watermark (TW) (I2), image-based watermark (IW) (I3), and shape-based watermark (SW) (I4).

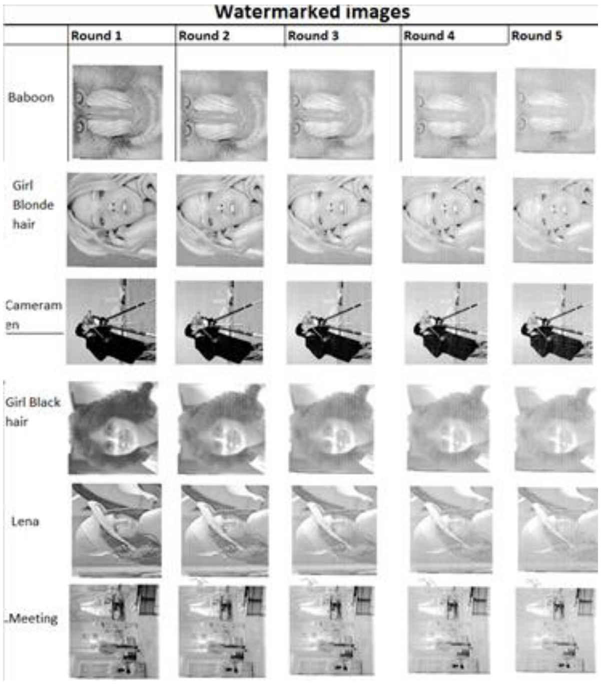
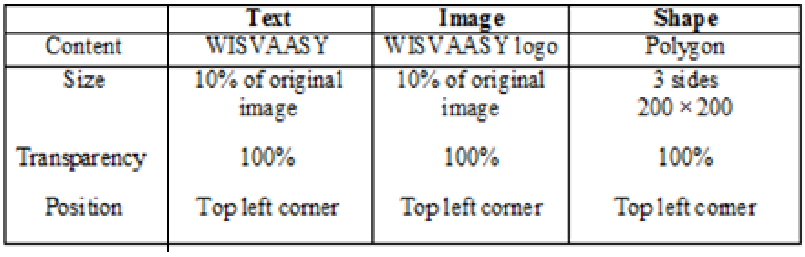


Figure 3.4 Image dataset for MPS

That results in the 1776 image dataset, consisting of 444 original, IW, SW and TW images. Table 3.1 represents the settings, i.e., content, size, position of the watermark used for embedded the text, image and shape as a watermark. All watermarks are embedded on the top-left corner of this image having 10% size of original image. The text watermark consists of text “WISVAASY”, image watermark is a logo of the same text “WISVAASY”, and the shape-based watermark is a polygon with size 200×200 pixels having three sides.

Table 3.1 Settings for embedding a text, image or shape-based watermark



	Text	Image	Shape
Content	WISVAASY	WISVAASY logo	Polygon
Size	10% of original image	10% of original image	3 sides 200 × 200
Transparency	100%	100%	100%
Position	Top left corner	Top left corner	Top left corner

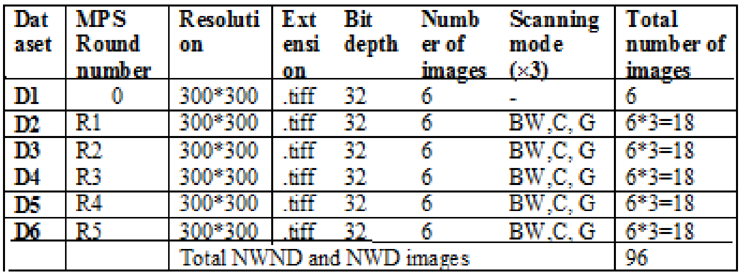
c. Near duplicate image dataset for BIIIA

This dataset was used as a second one for testing the BIIIA approach. This dataset was named as a near duplicate image dataset because near duplicate images are transformed versions of an image by using different operations like contrast enhancement, resizing, etc. The degradation from MPS was used to develop near duplicate images, i.e., the watermarked/non-watermarked images were degraded after MPS. Six standard images were used, namely, *Baboon*, *Cameraman*, *Meeting*, *Girl with black hair, Lena and Girl with Blonde hair* having 300×300 resolutions, and 32-bit depth as the original or non-watermarked images. A total of 12 datasets were prepared which further divided into two parts; the first part consisted of NWD images as well as NWND images (i.e., D1 to D6); the second part was WD images and WND images (i.e., D7 to D12).

(a) NWND and NW degraded (NWD) image dataset (D1 to D6)

Dataset D1 contained six NWND images as described, they were degraded after MPS degradation to five times with three different scan modes: black & white, color, grayscale. All scanning and printing operations were carried out by using the printer described in Section 3.8.2. Dataset D2 was created by using dataset D1 images and printing them first and then scanning with three different scanning modes, i.e., black & white (BW), colour (C) and greyscale (G). We call it as one round of MPS that resulted in 18 degraded images (i.e., six degraded images for each scanning mode). For creating dataset D3 and dataset D2, images were printed and then scanned with following settings: (a) D2, BW images were again scanned in BW mode, (b) colour scanned images from D2 were scanned again with colour mode, (c) greyscale scanned D2 images were scanned again with greyscale scanning mode; for Round 2 of MPS that resulted in 18 degraded images to generate dataset D3; and so on so forth for the third, fourth and fifth rounds of MPS, each round results in 18 degraded images to create datasets D4, D5 and D6 for Round 3, Round 4 and Round 5. In total, 96 images were present in the first part of this dataset (i.e., non-watermarked image dataset) as shown in Table 3.2 (i.e., six non-watermarked or original images, 90 degraded images from Round 1 to Round 5 with different scanning modes, 8 for each round).

Table 3.2 Settings for degrading the non-watermarked images after MPS



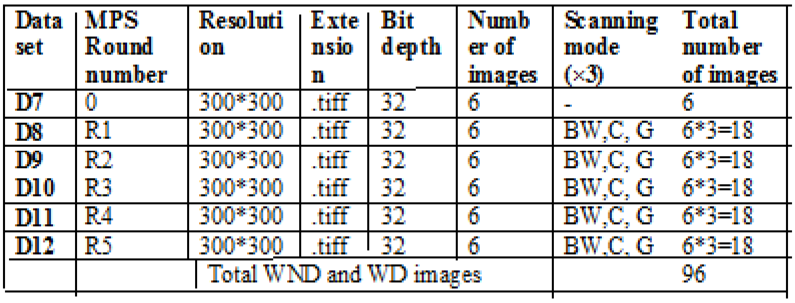
Dataset	MPS Round number	Resolution	Extension	Bit depth	Number of images	Scanning mode (×3)	Total number of images
D1	0	300*300	.tiff	32	6	-	6
D2	R1	300*300	.tiff	32	6	BW,C, G	6*3=18
D3	R2	300*300	.tiff	32	6	BW,C, G	6*3=18
D4	R3	300*300	.tiff	32	6	BW,C, G	6*3=18
D5	R4	300*300	.tiff	32	6	BW,C, G	6*3=18
D6	R5	300*300	.tiff	32	6	BW,C, G	6*3=18
		Total NWND and NWD images					96

(b) WND dataset and WD dataset (D7 to D12)

The second part of these datasets is comprised of watermarked and degraded images. Dataset D7 was created by using image-based watermark for six non-watermarked images as shown in Figure 3.4, namely, *Lena, Baboon, Cameraman, Meeting, Girl with black hair and Girl with Blonde hair* using the DWT watermarking algorithm. MPS degradation on watermarked images was applied by using the three different scanning mode for each round of print and scan, up to five rounds. All scan-and-print operations were conducted by using the printer described in Section 3.8.2.

Dataset D8, created by degrading dataset D7, six watermarked images are obtained by printing first and then scanning in the three different scanning modes; this represents one round of degradation from MPS. In Round 1, it results in 18 watermarked and degraded images. For creating dataset D9, the images of dataset D8 were printed and then scanned with following settings: (a) Images of D8 scanned in BW mode were again scanned with BW mode; (b) Images from D8 were also scanned again with colour mode; (c) D8 images were scanned again with the greyscale mode. For Round 2, that results in 18 degraded images to construct dataset D9; Similar procedures were followed for the third, fourth and fifth round of MPS, each round results in 18 degraded images to create datasets D10, D11 and D12 for Round 3, Round 4 and Round 5 of degradation. In total, 96 images were present in the second part of the dataset (i.e., watermarked degraded datasets) as shown in Table 3.3 (i.e., six non-watermarked or original images, 90 degraded and watermarked images by using multiple scanning modes, i.e., BW, colour and greyscale, totally 30 for each).

Table 3.3 Settings for degrading the watermarked images after MPS



Data set	MPS Round number	Resoluti on	Exte nsio n	Bit depth	Numb er of images	Scanning mode (×3)	Total number of images
D7	0	300*300	.tiff	32	6	-	6
D8	R1	300*300	.tiff	32	6	BW,C, G	6*3=18
D9	R2	300*300	.tiff	32	6	BW,C, G	6*3=18
D10	R3	300*300	.tiff	32	6	BW,C, G	6*3=18
D11	R4	300*300	.tiff	32	6	BW,C, G	6*3=18
D12	R5	300*300	.tiff	32	6	BW,C, G	6*3=18
		Total WND and WD images					96

d. Near duplicate image datasets for BIIGA

This dataset was used for testing the BIIGA approach. Six standard images were used, namely, *Baboon, Cameraman, Meeting, Girl with black hair, Lena and Girl with*

Blonde hair having 300×300 resolutions, 32-bit depth as original or non-watermarked images. A total of 8 datasets were prepared which were further divided into the two parts: the first part consists of NWD images and NWND images (i.e., G1 to G4); the second part was WD images and WND images (i.e., G5 to G8).

(a) NWD image dataset (G1 to G4)

These datasets were created by using Section 3.8.4 (C) image set at part (a). Datasets D1 to D6 images were arranged in four categories: original images, black and white NWD images, colour NWD images and greyscale NWD images. NWND category images were grouped under dataset G1 having six images. Black and white NWD images were assembled as dataset G2, having 30 degraded images (i.e., six from each round, five rounds in total). Color NWD images were collected as dataset G3, having 30 degraded images. Greyscale NWD images were gathered as dataset G4, having 30 degraded images. In total, 96 images, 90 were NWD images and six were non- NWND images.

(b) WND and WD image datasets (G5 to G8)

These datasets were created by using Section 3.8.4 (C) part (a) images. Datasets D7 to D8 were arranged in four categories: WND images, black and white WD images, colour WD images and greyscale WD images. WND category images were grouped under Dataset G5 having six images: black and white WD images as Dataset G6, having 30 degraded images (i.e., six from each round, totally five rounds). Color WD images were collected as Dataset G7, having 30 degraded images. Greyscale WD images were gathered as Dataset G8, having 30 degraded images. In total, 96 images, out of which 90 were degraded (WD) images, six were WND images.

3.8.4 WUtils.com & Tomeko.net Web Tools

The next generation starts using DNA for storing digital data because it is considered as the most dense and stable media (George, Church, & Kosuri, 2012). In this thesis, image analysis was performed by using biological sequences (i.e., DNA). The main problem was how an image could be encoded in DNA. Naidu and Narayanan developed an approach for identifying polymorphic malware variants (Naidu & Narayanan, 2014; Naidu & Narayanan, 2016) by using biological representation and bioinformatics sequence for alignment. The Naidu and Narayanan approach did not address the issue: "how can we convert an image to biological DNA?", without encoding the image biologically, the subsequent analysis could not be performed. That is, encoding the image biologically is beyond the scope of their work.

The proposed biological representation of images was inspired from Naidu and Narayanan's biological representation for viruses (Naidu & Narayanan, 2016). Steps required for encoding images in DNA were added, Naidu and Narayanan approach (Naidu & Narayanan, 2016) was followed. In their first step of converting the virus to biological sequences (i.e., DNA), they need to transform the virus into a hexadecimal. That indicates a hex dump of the images is a must to follow their approach. To the best of our knowledge, there is no direct method reported in the literature to extract a hex dump from an image.

For converting an image to its hex dump, multiple ways were investigated. It was found that if an image was converted to a base64 byte array, then it was easy to convert the base64 byte array to a hexadecimal. For converting an image to the base64 array, we used a *free online web utility WUtils.com* that converted the image to the base64 byte array. Then, the base64 byte array was converted to the hexadecimal by using another online tool developed by Tomasz Ostrowski. Now, the extracted hexadecimal dump of an image could be employed for converting it to DNA by using Naidu and Narayanan's approach. In the literature, there is a lack of evidence showing the use of bioinformatics sequence alignment in the analysis of images. After getting the images DNA sequences, sequence alignment tools were required which are explained in the next subsection.

3.8.5 JAligner

After representing images as DNA, the role of bioinformatics sequence alignment becomes very crucial for automatic signature extraction. Sequence alignment acts as the heart of bioinformatics for identifying the similarities between two biologically-represented image sequences or any other biological sequence. Two main methods are used for sequence alignment global (i.e. alignment from start to end of the sequence) and local (i.e. find local regions having high similarity) alignment. The Needleman-Wunch algorithm (NWA) was used for global sequence alignment (Needleman & Wunsch, 1970) and the Smith-waterman algorithm (SWA) for pairwise local alignment (Smith & Waterman, 1981). For overcoming the processing speed problem, Gotoh proposed (Gotoh, 1982) an improved version of SW and NW algorithms that we used in our approach.

Moustafa developed an open source tool for sequence alignment known as JAligner. This tool was written in Java and used the affine gap penalty model, having Gotoh's upgradation for local pairwise sequence alignment. In this thesis, JAligner was used for local pairwise alignment, an in-house modified code of JAligner for global alignment used the Needleman-Wunch (NW) algorithm with Gotoh's improvement. Areas of similarity between two biological sequences are recognised by applying sequence alignment algorithms. That will help to identify functional, structural and evolutionary relationships. If the number of biological sequences is three or more having the same length for understanding homology and the evolutionary relationship, multiple tools of sequence alignment are needed that are explained in the next subsection.

3.8.6 MAFT

Multiple sequence alignment (MSA) is a process of aligning three or more biological sequences of identical length to hypothesize homology and an evolutionary relationship (Weizhong, et al., 2015). There are different tools for MSA like Clustal omega (Sievers, et al., 2011), Kalign (Lassmann, Frings, & Sonnhammer, 2009), MAFT (Katoh & Standley, 2013), T-Coffee (Notredame, Higgins, & Heringa, 2000), MUSCLE (Edgar, 2004), MView (Brown, Leroy, & Sander, 1998) and WebPRANK (Loytynoja & Goldman, 2010), etc. Each tool is good for certain sets of conditions: Clustal omega and MAFT are suitable for medium to large alignments, Kalign for large, MUSCLE for medium and T-Coffee is suitable for small alignments. MView is used to transform the results of sequence similarity search into a MSA. WebPRANK has a phylogeny-aware MSA program. The lengths of biologically-encoded DNA sequences of images fall into medium or large sequences that motivated us to use MAFT for MSA in this thesis. A literature survey about MAFT automatically gives other reasons for choosing it, those are explained in the next paragraph.

MAFT has a long history of development and evolution. In 2002, this tool was launched for rapid MSA based on Fourier transform (Katoh, Misawa, Kuma, & Miyata, 2002). For improving the accuracy of MSA in 2005, they released MAFT version 5.3 (Katoh, Kuma, Toh, & Miyata, 2005). To overcome the time-consuming process for phylogenetic tree development, they developed a PartTree algorithm for MAFT in 2006 (Katoh & Toh, 2007). In 2008, they upgraded MAFT to version 6, where they added the PartTree algorithm to improve scalability and a four-way consistency objective function to improve the accuracy of ncRNA alignment (Katoh & Toh, 2008; Katoh & Toh, 2008). Multiple alignments of DNA sequences were introduced into MAFT in 2009 (Katoh, Asimenos, & Toh, 2009). For further improvement in MSA, in 2010, parallelization techniques were integrated into MAFT (Katoh & Toh, 2010). A sequence adding method was implemented into MAFT in 2012 (Katoh & Frith, 2012), a new web server, aLeaves, that provides an on-demand exploration of metazoan gene family trees with enhanced interactivity on MAFT (Kuraku, Zmasek, Nishimura, & Katoh, 2013). In 2013, MAFT version 7 was released for improving performance and usability by adding features like adjustment of the direction of nucleotides, constrained alignment, adding unaligned sequence into an existing alignment, and parallel processing (Katoh & Standley, 2013). Sometimes MAFT aligned unrelated segments (i.e., over alignment); in 2016, they added a feature to put down over alignment (Katoh & Standley, 2016) and chained the guide trees for enhancing the MSA's accuracy within the structurally-conserved regions (Yamada, Tomii, & Katoh, 2016). In this thesis, the biological sequences are large, and all the literature about MAFT motivated us to use it in our research project. That is why MAFT version 7 was used throughout our research work for MSA. After getting local, global and multiple aligned sequences, we needed tools to perform pattern matching and pattern recognition that are explained for BIIIA and BIIGA.

3.8.7 MEGA7

For grouping species and establishing evolutionary relationships or phylogeny between species, we need a particular tool. There are many tools available nowadays for phylogenetic tree reconstructions like T-Rex (Boc, Diallo, & Makarenkov, 2012), phyloT, MEGA (Kumar, Stecher, & Tamura, 2016), etc. MEGA is one of the oldest and most robust tools with progressive development. The full form of MEGA is Molecular Evolutionary Genetics Analysis software that was introduced in 1994 for approximating evolutionary distances, calculating basic statistical values and reconstructing phylogenetic trees from molecular data (Kumar, Tamura, & Nei, 1994). MEGA 2 was released in 2001 with added features like large dataset analysis, creating groups of sequence, specifying domains and genes, expanding the collection of statistical methods and visual representation of input data and output results on the Microsoft Windows platform (Kumar, Tamura, Jakobsen, & Nei, 2001).

In 2004, automatic and manual sequence alignment, evolutionary distance estimation, phylogenetic tree inference, mining of web-based dataset and evolutionary hypothesis test were added to MEGA3 (Kumar, Tamura, & Nei, 2004). MEGA4 released with Maximum Composite Likelihood for calculating evolutionary distance between all sequences pairs concurrently and creating captions in 2007 (Tamura, Dudley, Nei, & Kumar, 2007). In 2011, Maximum Likelihood analysis for evolutionary trees, hypothesizing sequences and ancestral states with probability, best-fit substitution model selection (DNA), and evolutionary rate calculation were added in MEGA5 (Tamura, et al., 2011). A stand-alone executable of MEGA software was released with the name MEGA-CC (i.e., MEGA-Computational Core) that had all the functionalities of MEGA in 2012 (Kumar S. , Stecher, Peterson, & Tamura, 2012). Timetree Wizard was included in MEGA6 for timetree inference to describe phylogeny and calibration constraints, with more advanced memory management and enhanced algorithms in 2013 (Tamura, Stecher, Peterson, Filipski, & Kumar, 2013). The 64 bit MEGA7 was made available in two interfaces: command line and GUI (Kumar, Stecher, & Tamura, 2016) for processing larger datasets in 2016. The developmental history of MEGA inspired us to use this as a tool for BIIGA and phylogenetic tree. In this thesis, we employed MEGA7 for our research work.

3.8.8 Clamscan

From aligned sequences, common substrings were extracted as a pattern or meta-signature for a specific category of images. Categories of images mean that images have a particular type of watermarks or are degraded by print and scan or any other category that is under the scope of this thesis in Section 3.8.3. Clamscan is an open source software that is used for pattern matching in viruses. Common substring patterns or signatures were matched by using Clamscan for images identification and BIIIA development.

3.9 Summary

This chapter introduces a scientific research methodology. It is an ideal approach to emphasise on the research procedure and compose the research by planning and characterizing a research issue, making inferences that reflect the present reality. The research methodology recognised in this chapter makes a vital supposition: our contribution of motivation from bioinformatics tools and techniques for image analysis by using pattern matching.

In Chapter 2, little information and knowledge regarding bioinformatics-inspired image analysis is available in the literature. The task of this thesis is to consider the aspects of evaluation of watermarked and non-watermarked images, their degraded copies, their identification that are well interpreted and related to our area of research. After that, we can connect them to bioinformatics tools and techniques during the research process when it is most suitable.

In addition, we see how far we can get with research questions and current systems, make them more suitable for identifying and grouping the watermarked / non-watermarked and degraded images. We trust that we have gone as far as we can, with these traditional techniques in image analysis, following pattern matching for biologiacl images, perceiving how motivation from bioinformatics can lead to new bits of knowledge and methods for pattern matching in image analysis.

We, in this way, need to begin with what is at present, known about bioinformatics-inspired image analysis which leads to the experience of our first issue. As appeared in the literature review, in spite of all the work so far in bioinformatics-inspired image analysis, there is very little work addressing what bioinformatics-inspired image analysis is and how we can use it for degraded image analysis. In addition, none of the work addressed the issue of evaluation of image degradation for watermarked/non-watermarked images.

The next chapter explores image degradation for watermarked/non-watermarked images; the experimental results will answer the Research Question 1.

Chapter 4

Evaluations of Image Degradation from MPS

The aim of this chapter is to explore the first research question established in Chapter 3: "Is it possible to investigate and measure degradation of non-watermarked / watermarked images from MPS; if so, what are the suitable metrics?"In Section 4.1, we explain the issues in the existing evaluation approaches. Section 4.2 includes the metrics for measuring the degradation of watermarked/non-watermarked images with the hypothesis. In Section 4.3, a novel model is presented for resolving the problem of degradation evaluation. The calculated values of different metrics are plotted graphically and analysed in Section 4.4. *The performance of the metrics for the evaluation is discussed in Section 4.5. Lastly, Section 4.6 contains a summary of this chapter.*

4.1 Introduction

The area with regard to degradation of non-watermarked images from MPS has a deal of unanswered problems. This chapter will explain an approach to evaluate the degradation in non-watermarked / watermarked images by using MPS that can be used as a real-time application. For performing this, it is important to know about image quality metrics for image degradation evaluation. If a standard and well-established image quality metrics can be demonstrated to work with the evaluation of non-watermarked / watermarked and degraded images through MPS that will be considered as the proofs that our evaluation approach may be correct. In this chapter, our previous eight metrics of image quality evaluation for MPS degradation were used i.e., bias, correlation coefficient (cc), root mean square error (rmse), root average spectral error (rase), ergas, universal image quality index (uiqi), structural similarity metrics (ssim), and structural dissimilarity metric (dssim) (Garhwal & Yan, 2015).

Those existing evaluation approaches for watermarked images ignore the degradations due to MPS. The research discussed in this thesis investigates image degradation from a digitization viewpoint into consideration. MPS operations are iteratively committed to a hardcopy or a softcopy from the same source, i.e., taking a hardcopy of the original document or photo, the document is scanned and converted to a softcopy, then the softcopy is printed; the printout is treated as the printed hardcopy. The operation is iteratively in use of the same photocopy machine for many rounds. However, to the best of our knowledge to the date, (Garhwal & Yan, 2015) there is no other work reported to conduct such a comprehensive evaluation of image degradation from MPS.

4.2 Image Quality Metrics

The perceptual similarity between watermarked/non-watermark and original / degraded images after print and scan is calculated by quantifying using image quality metrics which were tested and verified for evaluating degradation of non-watermarked images in our previous work (Garhwal & Yan, 2015). These eight metrics are shown in a set of v in equation (4.2.1).

$$v = (Bias, CC, RMSE, RASE, ERGAS, UIQI, SSIM, DSSIM) \qquad (4.2.1)$$

These eight were the objective metrics explained in detail in Section 2.4.2 with the relevant literature and formulae of the metrics, they will not be repeated here.

Our motive here is to investigate and analyze again these eight objective metrics for non-watermarked images that are different from our previous publication (Garhwal & Yan, 2015) except one image, Lena, for checking and verifying that the metrics values with different groups of images work well and their watermarked versions (i.e., watermarked images). The suitability of metrics is tested by considering three aspects: objective reference method, robustness, consistency for further evaluation of image distortion caused by MPS. There are two more criteria for further verification of the metrics: practicability and solvability. This is explained below:

a. **Practicability**: Is our evaluation relevant and applicable to MPS scenarios?

b. **Solvability**: Does the proposed approach succeed in finding a valid metrics for evaluating degradation of non-watermarked images from MPS?

Further, the hypothesis test for evaluating images degradation is stated below:

Hypothesis:

Null hypothesis. The selected eight image metrics are suitable for measuring the image degradation from MPS.

Alternative hypothesis. None of the image quality metrics is appropriate for measuring the image degradation from MPS.

For executing the experiments, a novel method is proposed that will be described in the next section.

4.3 A Novel Method for Evaluating Images from MPS

For performing experiments, each research method needs datasets, test conditions and a proper method. More details about these three items are provided in the next paragraphs.

A. Dataset

To the best of our knowledge, there are no publicly available image datasets for evaluating image degradation from, so we have to create our own dataset. The preparation of a dataset used for evaluation of image degradation was explained in Section 3.8.3.

B. Test Conditions

The test conditions for the criteria of image selection are based on the print and scan, which depends on the selected image resolution in pixels per inch, namely, dpi; the parameters also are used for scanning and bit-depth (R. Villán, 2006). Some additional criteria are also employed as the requirements of this research project. More details about test conditions are given below in terms of variables for the evalaution.

Variable Definitions

a. *Dependent variable.* In the proposed research the image quality metrics used for quantification of image degradation like Universal Image Quality Index (Q) etc. are the dependent variables. In this thesis, the dependent variable was fixed to eight metrics on the basis of our previous publication (Garhwal & Yan, 2015). Image degradation also acts as the dependent variable.
b. *Independent variable.* This variable defines the yardstick on which dependent variable depends, i.e., the number of rounds of print and scan. This independent variable value is fixed to five rounds after MPS, the degradation is very prominent and no use for analysing further rounds of MPS.
c. *Control variable:* The number of print-and-scan rounds will work as the control variable because it is the process that will control the quantity and quality of image degradation. Other variables are the resolutions of watermark and image that are used for watermark embedding. Bit depth, image type and compression of the image are also used as control variables. All these control variables are explained below after the normalisation of all images of the dataset to 512 × 512 pixels.
    - *Resolution.* How many pixels, an image has, is represented by the concept resolution. A higher resolution for an image shows more information content. For the purse of experimentation, the resolution is measured in dots per inch (dpi). The resolution that we were using for images were horizontal 300 dpi and vertical 300 dpi.
    - *Compression of Image.* Provided by the standard images for testing, ITU-T provides a dataset of mages for testing with compression. In this thesis,

we have used the image format of .tiff extension, which retains the maximum information of an image after LZW compression.

- *Image Type:* Greyscale image.
- *Bit Depth:* 32
- *Numbers of print and Scan:* The round number of MPS was fixed to five for the independent variables.

Other criteria, like content, file format and metadata in word documents, etc., are considered for future work in this thesis.

d. Confounding Variable. External influences like fluctuation of voltage in the scanners or printers, any types of mechanical problems, material quality, etc. are considered as the potential confounding variables.

Calculations. The eight image metrics are adapted to the evaluations of watermarked / non-watermarked and degraded images by using the image processing toolbox of the famous platform MATLAB.

C. Method

For accomplishing this thesis evaluation, a novel method is proposed as shown in Figure 4.1 where a watermarked/non-watermarked image is an input image that acts as an original image on which the print-and-scan operations are applied to obtain a degraded image. For that degraded image, all eight-image quality metrics were calculated using Matlab. After the first round of print-and-scan operations, the output degraded image will be acted as an input of test image, namely R1, the degraded image for the second round of print-and-scan image degradation, and again all of the eight-metrics are calculated for the degraded image after the second round of print and scan.

This process is repeated in a similar manner till five round of print-and-scan process and then stopped because the degradationis so prominent that it can be seen clearly by using our naked eyes. At the end, we will get five different values of each metric for a test image. Now, these values can be checked and analyzed by plotting the graphs. Whether these values show the degradation or not will be investigated in the next section.

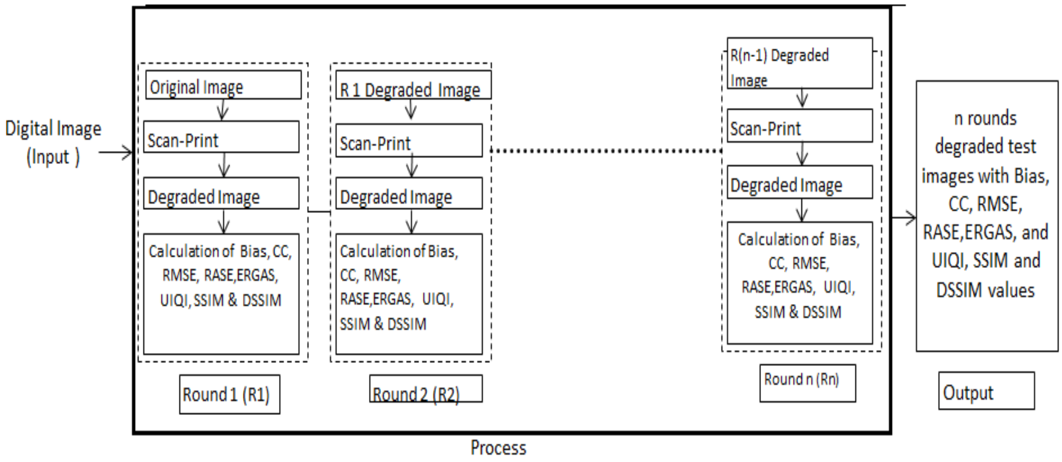

Digital Image (Input)
Original Image
Scan-Print
Degraded Image
Calculation of Bias, CC, RMSE, RASE,ERGAS, UIQI, SSIM & DSSIM
Round 1 (R1)
R 1 Degraded Image
Scan-Print
Degraded Image
Calculation of Bias, CC, RMSE, RASE,ERGAS, UIQI, SSIM & DSSIM
Round 2 (R2)
R(n-1) Degraded Image
Scan-Print
Degraded Image
Calculation of Bias, CC, RMSE, RASE,ERGAS, UIQI, SSIM & DSSIM
Round n (Rn)
Process
n rounds degraded test images with Bias, CC, RMSE, RASE,ERGAS, and UIQI, SSIM and DSSIM values
Output

Figure 4.1 A method for evaluating degradation from MPS

4.4 Results

In the literature, no other work has been reported to evaluate image degradation (Garhwal & Yan, 2015) using fully-referenced methods of image quality assessment on all the available datasets. The dataset, we created in Section 3.8.3, was used to conduct experiments and the study of evaluating degraded/non–degraded images after MPS. Then, the verified eight metrics from our previous publication (Garhwal & Yan, 2015) were employed to this dataset so as to calculate metrics for degraded image from MPS. After that, the results were analysed to ascertain their validity. We analysed the image degradation from MPA by using the differences between the original and the degraded images. These are explained in more detail for each of the eight metrics with graphical representation and comparison.

Bias

In our publication, bias was successfully verified to be used as one of the metrics for evaluating the MPS degraded NWD image (Garhwal & Yan, 2015). Bias values for the WD images by MPS were calculated by using the formula that was discussed in Section 2.4.2. Unnecessary information is added and necessary information is removed from an image during MPS process; this leads to the degraded images having mean greater than the original. This indicates that the bias for all the cases is negative.

The results are depicted graphically in Figure 4.2 (a) bias of NWD, (b) bias WD, and (c) average bias of NWD / WD. Bias shows its consistency in its value, which proves its suitability for measuring the degradation in the watermarked image by using MPS. Another validation was done by comparing the average bias of non-watermarked / watermarked and degraded images by using MPS. The trends delineate that the bias-based quality assessment is apt for measuring the WD images since the image differences are all convergent to the same mean with minor variations as shown in Figure 4.2 (c).

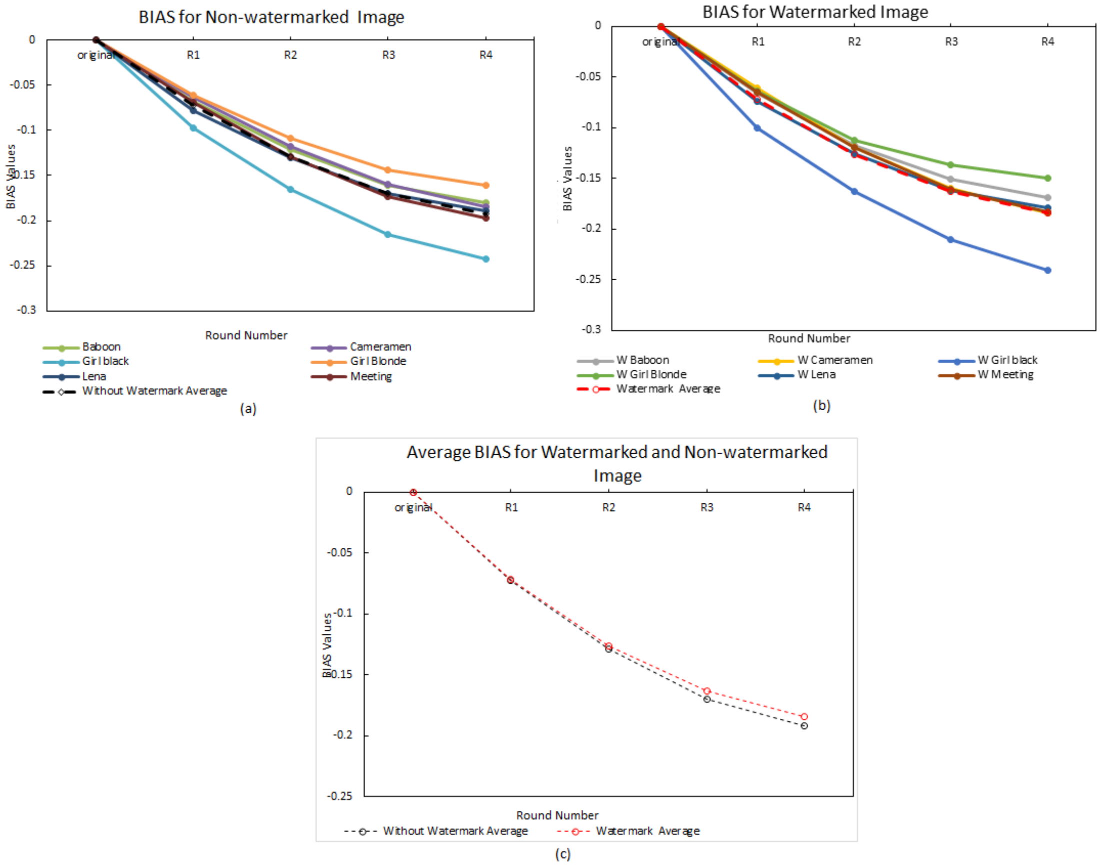


Figure 4.2 Image metrics Bias: (a) non-watermarked image (b) watermarked image (c) average for watermarked / non-watermarked image

Correlation Coefficient (CC)

After the confirmation of correlation coefficient (CC) as one of the metrics for evaluating the degraded and non-watermarked image after MPS (Garhwal & Yan, 2015); CC for the WD images was calculated; CC average values for WD and NWD were compared for validity. The calculation was done by using Matlab and the results were plotted in Figure 4.3 (a) CC of NWD, (b) CC of WD, and (c) average CC of NWD /WD images. Figures 4.3 (a) and (b) illustrates that there is continuous dwindling in the use of the metrics CC. The results is supported by taking advantage of the metrics CC; all the patterns of image differences between the original and its degraded images from MPS are approaching to the same mean as shown in Figure 4.3

(c). This shows that CC metrics are constantly convergent for all the five groups of degraded images with the merits of robustness and consistency.

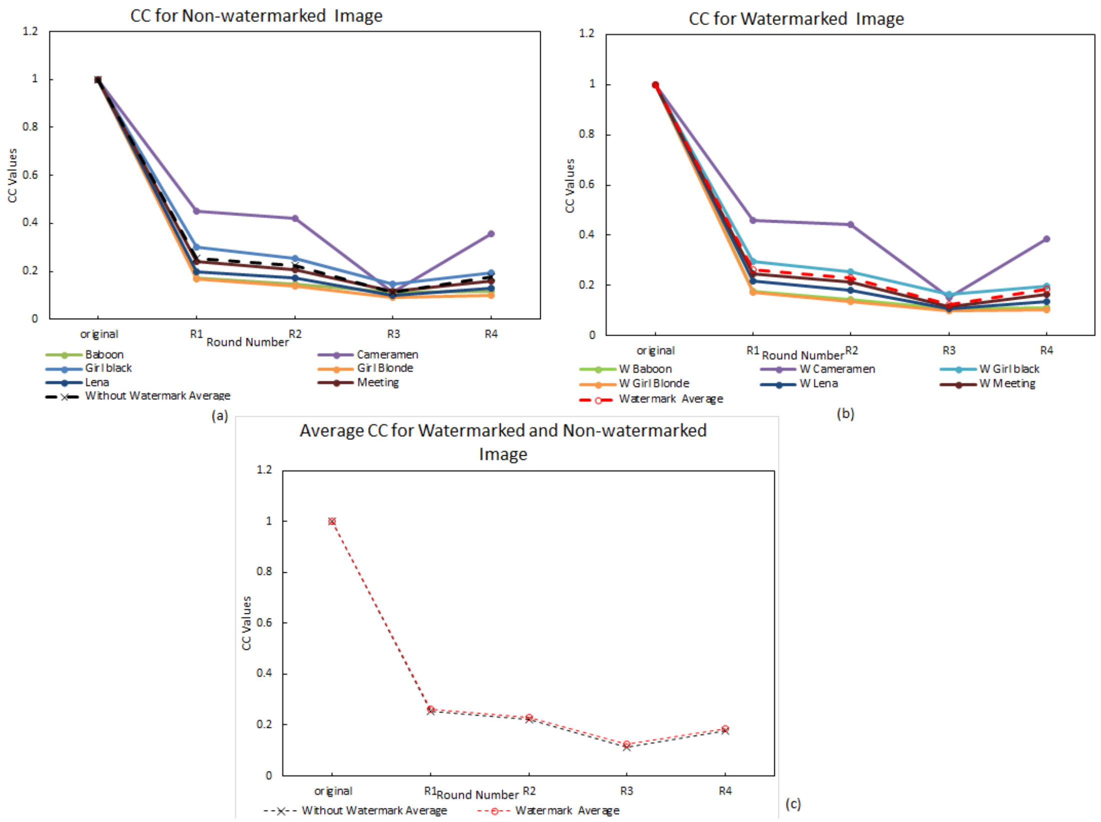


Figure 4.3 Image quality metrics CC: (a) non-watermarked image (b) watermarked image (c) average for watermarked or non-watermarked image

RMSE

From Figures 4.4 (a) and (b), we see that the metrics *RMSE* were also able to judge the differences between a test image and its degraded ones for all groups of non-watermarked / watermarked images. The patterns were convergent to the same mean except the image *Cameramen* was a bit of a diversion from the expectation as shown in Figures 4.4 (a) and (b). A comparison of the RMSE average value of the WD and NWD images is shown in Figure 4.4 (c) which shows a consistent increase and a minor difference between them. This analysis further confirms our previous work (Garhwal & Yan, 2015) that we can use it as a metric to evaluate degraded images generated from MPS.

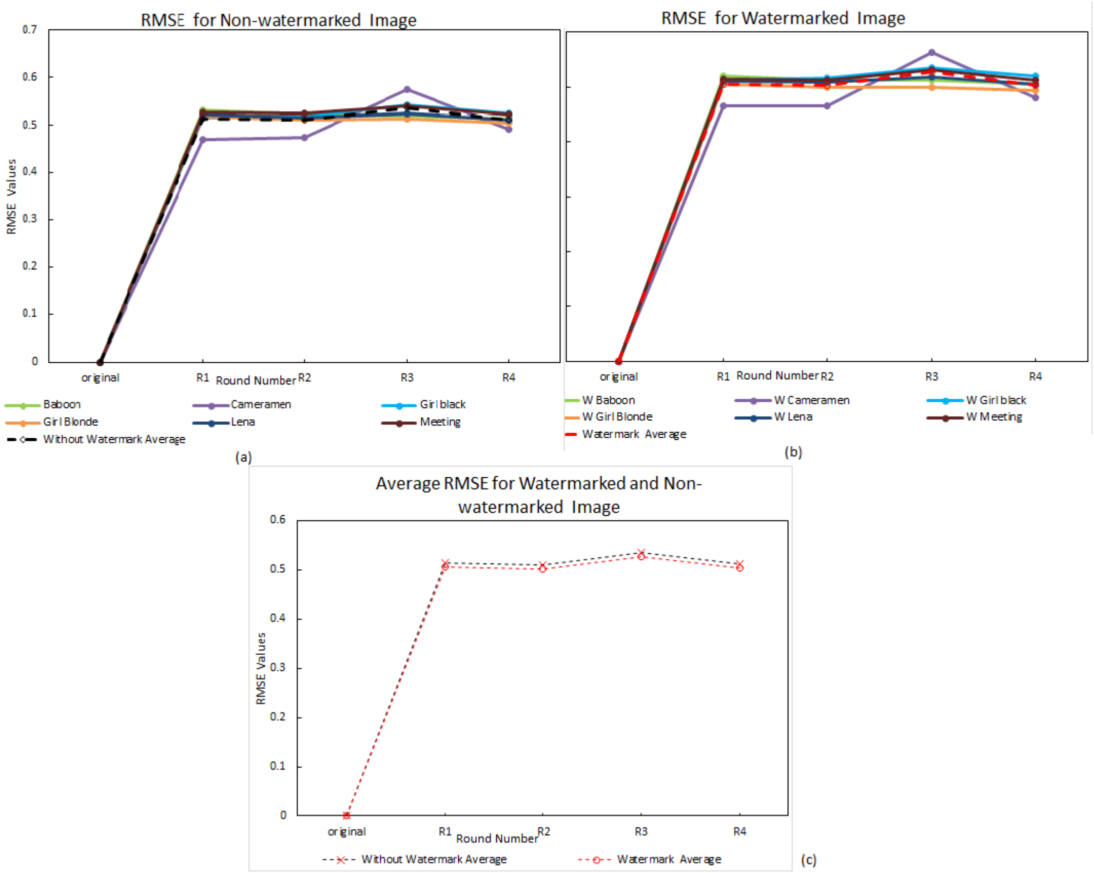


Figure 4.4 Image quality metrics RMSE: (a) non-watermarked image (b) watermarked image (c) average for watermarked / non-watermarked image

RASE

Figures 4.5 (a) and (b) were used to show that the metric *RASE* also could evaluate the image degradation by using NWD/WD images. All the data led to the point that the image metric *RASE* correctly reflects the differences between an original and its degraded images from five rounds of print and scan. Analysis of the RASE average values of WD and NWD images in Figure 4.5 (c) shows a consistent increase with almost no difference which supports that RASE is a suitable metric for evaluating WD images which is in coherence with our previous work (Garhwal & Yan, 2015).

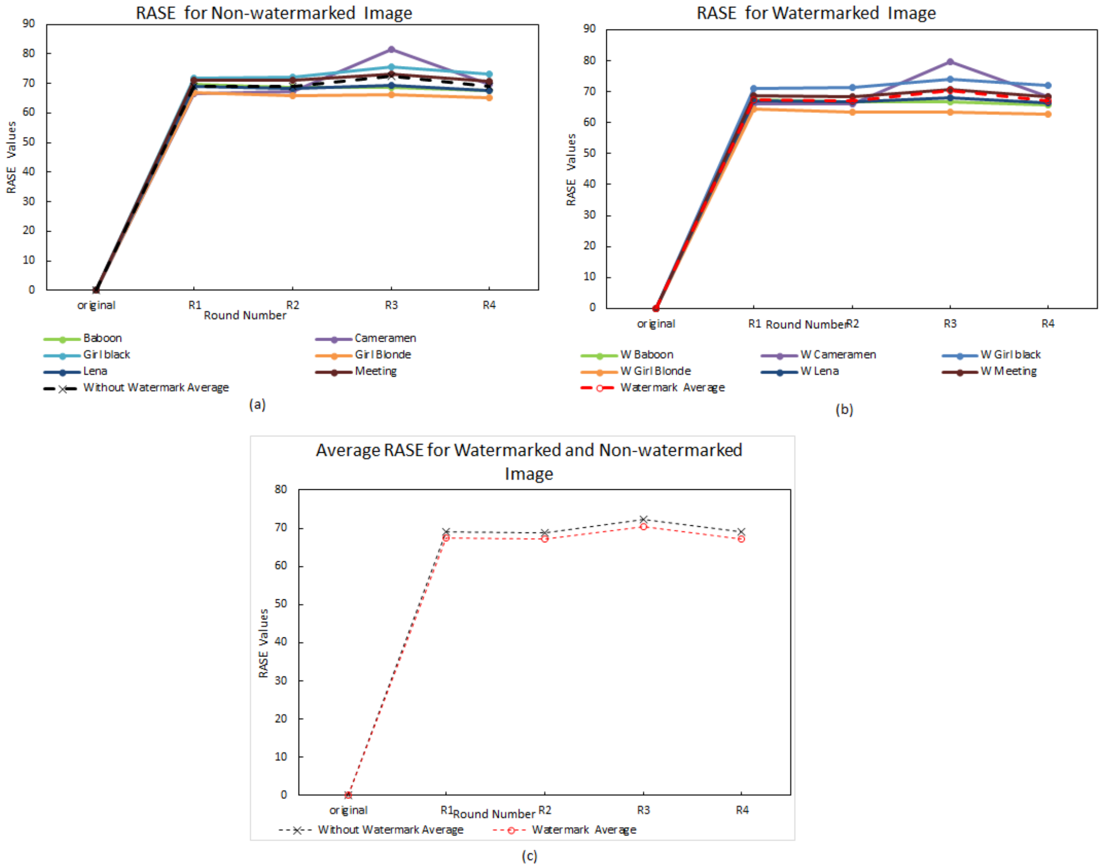


Figure 4.5 Image quality metrics RASE: (a) non-watermarked image (b) watermarked image (c) average for watermarked / non-watermarked image

ERGAS

Figures 4.6 (a) and (b) reveal the differences between a reference image and its five degraded images by using the metric ERGAS. The figure discovers that ERGAS satisfies the criteria as a good metric: consistency and robustness. The polylines in the figure converge to the same mean from the very beginning to the end. Figure 4.6 (c) compares the ERGAS average values for WD/NWD images which were consistently increasing and there is an ignorable difference between them that proves the suitability of ERGAS as a degradation evaluation metric for the WD / NWD images from MPS.

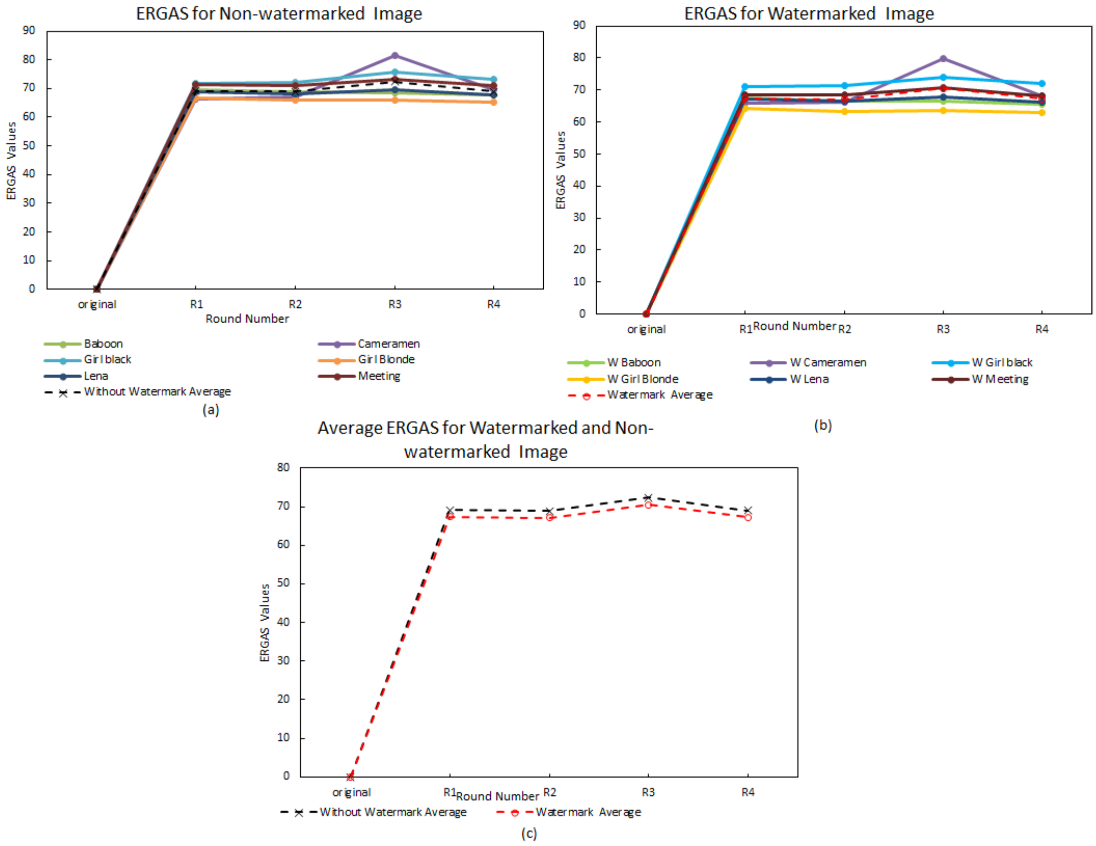
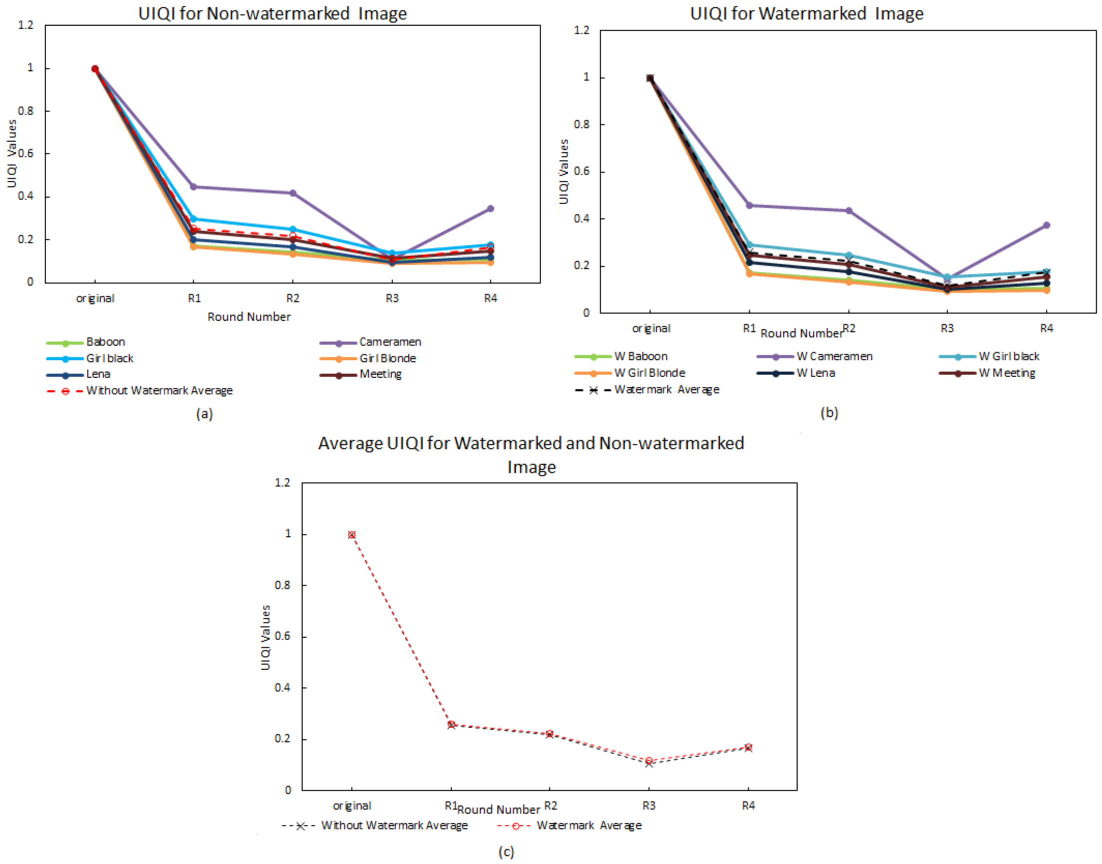


Figure 4.6 Image quality metrics ERGAS: (a) non-watermarked image (b) watermarked image (c) average for watermarked / non-watermarked image

UIQI

In Figures 4.7 (a) and (b), the metrics *UIQI* are employed to quantify the degradation from MPS for non-watermarked / watermarked images. The data points on the figure stand for a uniform pattern in decline. It is very interesting to see that the comparisons of average values of the UIQI in Figure 4.7 (c) are excellent to reflect the essence of the degradation from MPS for watermarked images.



(a)

(b)

(c)

Figure 4.7 Image quality metrics UIQI: (a) non-watermarked image (b) watermarked image (c) average for watermarked / non-watermarked image

SSIM

Figures 4.8 (a) and (b) reveal that the metric SSIM was consistently decreasing for all test WD / NWD images. SSIM values are between 1 and -1; from Figures, 4.8 (a), (b) and (c), we see that all values lie in between 1 and -1. Additionally, we validate SSIM by comparing the consistent decrease in the average value of NWD/WD images as shown in Figure 4.8 (c). We would like to say that SSIM is a suitable metric for MPS image degradation evaluation for WD images.

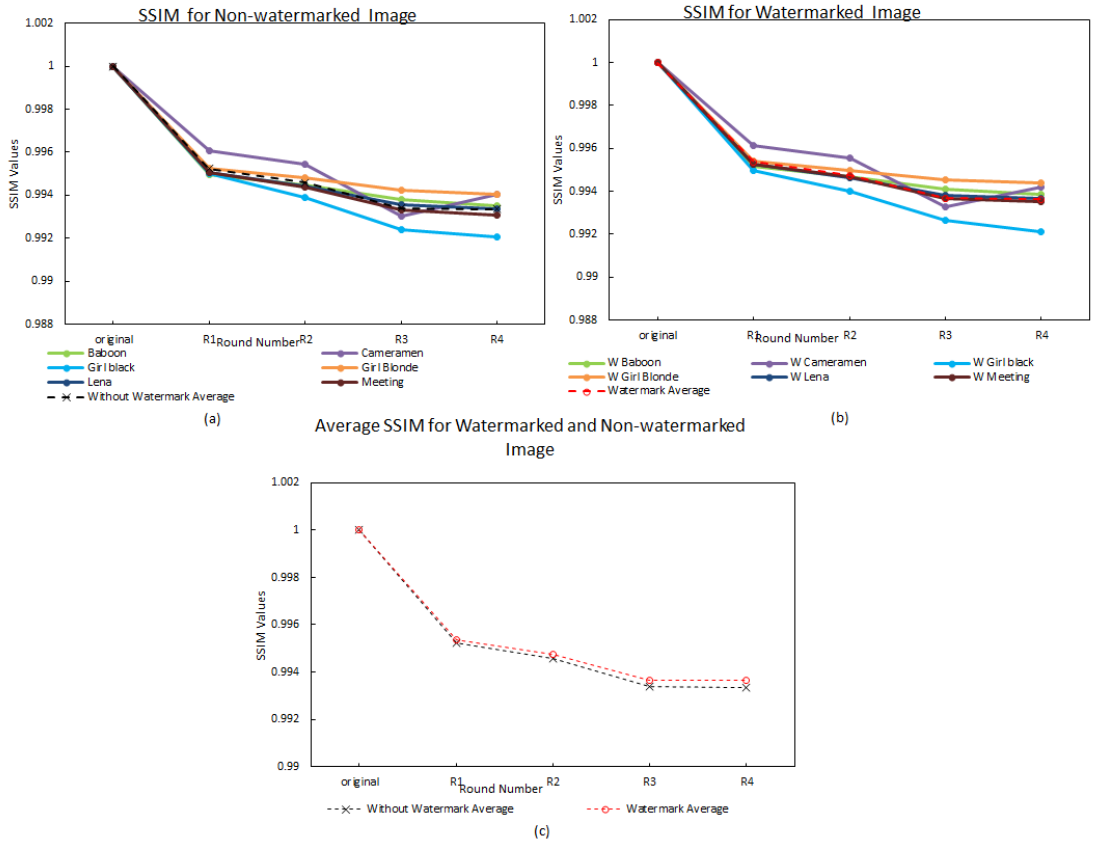
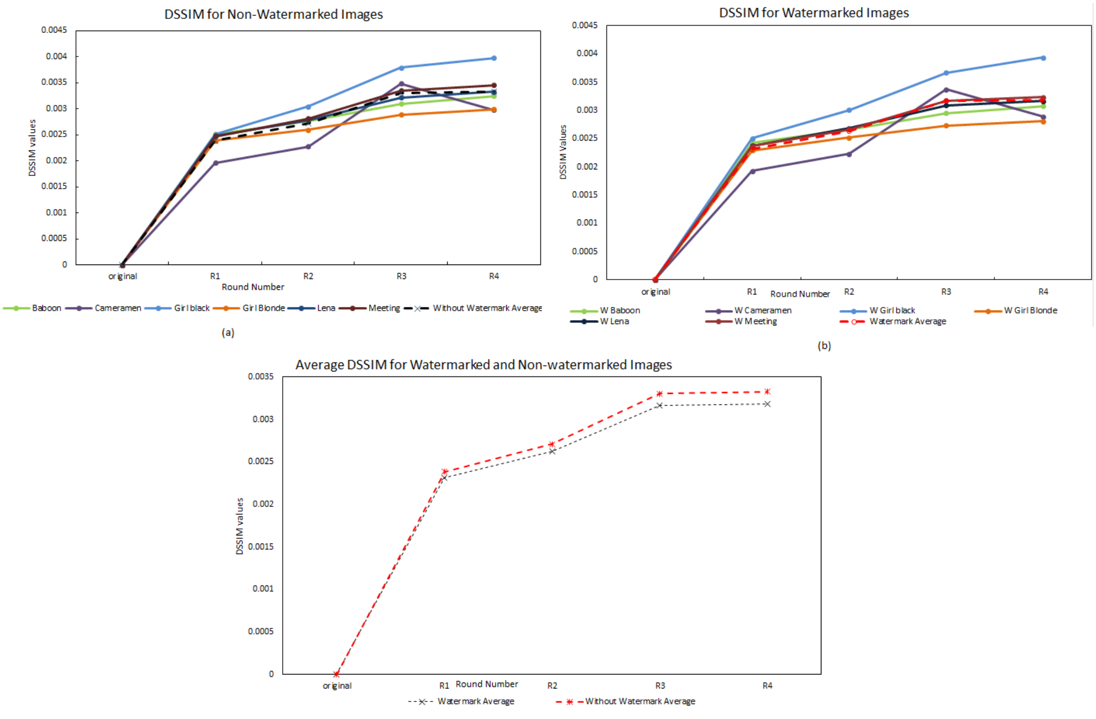


Figure 4.8 Image quality metrics SSIM: (a) non-watermarked image (b) watermarked image (c) average for watermarked / non-watermarked image

DSSIM

Figures 4.9 (a) and (b) show that the metric DSSIM is used to find degradation from MPS for NWD/WD images. The values of DSSIM show consistency in increasing for all test NWD/WD images which indicate that it also satisfies the criteria of a superior metric for measuring degradation. Comparison of DSSIM average values for WD/ NWD images is shown in Figure 4.9 (c) which consistently decrease with little difference. It verifies that DSSIM can be used as one of the metrics for evaluating WD images degradation by MPS.



(a)

(b)

(c)

Figure 4.9 Image quality metrics DSSIM: (a) non-watermarked image (b) watermarked image (c) average for watermarked / non-watermarked image

Summing up, the experiment results unveiled that all eight metrics have very good outcomes and completely qualify for being applied to evaluate degraded images after MPS.

4.5 Discussion

In this chapter, image degradation was taken for inspection. Evaluation of MPS degraded images have been done by using the eight objective metrics as our previous work (Garhwal & Yan, 2015). The results disclose that all eight metrics are suitable for calculating the differences between an input image and its degraded copies from MPS. That answers Research Question 1 with yes, it is possible to investigate and measure image degradation using the eight image quality metrics: bias, CC, RMSE, RASE, ERGAS, UIQI, SSIM, and DSSIM. Additionally, the hypothesis proposed was proven correct after results analysis. i.e., the selected eight metricsi of mage quality were found suitable for measuring the image degradation from MPS. Furthermore, this evaluation approach is practical and solvable indicating that this evaluation approach is relevant to MPS degradation as well as successful in finding the solution to the issue of evaluation of the degraded images after MPS.

This evaluation also provides a numerical reason for the identification and grouping of the degraded images. That will motivate us for further analysing the degraded images that are performed in the next chapter by developing the BIIIA and BIIGA algorithm for the degraded images after MPS.

4.6 Summary

Image degradation is a fundamental problem in digital image processing since it has occurred in print and scan very often. The multiple rounds of print and scan cause image degradation that leads to problems in image identification. For inspecting the image degradation, eight fully-referenced objective metrics were employed on degraded images generated from MPS. The results verified that all eight metrics outperformed and were found suitable for measuring image degradation. These numerical values of image degradation provide a strong theoretical background for further research in the area of identification and grouping of the images by using suitable approaches.

In the next chapter a bioinformatics-inspired approach will be proposed for the degraded images, we call it as BIIIA algorithm.

Chapter 5

BIIIA: Bioinformatics Inspired Image Identification Approach

This chapter explores a new approach, i.e., BIIIA, to identify watermarked images from non-watermarked images, by utilizing a sequence alignment technique in bioinformatics to align variable lengths of two extracted DNA from watermarked / non-watermarked and degraded / non-degraded images after MPS. Section 5.1*, briefs the introduction and background of the image identification. A hypothesis and BIIIA method overview are provided in Section* 5.2 *and Section* 5.3*, respectively. In Section* 5.4*, we explain the steps employed in the BIIIA method. Experiments and results of the BIIIA method are explained in Section* 5.5*. A discussion about the result is presented in Section* 5.6*. Finally, in Section* 5.7*, a summary of the chapter is given.*

5.1 Introduction and Background

Previous work, proposed by Naidu and Narayanan in syntactic signature extraction, inspired us using bioinformatics techniques, like sequence alignment techniques, to extract signatures (common sub-string) for identifying polymorphic malware variants. Furthermore, they extended their work by using different substitution matrices (Naidu & Narayanan, 2016); after that, Needleman-Wunch (NW) and Smith-Waterman (SW) algorithms (Naidu & Narayanan, 2016) for identifying polymorphic malware variants. That motivated us to test whether bioinformatics tools can be used for image analysis. For using bioinformatics tools to image analysis, the images must be encoded biologically (i.e., in DNA). In this thesis, the used images refer to the watermarked / non-watermarked images.

Watermarking can be tracked from 1954 (Cox & Miller., 2002) for securing the ownership of data. In the last few years, there has been a big revolution of using media / non-media watermarks for securing the digital contents like documents, images, etc. Watermarking will secure the digital contents by putting some owner-related information in the form of text, audio, video or image. Different watermarking algorithms based on frequency transformation, e.g., Discrete Cosine Transformation (DCT), Discrete Wavelet Transformation (DWT), Discrete Fourier Transformation (DFT), DCT & DWT combined, etc. were used to watermarking. For watermarking algorithms, the main issue is to extract the embedded watermark with the highest similarity; the process is very specific to other traditional embedding algorithms. Modern watermarking algorithms are evolving because of non-media (i.e. software, relational data, natural language text, sensor streams, streaming data, etc.) watermarking. This indicates the challenges to detect watermarks in big data; for different variants of watermarks (i.e., media / non-media watermarks) (Panah, Schyndel, Sellis, & Bertino, 2016), until they are noticeable in the cloud. The automatic identification of watermarked images using watermark embedding algorithms and other watermark signatures in watermark identification remains a relatively unexplored area of counterfeiting information.

A watermark may be media (i.e. image, audio, video, etc.) / non-media data (i.e. spatial data, sequence data, spatiotemporal data) (Panah, Schyndel, Sellis, & Bertino, 2016). Watermark embedding algorithms for audio (Patent No. IPN WO 89/08915, 1989), images (Schyndel, Tirkel, & Osborne, 1994; Caronni, 1995; Brassil, Low, Maxemchuk, & O'Gorman, 1994), video (Matsui & Tanaka, 1994) and watermarking schemes based on SVD and DNA (Wu & Kan, 2015) are reported in literature, having a distinct watermarking (i.e. audio, image or video watermark etc.) and extracting algorithm framework as shown in Figure 5.1. This indicates that watermarked images were used as the input and particular watermark extraction algorithms for extracting the watermark. The traditional method of watermark extraction has the watermark as output, but lacks of automatic signature extraction based on a syntactic approach for watermark identification by using biological image representation and sequence alignment.

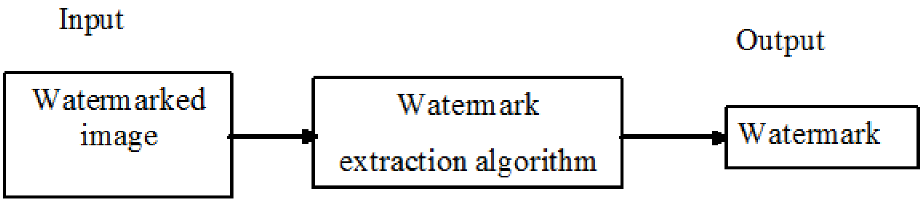


Figure 5.1 Watermark extraction framework

Signature extraction based on syntactic techniques for watermarks is not investigated in comparison to traditional watermark extraction techniques. The relevant literature is almost absent or limited. The historical reason behind that is a variety of watermarks that can be embedded in different multimedia. Logically, specific watermark extraction algorithms will work for a specific case. For extracting commonalities, semantic analysis is the only way for efficient watermark signature generating.

In the last decade, big data watermarking increased rapidly (Panah, Schyndel, Sellis, & Bertino, 2016) and emphasised on the need for watermark identification techniques with the Boolean answer: yes or no, which reflects the presence or absence of watermarks in media / non-media data. We are trying to solve that problem by using a biological representation of the image and bioinformatics alignment techniques as shown in Figure 5.2.

The work reported in Chapter 2 about the use of DNA for image representation will focus on how we can store and retrieve the error-free data by using DNA for the biological image encoding and decoding; but, lacking the use of image to DNA encoding for identifying the watermarked / non-watermarked images. This literature gap of previous research motivated us to check whether DNA image encoding would work for watermarked/non-watermarked images.

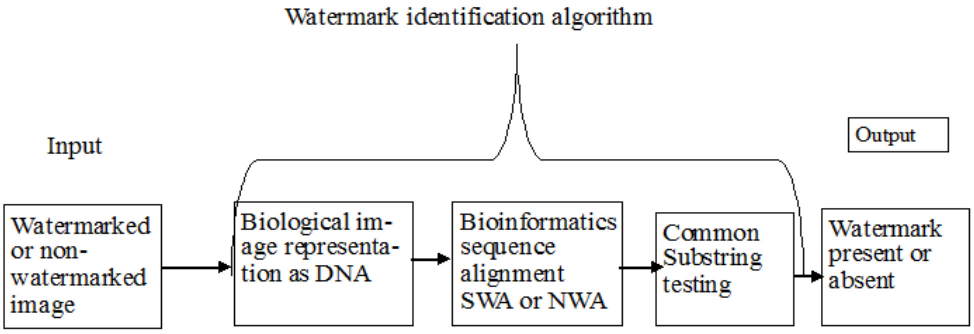

Watermark identification algorithm
Input
Output
Watermarked or non-watermarked image
Biological image representation as DNA
Bioinformatics sequence alignment SWA or NWA
Common Substring testing
Watermark present or absent

Figure 5.2 Watermark identification framework

After, representing the W or NW image as DNA, bioinformatics tools and techniques can be employed. The technique we were using for this research work is sequence alignment; the role of bioinformatics sequence alignment is crucial for automatic signature extraction. Sequence alignment acts as the heart of bioinformatics for identifying the similarity between two biologically-represented image sequences or any other biological sequence. Two main methods were explained in Chapter 2, i.e., the Needleman-Wunch (NW) algorithm for global sequence alignment (Needleman & Wunsch, 1970) and Smith-waterman (SW) for pairwise local alignment (Smith & Waterman, 1981).

As discussed in Chapter 2 for applying sequence alignment to the images, the first features of the images were extracted and then these features were encoded biologically. After that, the sequence alignment was applied to further experimental work; for instance, 2D shape matching (Kim, Chang, Lee, & Lee, 2010; Kim, Chang, Liu, Lee, & Lee, 2009), etc. To the best of our knowledge, none of the approaches directly encoded the whole image into DNA sequence without extracting any features of the image; then, bioinformatics sequence alignment algorithm was applied to image identification. This resulted in our interest continuing to grow for a method to develop an automatic signature extraction for watermarked/non-watermarked images by using bioinformatics tools and techniques.

5.2 Hypothesis Test

The research hypothesis for identifying watermarked/non-watermarked and degraded images is that it is possible to identify syntactic patterns using bioinformatics tools and techniques that help to determine whether a degraded / non-degraded image contains a type of watermark or not; that helps us to identify the images of their expected category. If this research hypothesis does not apply to image identification, it is highly unlikely that syntactic structures extracted by using bioinformatics tools will be used for image identification.

5.3 BIIIA Method: An Overview

In any dataset, test conditions and a method are required for successful execution of the experiments.

Dataset

In this study, we focus on identifying images by using string-based automatic signature extraction techniques. However, at this stage, we were not clear what the implications would be used for automatic signature extraction from watermarked / non-watermarked images. This rationale limited us to three different watermarks (i.e., image, text and shape-based watermarks) for watermarking from a potentially large number of watermarked images.

Four image datasets were used, i.e., I1, I2, I3, and I4, where I1 is comprised of non-watermarked 444 images, I2, I3, and I4 are watermarked datasets, each has 444 images, i.e., I2 is a text watermark (TW), I3 is an image watermark (IW) and I4 is a shape-based watermark (SW). More details about the dataset preparation are given in Section 3.8.3.

Novel Method BIIIA and Test environment conditions

The BIIIA method consists of three main steps: biologically-based encoding of images and local & global alignment as well as pattern matching as shown in Figure 5.3. The test environment conditions along with the steps of BIIIA method are explained in the next few paragraphs.

Biology-Based Encoding of the Image

Four letters of DNA A, T, C and G were used for encodings the image in biology to find out its suitability for watermarked/non-watermarked image identification.

Local and Global Sequence Alignment

The sequence alignment algorithms SWA and NWA were employed on biologically -encoded images in the DNA sequence. Consequently, the biologically-encoded images for sequence alignment needed a scoring matrix. To find the optimal alignment of the SWA and NWA algorithms was used (i.e., substitution strategy and gap scoring). Different scoring matrices are available like point accepted mutation (PAM), blocks substitution matrix (BLOSUM), and IDENTITY (ID) that are used by SWA and NWA. For our experiments, we used ID to achieve exact matching. The identity (ID) scoring matrix dispenses the most parsimonious method for finding a relation between one symbol and another symbol in the string, no assumptions are made. That motivated us to employ the ID matrix for the experiments rather than well-known biological scoring matrices such as PAM (Point Accepted Mutation and BLOSUM (Block Substitution Matrix) (United states Patent No. 20070244652, 2007). Additionally, we used a simple match-mismatch scoring matrix for more analysis.

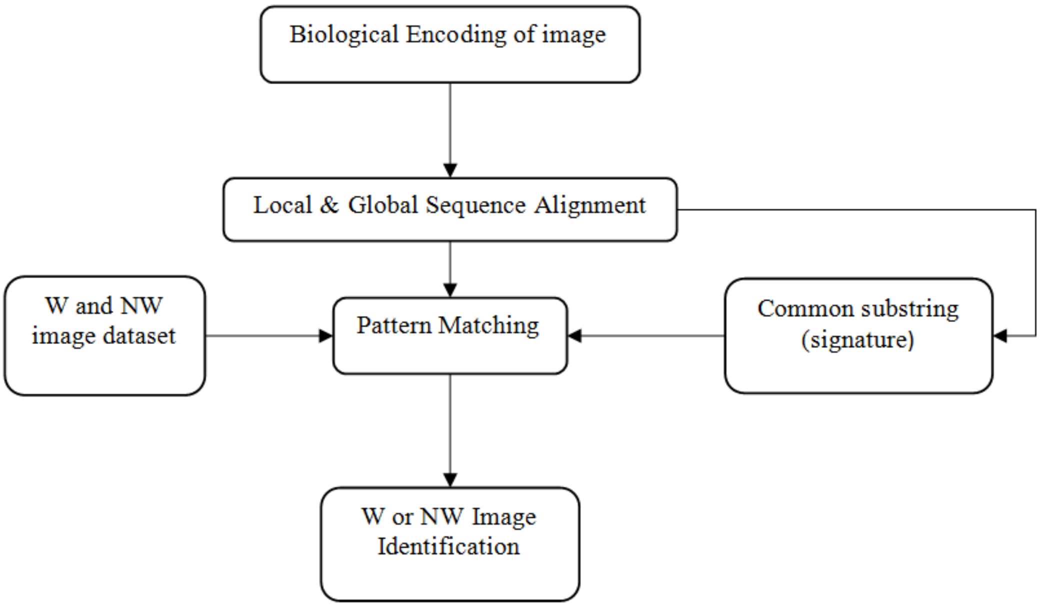

Biological Encoding of image
Local & Global Sequence Alignment
W and NW image dataset
Pattern Matching
Common substring (signature)
W or NW Image Identification

Figure 5.3 BIIIA method: an overview

For generating scoring matrices, we had to assign gap penalties for every insertion or deletion of any characters during the alignment process. Initially, NWA had no control over the penalty assigned to definite gap length. For increasing the computational speed, subsequently a linear cost function is assigned for some insertion (deletion) remainder (Sankoff, 1972; Sellers, 1974). So, for every indel (insertion or deletion), a penalty is assigned. However, in the meantime, if an indel is penalised in such a way that it becomes substantially less than the mismatch, this results in quite expensive longer gaps. For solving this problem, the gap penalty function has a gap open penalty that charges for every gap initiated and a gap extension penalty for penalising the length for every indel. The resulting affine linear function for gap length is $f(i)=c+d\times i$, where the gap penalty function is $f(i)$ having gap length i, a penalty c for gap initiation, and d is the smaller penalty for gap extension. A further breakthrough was made by Gotoh (Gotoh, 1982), where optimal alignment has affine linear gap penalties, the computation time is proportional to the multiplication of the length of the two sequences to be aligned. For our research work, we used NWA with Gotoh and SWA with Gotoh to speed up the computations.

The next was the gap open and gap extended penalties: fixed gap open and gap extended penalties were used for all combinations to determine whether we could identify watermarked/non-watermarked images correctly and efficiently.

Pattern Matching

In this thesis, pattern matching was performed by using Clamscan between common sub strings (signatures), extracted after sequence alignment of two biologically-encoded images, tested against the watermarked/non-watermarked image datasets to check whether it identifies the expected category. That indicates, if the common substring belongs to the IW images, then it will only identify the IW images and not any other group of images.

The aim of this thesis is to examine whether Naidu and Narayanan's syntactic approach (Naidu & Narayanan, 2016) can be used for identifying watermarked / non-watermarked images. The next section will describe in detail how biologically encoded images into a DNA sequence was performed, how sequence-aligned algorithms with Identity (ID) and match-mismatch (MM) scoring matrices were employed.

5.4 BIIIA Method: Steps

For finalising the proposed methodology prototype, different combinations of biologically-encoded images and sequence alignment algorithms with scoring matrices were tested as shown in Figure 5.4 and summarised in the six steps as follows:

Step 1. Watermarking original images (I1) with text (I2), image (I3) and shape –based watermark (I4), resizing and selecting two test images at random from each dataset**.**

Step 2. Explain how we convert images into an acceptable form of sequence alignment.

Step 3. Explain the DNA-representation method.

Step 4. Explain pairwise local & global alignment of DNA.

Step 5. Explain the procedure of common substring (signature) extraction from aligned DNA.

Step 6. Deal with the procedure of testing the signatures by converting back to hexadecimal (hex) code. Each of the six steps is explained in the following text.

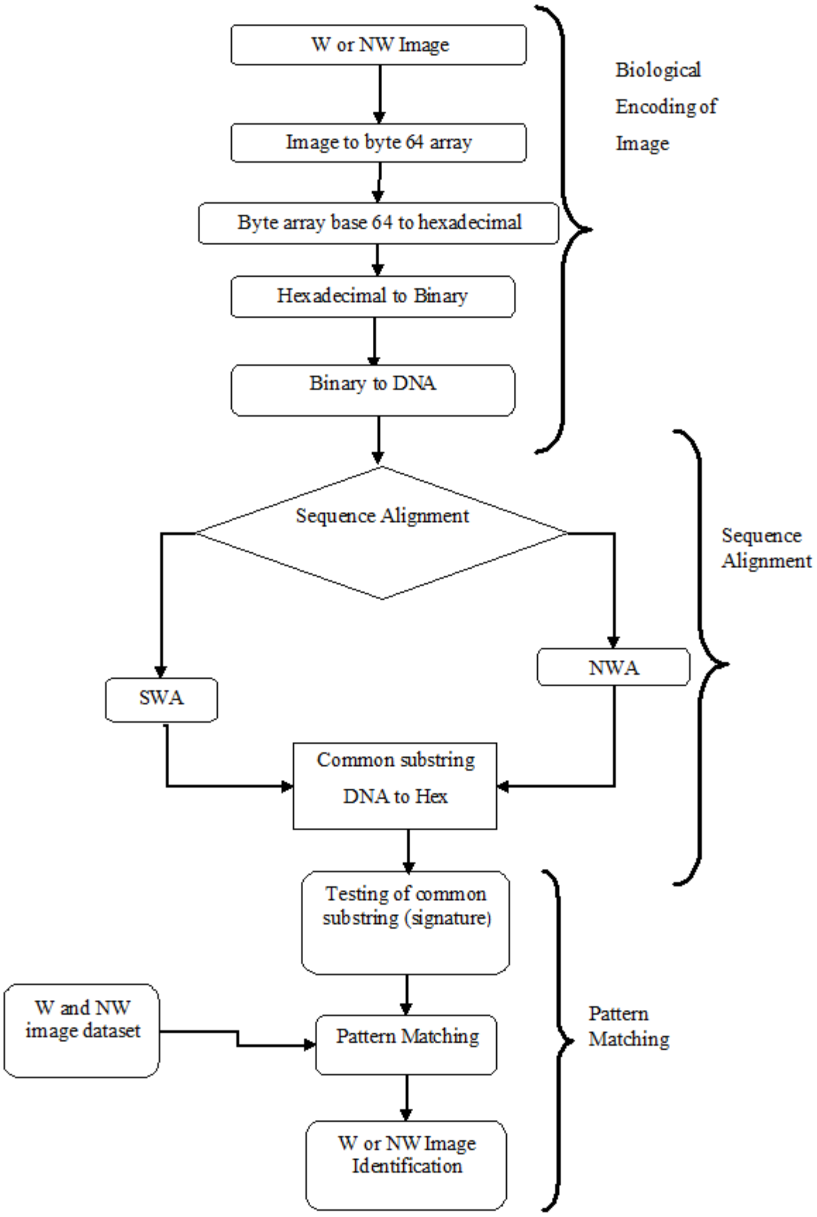


Figure 5.4 Initial BIIIA method

A. *Biology-based encoding of image into DNA:*

Step 1. Digital image. The preparation methods for Datasets I1, I2, I3 and I4 were explained in Section 3.8.3, this is not repeated here. Randomly, any of the two smallest size watermarked/non-watermarked images were selected from any of the datasets I1, I2, I3 or I4 because the available conversion tools could not handle bigger size images for converting an image to the required format (i.e., base64 byte array, hexadecimal etc.). That acts as input for the BIIIA system.

Step 2. Hexadecimal conversion. In this step, the selected watermarked/non-wateramrked images from Step 1 were converted into base64 byte array and then to hexadecimal. This step is an additional step to the Naidu and Narayanan approach (Naidu & Narayanan, 2016), to the best of our knowledge, there is no method reported in the literature to extract the hex dump directly from an image.

- Conversion of selected test images to the byte array (i.e., base64). Selected test images are converted into a byte array by using the website WUtils.com which provides facility to convert an image to base64.
- Conversion of the byte array to hexadecimal. Base64 values of the image were converted to the hexadecimal by using tomeko.net website. A small example is presented in Table 5.1 for representing an image to a byte array to the hexadecimal.

Table 5.1 Example of converting an image to a Hex code

Image	Byte 64 array	Hexadecimal
	SUkqAAgAAAASAP4ABAAB AAAAAAAAAAAABBAABAA AAAAEAAAEBBAABAAAA AAEAAAIBAwAEAAAA5gA AAAMBAwABAAAABQAAA AYBAwABAAAAAgAAABE …………………………………… ……………………. .ETC.	49492a00080000001200fe 00040001000000000000000 00010400010000000000100 00010104000100000000001 00000201030004000000e6 …………………………… ………………………….. .ETC.

Step 3. *Hexadecimal to DNA.* In this step, the extracted hexadecimal from the images in Step 2, converted to the binary code and then into the DNA sequences. The sub-steps for this conversion are explained below:

- Conversion of the hexadecimal to binary codes. The hexadecimals are converted into the binary codes by using in-house developed macros in EmEditor. The rules for transforming the hexadecimal into the binary code are followed from the Naidu and Narayanan approach (Naidu & Narayanan, 2016) that are: ‘0’ → ‘0000’; ‘1’ → ‘0001’; ‘2’ → ‘0010’; ‘3’ → ‘0011’; ‘4’ → ‘0100’; ‘5’ → ‘0101’; ‘6’ → ‘0110’; ‘7’ → ‘0111’; ‘8’ → ‘1000’; ‘9’ → ‘1001’; ‘a’ → ‘1010’; ‘b’ → ‘1011’; ‘c’ → ‘1100’; ‘d’ → ‘1101’; ‘e’ → ‘1110’; and ‘f” → ‘1111’ .
- Conversion of binary codes to DNA. Binary codes are converted into DNA by using in-house developed macros. Rules for conversion are followed

from Naidu and Narayanan approach (Naidu & Narayanan, 2016), that are: ‘00’ → ‘A’; ‘11’ → ‘T’; ‘10’ → ‘G’; and ‘01’ → ‘C’. A small example how to convert a hexadecimal to 64-bit binary code to DNA

01252012 (8 hexadecimal characters)

00000001001001010010000000010010 (32- bit binary code)

AAACAGCCAGAAACAG (16 DNA character)

B. Sequence Alignment

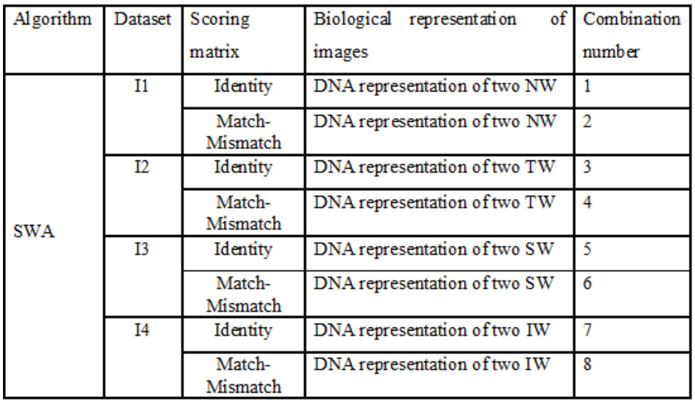
SWA and NWA algorithms were used for string matching regular alphabet between two DNA sequences of biologically encoded image from Step 3. The two DNA sequences extracted from the two test images (watermarked / non-watermarked) were used as an input to JAligner for sequence alignment. These string-matching algorithms recognise one or more positions in one string, where left strings known as patterns, are found. Suppose ‘∑’ be an alphabet (a character), i.e., a finite set. Traditionally, patterns and searched strings are vectors of part of ‘∑’. The ‘∑’ could possibly be a regular character or alphabet, i.e., for example, the Latin format A to Z letters. Some algorithms perhaps use binary codes (∑ = {0, 1}) in bioinformatics (∑ = {A, T, C, G}). SWA and NWA use dynamic programming alignment. Dynamic programming alignment is a more quantitative approach where scores are assigned for matches and mismatches (scoring matrices), instead of applying dots (Needleman & Wunsch, 1970; Waterman, Smith, & Beyer, 1976; Smith & Waterman, 1981). The results are called as ‘alignments’ as any one of them, or both. It may be changed by inserting gaps to get optimal patterns. For comparing with another scoring approach, we use a simple match-mismatch (MM) scoring matrix where we assign a score 2 if there is a match and for mismatch we assign -1, with a fixed gap open equal to 10 and the gap extend value is 1. The highest scores in scoring matrices indicate accurate alignment.

Step 4. In this step, global and local alignment were performed using the following combination:

a. Pairwise local alignment (SWA). SWA tries to find the most matched substrings between the pattern and search string, i.e., instead of looking in the completed sequence, the SWA selects the parts of all realisable length, then matches and improves the resemblance rate.

In total, eight pairwise local alignments were performed by using the eight different combinations, made up from four datasets, SWA with an ID and match-mismatch (scores for match=2, mismatch=-1, and gap open=10 and gap extension=1) as shown in Table 5.2 where NW represents non-watermarked images, TW for text-based watermark, IW with regard to image-based watermark and SW stands for shape-based watermark.

Table 5.2 Combinations of pairwise sequence alignment

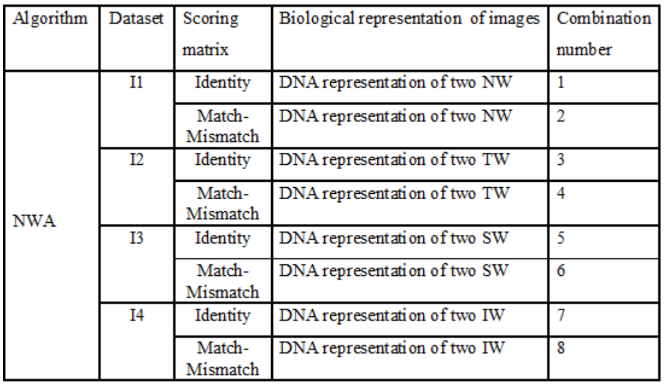


Algorithm	Dataset	Scoring matrix	Biological representation of images	Combination number
SWA	I1	Identity	DNA representation of two NW	1
		Match-Mismatch	DNA representation of two NW	2
	I2	Identity	DNA representation of two TW	3
		Match-Mismatch	DNA representation of two TW	4
	I3	Identity	DNA representation of two SW	5
		Match-Mismatch	DNA representation of two SW	6
	I4	Identity	DNA representation of two IW	7
		Match-Mismatch	DNA representation of two IW	8

The four datasets are I1, I2, I3, and I4, where I1 is the original image (non-watermarked) dataset, I2 to I4 are its watermarked variants. For these combinations, sixteen test images were encoded biologically as DNA sequences, two from each dataset I1, I2, I3 and I4. This indicates that in total, we had 16 DNA encoded images for pairwise local alignment. For executing the experiments with the above settings, the JAligner programs were customized for two combinations: SWA with an ID matrix and SWA with an MM matrix.

b. *Global alignment.* NWA tries to get the best possible alignment by looking out the sequences from the beginning to the end of the sequence.

Table 5.3 Combinations for global sequence alignment



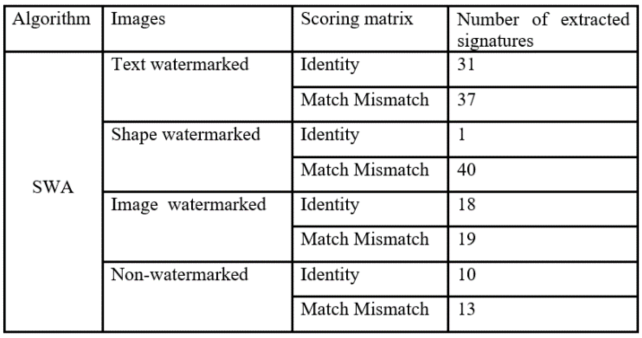
Algorithm	Dataset	Scoring matrix	Biological representation of images	Combination number
NWA	I1	Identity	DNA representation of two NW	1
		Match-Mismatch	DNA representation of two NW	2
	I2	Identity	DNA representation of two TW	3
		Match-Mismatch	DNA representation of two TW	4
	I3	Identity	DNA representation of two SW	5
		Match-Mismatch	DNA representation of two SW	6
	I4	Identity	DNA representation of two IW	7
		Match-Mismatch	DNA representation of two IW	8

In total, eight global alignments were performed by using the eight different combinations made up from four datasets: NWA with an ID and match-mismatch (scores for match=2, mismatch=-1, and gap open=10 and gap extension=1) as shown in Table 5.3, where NW represents non-watermarked images, TW refers to text-based watermark, IW stands for image-based watermark and SW means shape-based watermark. The four datasets were I1, I2, I3, and I4, where I1 is the original image (non-watermarked) dataset, I2 to I4 are its watermarked variants. For these combinations, sixteen test images were encoded biologically as DNA sequences, two from each dataset I1, I2, I3 and I4. It indicates that we had 16 biologically-encoded images in total for eight global alignment. For the purpose of execution, the JAligner programs were customized for two combinations: NWA with an ID matrix and NWA with an MM matrix.

C. *Pattern Matching by using Signature*

Step 5. After the local and global alignment process, common substrings were extracted from aligned sequences, we call it as signatures that were used for detecting a family of watermarked / non-watermarked variants of images. The method for the pattern matching was from the previously reported approach (Naidu & Narayanan, 2014; Naidu & Narayanan, 2016; Naidu & Narayanan, 2016).

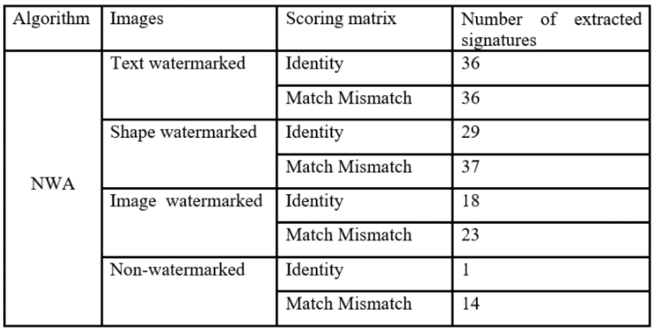
Table 5.4 The number of signatures using pairwise local sequence alignments with SWA



Algorithm	Images	Scoring matrix	Number of extracted signatures
SWA	Text watermarked	Identity	31
		Match Mismatch	37
	Shape watermarked	Identity	1
		Match Mismatch	40
	Image watermarked	Identity	18
		Match Mismatch	19
	Non-watermarked	Identity	10
		Match Mismatch	13

In Step 4, eight pairwise local alignments were employed to the eight combinations of Table 5.2, to extract signatures that were summarized in Table 5.4.

Table 5.5 The number of signatures by using global sequence alignments with NWA



Algorithm	Images	Scoring matrix	Number of extracted signatures
NWA	Text watermarked	Identity	36
		Match Mismatch	36
	Shape watermarked	Identity	29
		Match Mismatch	37
	Image watermarked	Identity	18
		Match Mismatch	23
	Non-watermarked	Identity	1
		Match Mismatch	14

The maximum number of signatures for shape-watermarked images is 40 with the match-mismatch scoring matrix. A minimum number of signatures is 1 for shape-watermarked images with identity scoring matrix.

The signatures, extracted after employing the eight global alignments to eight combinations of Table 5.3, were represented in Table 5.5. The maximum number of signatures for shape-watermarked images is 37 with the match-mismatch scoring matrix. The minimum number of signatures is 1 for the non-watermarked images with identity scoring matrix.

Step 6. In this final step, all signatures, obtained in **Step 5** in their DNA sequence representation, were converted back to hexadecimal format. These signatures were tested against the watermarked / non-watermarked variants using ClamAV (Clamscan antivirus scanner) software. For example, if the signature was obtained from the text-watermarked images, the text- watermarked image dataset will be considered.

5.5 Experimental Results

Two experiments were performed. The first experiment determined DNA encoding and sequence alignment algorithm (i.e. SWA and NWA) for BIIIA by using four (I1, I2, I3 and I4) image datasets. The second experiment is further used to verify the validity of the encoding and sequence alignment algorithm for BIIIA by testing it on the MPS degraded image datasets that have the different images and different watermarking schemes with different types of watermark.

Experiment 1. Selected test images from I1, I2, I3 and I4 datasets were encoded biologically into DNA sequence as the BIIIA method Step 1 to Step 3. After that, sequence alignment algorithms, the SWA and the NWA are applied to the DNA - encoded images to extract the common substring (signature) by employing BIIIA Step

4 and Step 5. Lastly, Step 6 was implemented to test whether the signature would identify the image datasets with their expected signatures. The results of Experiment 1 are discussed below.

Results of Experiment 1 will be discussed in two parts: the first part (Result 1) explains the sixteen combinations (i.e., eight for NWA alignment and eight for SWA alignment) with identity percentage, similarity percentage, gap percentage, alignment length and alignment score that will partially answer sub-question 3(b) by using conserved region analysis; the second part (result 2) will answer the sub-question 3(a), and will fully answer the sub-question 3 (b).

Result 1. Conserved regions analysis and most suitable sequence alignment algorithm Table 5.6 shows that the percentages of identities and similarities range from 5.79% (lowest) to 100% (highest). Lower values of identities and similarities reflect that lower percentages of DNA residues were conserved in the biologically-represented images and vice versa. In the case of the watermarked images aligned with SWA and ID matrix, the percentage of identities and similarities was 100%. These results reflect that by converting or representing images into biological representation, we can extract common substrings or subsequences. Moreover, from Table 5.6, the percentage of gaps ranges from 0% to 37.64%, which represents that the quantity of insertion and deletion varies from 0% to 37.64%. 0% gap percentage shows 100% match, and the matching percent decreased by an amount greater than the 0% gap percentage.

The work reported here followed the method adopted by Naidu and Narayanan (Naidu & Narayanan, 2016), i.e., a fixed combination of gap open (i.e. 10) and gap extended, i.e., 1 (we changed the gap extend from 0.5 to 1 penalty). We explored various combinations of SWA and NWA with ID and MM scoring matrices on watermarked image detection. From Table 5.6 and Table 5.7, an 85% and over similarity and identity percentage will reflect that higher amounts of conserved regions, which are conserved during the DNA representations of images. The 85% overall similarity and identity is found for the following combination of sequence alignment:

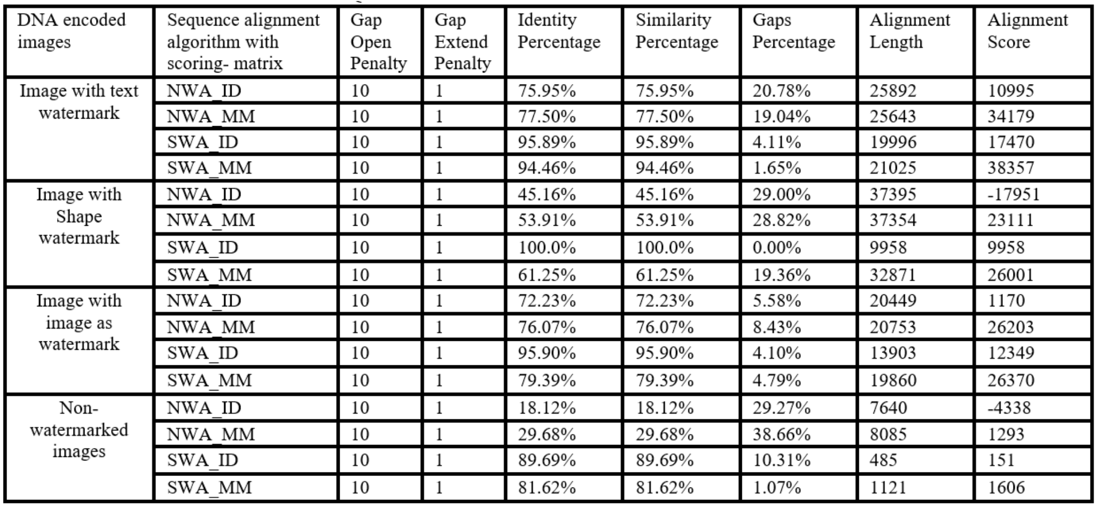
a. SWA with ID matrix (SWA_ID) and SWA with MM matrix (SWA_MM) for text-watermarked image representation in DNA representation.

b. SWA_ID for shape-watermarked DNA representation.

c. SWA_ID for image-watermarked DNA representation.

d. SWA_ID for non-watermarked DNA representation.

Table 5.6 The results of sequence alignments



DNA encoded images	Sequence alignment algorithm with scoring- matrix	Gap Open Penalty	Gap Extend Penalty	Identity Percentage	Similarity Percentage	Gaps Percentage	Alignment Length	Alignment Score
Image with text watermark	NWA_ID	10	1	75.95%	75.95%	20.78%	25892	10995
	NWA_MM	10	1	77.50%	77.50%	19.04%	25643	34179
	SWA_ID	10	1	95.89%	95.89%	4.11%	19996	17470
	SWA_MM	10	1	94.46%	94.46%	1.65%	21025	38357
Image with Shape watermark	NWA_ID	10	1	45.16%	45.16%	29.00%	37395	-17951
	NWA_MM	10	1	53.91%	53.91%	28.82%	37354	23111
	SWA_ID	10	1	100.0%	100.0%	0.00%	9958	9958
	SWA_MM	10	1	61.25%	61.25%	19.36%	32871	26001
Image with image as watermark	NWA_ID	10	1	72.23%	72.23%	5.58%	20449	1170
	NWA_MM	10	1	76.07%	76.07%	8.43%	20753	26203
	SWA_ID	10	1	95.90%	95.90%	4.10%	13903	12349
	SWA_MM	10	1	79.39%	79.39%	4.79%	19860	26370
Non-watermarked images	NWA_ID	10	1	18.12%	18.12%	29.27%	7640	-4338
	NWA_MM	10	1	29.68%	29.68%	38.66%	8085	1293
	SWA_ID	10	1	89.69%	89.69%	10.31%	485	151
	SWA_MM	10	1	81.62%	81.62%	1.07%	1121	1606

As stated earlier, identity and similarity percentages are more than 85%, which shows that conserved regions are more preserved. SWA_ID had always an 85% or over percetahe of similarity and identity for the DNA representations of watermarked / non-watermarked images. Because of that, from the above results, we conclude that SWA_ID is the best combination of sequence alignment for the two images represented, biologically.

Result 2. For checking the suitability of DNA-based encoding in biology

Table 5.7 provides the rates for detection of the three watermarked datasets (i.e., I2, I3, I4) and non-watermarked dataset (I1), employed in these experiments. Signatures from the text-watermarked images were tested against the text-watermarked images using Clamscan.

Table 5.7 shows a 100% detection rate for the individual signatures, except the SWA_ID of the image is 74.54%. This is less than the 100% detection rate but was resolved by employing multiple signatures together; we will obtain 100% detection rate.

Shape-watermarked image signatures were tested against the shape-watermarked image using Clamscan. From Table 5.7, except for the SWA_ID DNA representation (i.e., 0%), all combinations have a 100% detection rate. This 0% detection rate was not resolved because we got only one signature.

Image-watermarked images were tested by using Clamscan without any exception.

From Table 5.7, it is clear that the individual signatures had a 100% detection rate.

Table 5.7 The detection rates of watermarked/non-watermarked images identification by using Clamscan

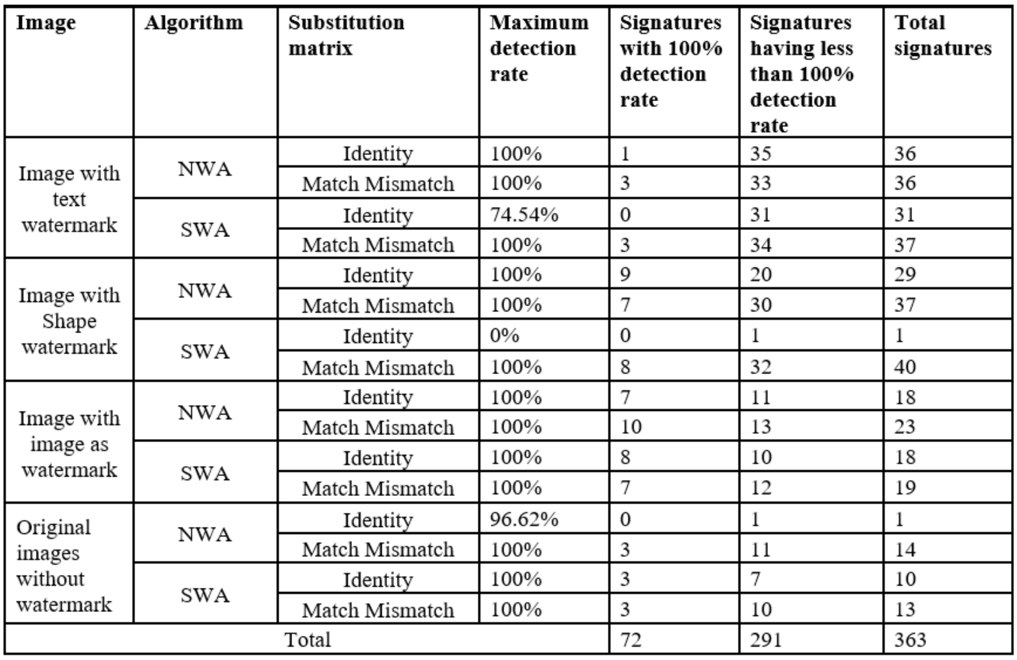


Image	Algorithm	Substitution matrix	Maximum detection rate	Signatures with 100% detection rate	Signatures having less than 100% detection rate	Total signatures
Image with text watermark	NWA	Identity	100%	1	35	36
		Match Mismatch	100%	3	33	36
	SWA	Identity	74.54%	0	31	31
		Match Mismatch	100%	3	34	37
Image with Shape watermark	NWA	Identity	100%	9	20	29
		Match Mismatch	100%	7	30	37
	SWA	Identity	0%	0	1	1
		Match Mismatch	100%	8	32	40
Image with image as watermark	NWA	Identity	100%	7	11	18
		Match Mismatch	100%	10	13	23
	SWA	Identity	100%	8	10	18
		Match Mismatch	100%	7	12	19
Original images without watermark	NWA	Identity	96.62%	0	1	1
		Match Mismatch	100%	3	11	14
	SWA	Identity	100%	3	7	10
		Match Mismatch	100%	3	10	13
Total				72	291	363

Signatures of non-watermarked images were tested using Clamscan against non-watermarked images. From Table 5.7, the lowest detection rate is 96.62% for NWA_ID. In this case, because one signature was obtained from pairwise alignment and it was not already tested, less than 100% detection rate could not be overcome. Rest of the cases have 100% detection rate.

With signatures extracted from eight local and eight global alignments, two local and one global alignments in total three alignment signatures had less than 100% detection rate. That indicates excellent detection performance for identifying watermarked / non-watermarked images. In total, 363 signatures were tested, 72 (i.e. 19.83%) signatures had a 100% detection rate, the rest had less than 100% detection rate.

The performance of this proposed watermark identification was not compared with any other approach, because of its novelty. Traditional watermark extraction and identification are very specialised for only one watermarking scheme, so there is always a 100% detection rate for untampered watermarked images. Our approach is more robust, as it can be used for different kinds of untampered or non-degraded and watermarked images.

Discussion on Experiment 1 results

The proposed syntactic approach under the aid of string matching NWA and SWA with different combinations of identity and match-mismatch scoring matrix experimented on four datasets, three watermarked (i.e., text (I2), shape (I3) and image (I4)) and one non-watermarked (I1) datasets for automatically generating specific image signatures. These signatures were detected from all images of the same dataset. Our analysis shows in Tables 5.6 and Table 5.7 that the current BIIIA approach successfully and consistently identifies all watermarked / non-watermarked datasets. The ultimate purpose of any syntactic technique is to find the potential 'grammar' of an image from a relatively small number of test sets for robust automatic signature extraction.

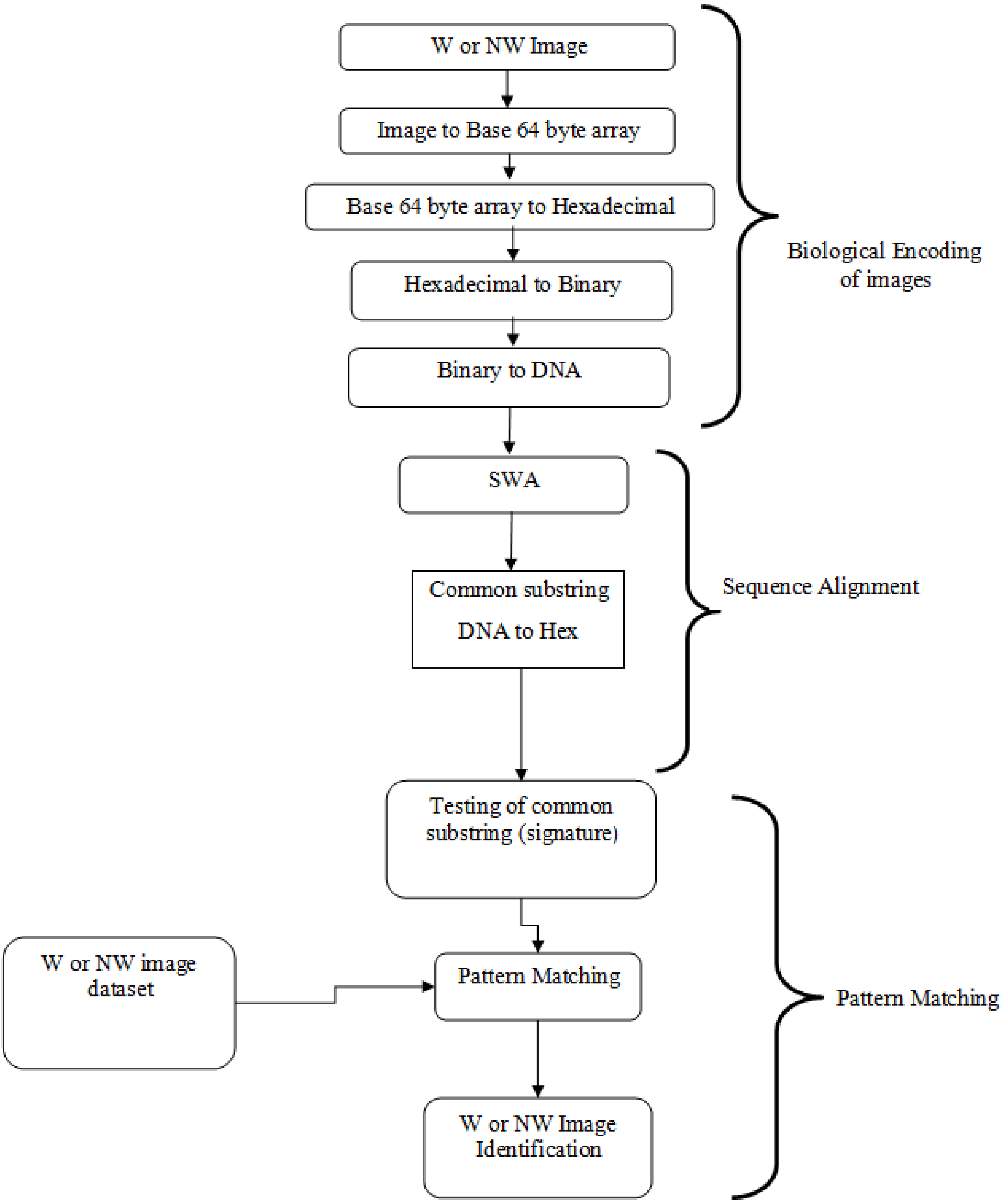

W or NW Image
Image to Base 64 byte array
Base 64 byte array to Hexadecimal
Hexadecimal to Binary
Binary to DNA
Biological Encoding of images
SWA
Common substring DNA to Hex
Sequence Alignment
Testing of common substring (signature)
W or NW image dataset
Pattern Matching
W or NW Image Identification
Pattern Matching

Figure 5.5 Final BIIIA method

The BIIIA has the notable concerns that it will identify unknown watermarked variants. The proposed work reveals the novelty that can identify different watermarks with one approach efficiently; the requirement for developing the watermark-specific approaches and identifying watermark becomes obsolete. Our findings for Result 1 answers sub-question 3 (a), i.e., DNA representation of the image is suitable for BIIIA in Table 5.6. Sub-question 3 (b) was answered by Result 1 and Result 2, i.e., the SWA algorithm with ID matrix is the best combination of sequence alignment for biologically-encoded images by using BIIIA.

Moreover, from the signatures tested by using the software Clamscan and the experimental results provided in Table 5.7, it can be concluded that there is a 100% detection rate except for a very few cases. Additionally, the hypothesis of BIIIA algorithm proved partially correctness as we successfully identified watermarked/non-watermarked images; testing of the final BIIIA method on image identification will be performed in Experiment 2. These findings led to finalize the proposed prototype of BIIIA in Figure 5.4; Experiment 1 for the BIIIA method finalized DNA and SWA with ID matrix. The proposed BIIIA method is shown in Figure 5.5 where all the findings of experiment 1 are considered.

Experiment 2. Experiment 1 proved the best encoding scheme in biology, i.e., the DNA; the best sequence alignment algorithm, i.e., the SWA with ID matrix, is employed for further validation of the final BIIIA method shown in Figure 5.5 by using the image datasets prepared in Section 3.8.3, i.e., datasets D1 to D12. In other words, Step 1 degraded NWD (D2 to D6) dataset created by using D1 (NWND image dataset) and WD (D8 to D12) image datasets, utilizing D7 (WND image dataset) in Section 3.8.3. Step 2 explains how we converted images into an acceptable form of sequence alignment. Step 3 explains the DNA representation method of images. Step 4 explains the pairwise local alignment of DNA using SWA. Step 5 explains the procedure of common substring (signature) extraction from aligned DNA. Finally, Step 6 deals with the procedure of testing the signatures by converting back to hexadecimal (hex) code.

A. Biology-based encoding of the degraded images from MPS:

Step 1. The dataset creation process was explained in Section 3.8.3, so we are not repeating it here.

Step 2. *Selection and conversion.* In this step, the test set was prepared; we selected two random images. Thirty-two test images were used as the test set for image datasets: two from D1 and six from each dataset D2, D3, D4, D5 and D6 (i.e., two images for black and white scanning mode (BW), two images for color scanning mode (C) and two images for greyscale scanning mode (G). In total, six images are for each dataset from D2 to D6). Similarly, thirty-two test images were selected from the degraded and watermarked images: six images from each dataset from D8 to D12 (two images for black and white scanning mode (BW), two images for color scanning mode (C) and two images for greyscale scanning mode (G)). Also, two from each NWND dataset D1 and D7 dataset. So, 64 images were selected. For resolving the issue of unavailability of conversion tools with regard to bigger images to a byte array for getting the required format, all dataset images were resized. The same conversion to hexadecimal of the BIIIA method was followed, i.e., an image to base64byte array, base64byte array to hexadecimal.

Step 3. In this step, the extracted hexadecimal from Step 2 was converted into binary code and then into DNA, as the BIIIA method was explained in Section 5.4.

B. Sequence Alignment

The DNA sequence extracted from the test images in Step 3 (watermarked / non-watermarked and degraded) were used as the input to JAligner for sequence alignment.

Step 4. In this step, pairwise local alignment was accomplished as explained below:

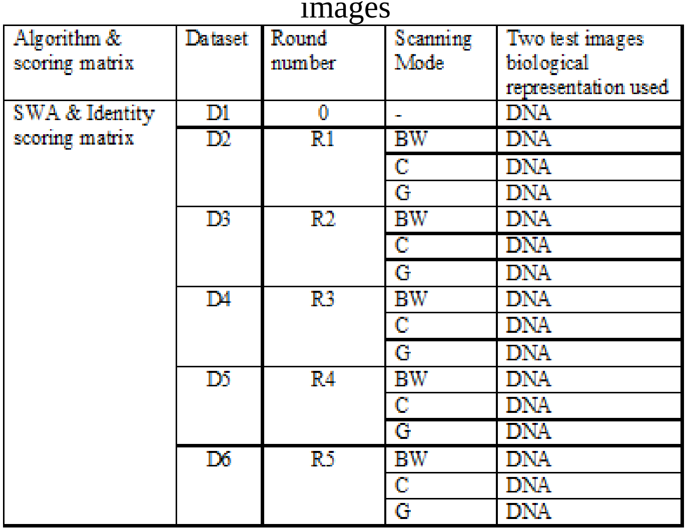
(a) *Pairwise local alignment.* SWA with an ID scoring matrix was used for pairwise local alignment.

Two extracted DNA sequences from the two images (i.e., from D1 to D12) in Step 3 was utilized as an input for the JAligner. In the proposed work, test experiments on combinations of each dataset are shown in Table 5.8, where R1, R2, R3, R4, and R5 are the round numbers 1, 2, 3, 4, and 5 of MPS. BW represents black and white, C stands for colour, and G refers to greyscale scanning mode. We have six datasets, say, D1, D2, D3, D4, D5, and D6 (where D1 is the original image dataset, D2 to D6 are its degraded variants of non-watermarked images) as shown in Table 5.8.

From these datasets based on the selected test images, pairwise local alignments (using SWA and ID matrix) will be performed on the extracted DNA of two images from Step 3, one alignment for D1 images. We applied a similar procedure to three alignments (i.e., one for each BW scanned, color and greyscale scanned images) for dataset D2; between the two test images, the DNA from Round 1 (R1) was degraded in black and white (BW) scanning mode; then, between the two images, R1 was degraded in color (C) scanning mode; lastly, between the two R1 degraded images in greyscale (G) scanning mode; so on for D3, D4, D5 and D6 having three alignments. In total, 16 pairwise local alignments were performed for non-watermarked and degraded datasets.

Thirty-two images were used as test sets represented biologically as DNA, two from D1 and six from each dataset D2, D3, D4, D5 and D6. This indicates that, in total, we had 32 biological representations of images (i.e., DNA) available for pairwise local alignment. Biological representation (i.e., DNA) was obtained from images to get the most common substring pattern or to find conserved regions. In our case, 32 DNA representations were available which result in 16 sequence alignments by using SWA algorithm with the ID matrix.

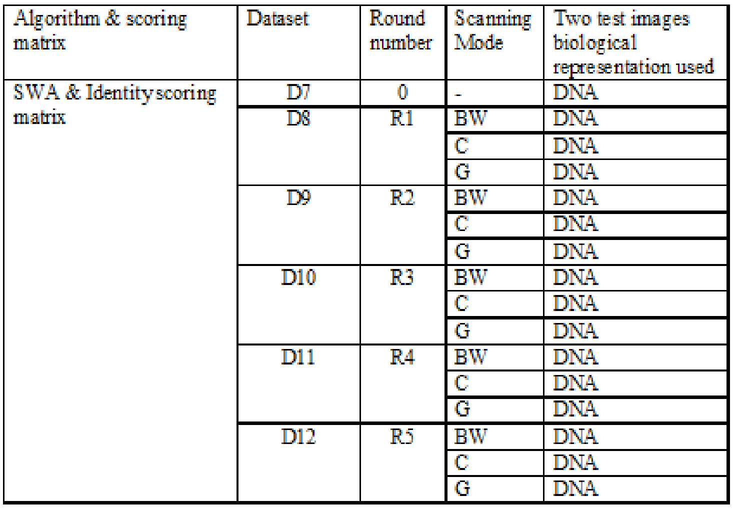
Table 5.8 Combination used for pairwise local alignment of non-watermarked images



Algorithm & scoring matrix	Dataset	Round number	Scanning Mode	Two test images biological representation used
SWA & Identity scoring matrix	D1	0	-	DNA
	D2	R1	BW	DNA
			C	DNA
			G	DNA
	D3	R2	BW	DNA
			C	DNA
			G	DNA
	D4	R3	BW	DNA
			C	DNA
			G	DNA
	D5	R4	BW	DNA
			C	DNA
			G	DNA
	D6	R5	BW	DNA
			C	DNA
			G	DNA

A similar combination and procedure was followed for the watermarked and degraded image datasets from D7 to D12 for pairwise local alignment (where D7 is the original watermarked images and D8 to D12 are its degraded variants) as shown in Table 5.9.

Table 5.9 Combination used for pairwise local alignment of watermarked images



Algorithm & scoring matrix	Dataset	Round number	Scanning Mode	Two test images biological representation used
SWA & Identity scoring matrix	D7	0	-	DNA
	D8	R1	BW	DNA
			C	DNA
			G	DNA
	D9	R2	BW	DNA
			C	DNA
			G	DNA
	D10	R3	BW	DNA
			C	DNA
			G	DNA
	D11	R4	BW	DNA
			C	DNA
			G	DNA
	D12	R5	BW	DNA
			C	DNA
			G	DNA

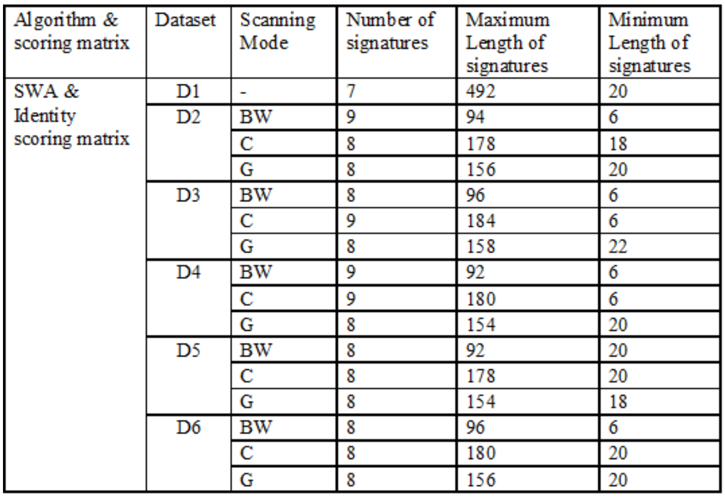
Thirty-two images were used as test sets that were represented biologically as DNA, two from D7, and six from each dataset D8, D9, D10, D11 and D12. This indicates that in total we had 32 biological representations of the images (i.e., DNA) which are available for the pairwise local alignment. In our case, 32 biological representations were available, which resulted in 16 sequence alignments by using SWA with the ID matrix.

In total, 32 local alignments were performed, 16 local pairwise alignments for non-watermarked and degraded images, 16 for watermarked degraded images.

C. *Pattern Matching using Signatures*

Step 5. The signature extraction process was followed from the previously reported Naidu and Narayanan approach (Naidu & Narayanan, 2016). Aligned sequences were obtained from Step 4; those common substrings were extracted, which we call it as signatures that were used for detecting a family of degraded watermarked / non-watermarked variants of the images.

Table 5.10 Signatures from non-watermarked and degraded images after MPS by using pairwise local alignment with SWA algorithm

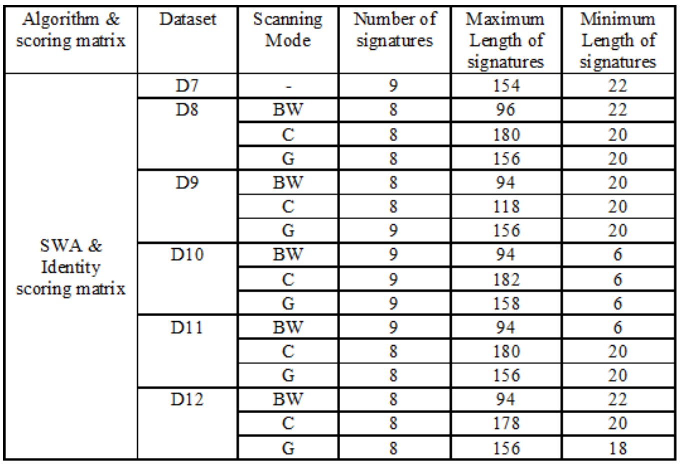


Algorithm & scoring matrix	Dataset	Scanning Mode	Number of signatures	Maximum Length of signatures	Minimum Length of signatures
SWA & Identity scoring matrix	D1	-	7	492	20
	D2	BW	9	94	6
		C	8	178	18
		G	8	156	20
	D3	BW	8	96	6
		C	9	184	6
		G	8	158	22
	D4	BW	9	92	6
		C	9	180	6
		G	8	154	20
	D5	BW	8	92	20
		C	8	178	20
		G	8	154	18
	D6	BW	8	96	6
		C	8	180	20
		G	8	156	20

Signatures were extracted from 16 aligned sequences of NWD and NWND images. Pairwise alignments, resulting for signature extraction, were summarized in Table 5.10, where the number of signatures were extracted for individual cases (i.e., alignments) with the maximum and minimum length of signatures. The maximum number of nine signatures were obtained for D2, the minimum number was 7 for Dataset D1. The minimum length of a signature was 6 and the maximum was 492 for D2 and D1. For WD and WND images, signatures were extracted from 16 aligned sequences of WD and WND images. Results, after pairwise alignment and signature extraction, are summarized in Table 5.11, where a number of signatures was extracted for individual cases (i.e. alignments), with the maximum and minimum lengths of signatures were assembled.

The maximum number of signatures was nine for the D7 dataset and the minimum was eight for many datasets like D8, BW scanned datasets. The minimum length of a signature was 6 for many datasets; the maximum was 182 for D10 color images.

Table 5.11 Signatures from watermarked and degraded images after MPS by using pairwise local alignment with SWA algorithm



Algorithm & scoring matrix	Dataset	Scanning Mode	Number of signatures	Maximum Length of signatures	Minimum Length of signatures
SWA & Identity scoring matrix	D7	-	9	154	22
	D8	BW	8	96	22
		C	8	180	20
		G	8	156	20
	D9	BW	8	94	20
		C	8	118	20
		G	9	156	20
	D10	BW	9	94	6
		C	9	182	6
		G	9	158	6
	D11	BW	9	94	6
		C	8	180	20
		G	8	156	20
	D12	BW	8	94	22
		C	8	178	20
		G	8	156	18

Step 6. In this final step, the DNA sequence representation of all signatures obtained in Step 5 were converted back to hexadecimal format. These signatures were tested against degraded variants of watermarked / non-watermarked images using ClamAV (Clamscan antivirus scanner); for example, the signatures, obtained from the watermarked images, are tested on original watermarked image datasets; the signatures, obtained from watermarked and degraded images from black and white (B&W) scanning modes (i.e. D8) of degradation, were tested against the dataset D8 of BW scanned images. In short, signatures, obtained from dataset D7 particular scanned mode images, are tested on the same dataset D7 with the same scanning mode images.

Results. The results of Experiment 2 will be presented in two parts: the first part (result 1) will explain each of the 32 combinations (i.e., 32 SWA alignments) related to identity percentage, similarity percentage, gap percentage, alignment length and alignment score that will analyses the conserved regions in the watermark/non-watermarked/degraded images and their non-degraded variants. The second part (result 2) will answer sub-question 3(c).

Result 1. Conserved regions analysis

Table 5.12 shows that the percentages of identities and similarities range from 90.28% (lowest) to 97.08% (highest).

Table 5.12 The results of sequence alignment of watermarked/non-watermarked DNA-based encoding

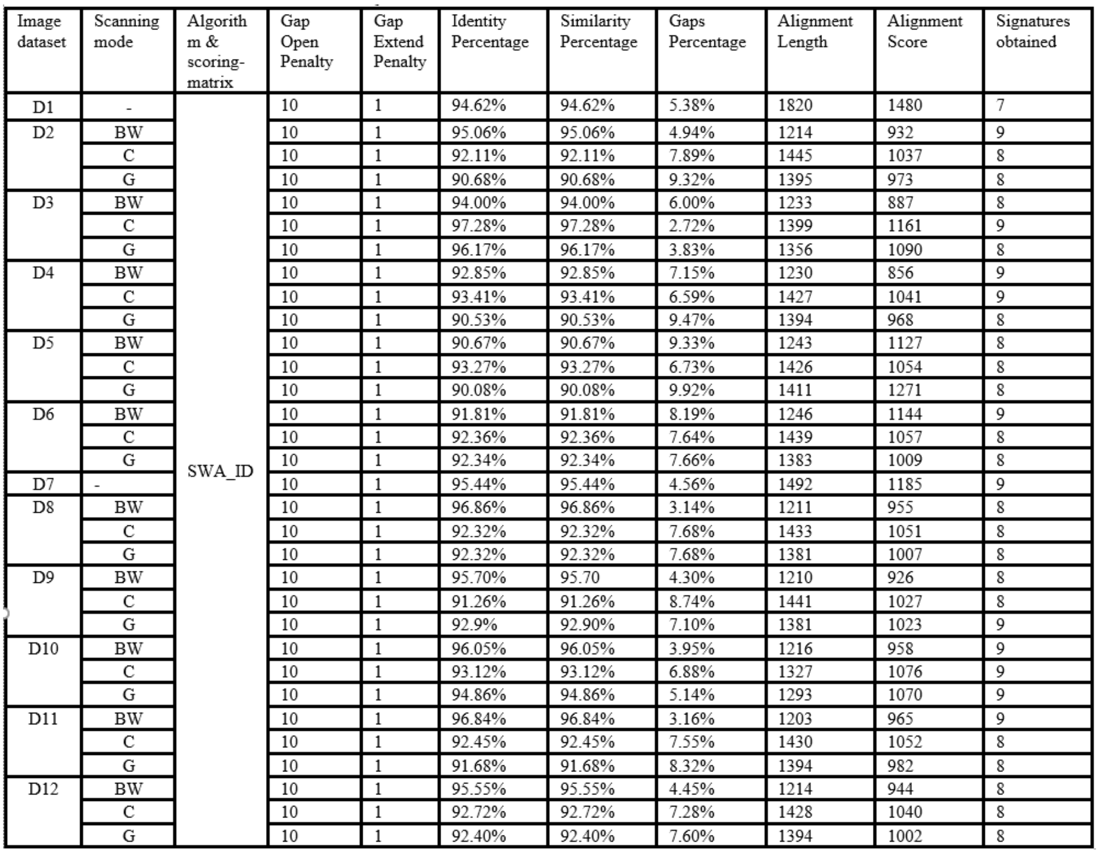


Image dataset	Scanning mode	Algorithm & scoring-matrix	Gap Open Penalty	Gap Extend Penalty	Identity Percentage	Similarity Percentage	Gaps Percentage	Alignment Length	Alignment Score	Signatures obtained
D1	-	SWA_ID	10	1	94.62%	94.62%	5.38%	1820	1480	7
D2	BW		10	1	95.06%	95.06%	4.94%	1214	932	9
	C		10	1	92.11%	92.11%	7.89%	1445	1037	8
	G		10	1	90.68%	90.68%	9.32%	1395	973	8
D3	BW		10	1	94.00%	94.00%	6.00%	1233	887	8
	C		10	1	97.28%	97.28%	2.72%	1399	1161	9
	G		10	1	96.17%	96.17%	3.83%	1356	1090	8
D4	BW		10	1	92.85%	92.85%	7.15%	1230	856	9
	C		10	1	93.41%	93.41%	6.59%	1427	1041	9
	G		10	1	90.53%	90.53%	9.47%	1394	968	8
D5	BW		10	1	90.67%	90.67%	9.33%	1243	1127	8
	C		10	1	93.27%	93.27%	6.73%	1426	1054	8
	G		10	1	90.08%	90.08%	9.92%	1411	1271	8
D6	BW		10	1	91.81%	91.81%	8.19%	1246	1144	9
	C		10	1	92.36%	92.36%	7.64%	1439	1057	8
	G		10	1	92.34%	92.34%	7.66%	1383	1009	8
D7	-		10	1	95.44%	95.44%	4.56%	1492	1185	9
D8	BW		10	1	96.86%	96.86%	3.14%	1211	955	8
	C		10	1	92.32%	92.32%	7.68%	1433	1051	8
	G		10	1	92.32%	92.32%	7.68%	1381	1007	8
D9	BW		10	1	95.70%	95.70	4.30%	1210	926	8
	C		10	1	91.26%	91.26%	8.74%	1441	1027	8
	G		10	1	92.9%	92.90%	7.10%	1381	1023	9
D10	BW		10	1	96.05%	96.05%	3.95%	1216	958	9
	C		10	1	93.12%	93.12%	6.88%	1327	1076	9
	G		10	1	94.86%	94.86%	5.14%	1293	1070	9
D11	BW		10	1	96.84%	96.84%	3.16%	1203	965	9
	C		10	1	92.45%	92.45%	7.55%	1430	1052	8
	G		10	1	91.68%	91.68%	8.32%	1394	982	8
D12	BW		10	1	95.55%	95.55%	4.45%	1214	944	8
	C		10	1	92.72%	92.72%	7.28%	1428	1040	8
	G		10	1	92.40%	92.40%	7.60%	1394	1002	8

Higher values of identities and similarities indicate that higher percentages of the DNA residues were conserved in the biologically-represented images (i.e., DNA) and vice-versa. In the case of DNA-represented images after MPS, variants are aligned with the SWA and ID matrix, the percentage of identities and similarities was between 90.28% and 97.08%.

These results indicate that we could successfully extract common substrings or subsequences (i.e. signatures) by converting or representing images into biological representations (i.e. DNA). The maximum number of signatures extracted was 9 and the minimum was 7. Moreover, from Table 5.12, the percentage of gaps, ranged from 2.72% to 9.92%, represents that the quantity of insertion and deletions, ranged from 2.72% to 9.92%. During the alignment between the two strings, a 0% gap indicates a

100% match as the matching percent is decreased; the amount of the gap will also increase.

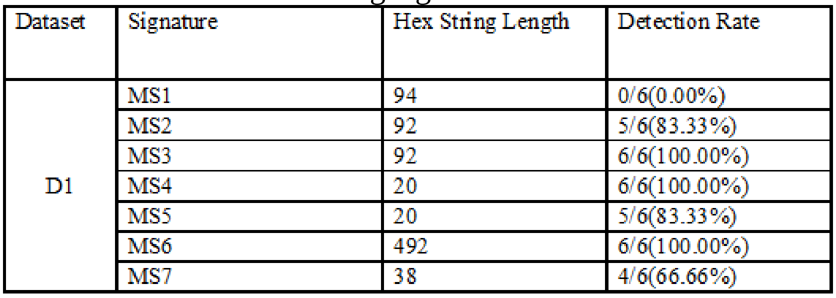
The BIIIA method followed the Naidu and Narayanan approach (Naidu & Narayanan, 2016) or gap open and gap penalties, i.e., fixed combination of gap open (i.e. 10) and gap extend (i.e., 1), we changed the gap penalty from 0.5 to 1). We explored SWA algorithm with the ID scoring matrices on watermarked / non-watermarked and degraded variants by using MPS for image detection. From Table 5.12, every dataset has the percentage of similarity and identity, which is more than 85%, that will reflect a higher number of conserved regions, are conserved during the DNA representation of the images.

From these results, we can conclude that the chances for identifying images for their particular categories will increase. In the next few paragraphs, we will show the identification rates for different cases of our research.

Result 2. Detection rates for watermarked/non-watermarked and degraded images and their original variants.

Tables 5.13, 5.14, 5.15 and 5.16 provide the detection rates for the detection of the NWND dataset (i.e., D1), NWD datasets (i.e., D2, D3, D4, D5, and D6), WND dataset (i.e., D7), and WD (i.e., D8, D9, D10, D11, and D12) images. In total, 32 pairwise alignments were performed by employing SWA algorithms, 250 signatures were extracted in Step 5 for test.

Table 5.13 The detection results of non-watermarked and non-degraded image by using signatures



Dataset	Signature	Hex String Length	Detection Rate
D1	MS1	94	0/6(0.00%)
	MS2	92	5/6(83.33%)
	MS3	92	6/6(100.00%)
	MS4	20	6/6(100.00%)
	MS5	20	5/6(83.33%)
	MS6	492	6/6(100.00%)
	MS7	38	4/6(66.66%)

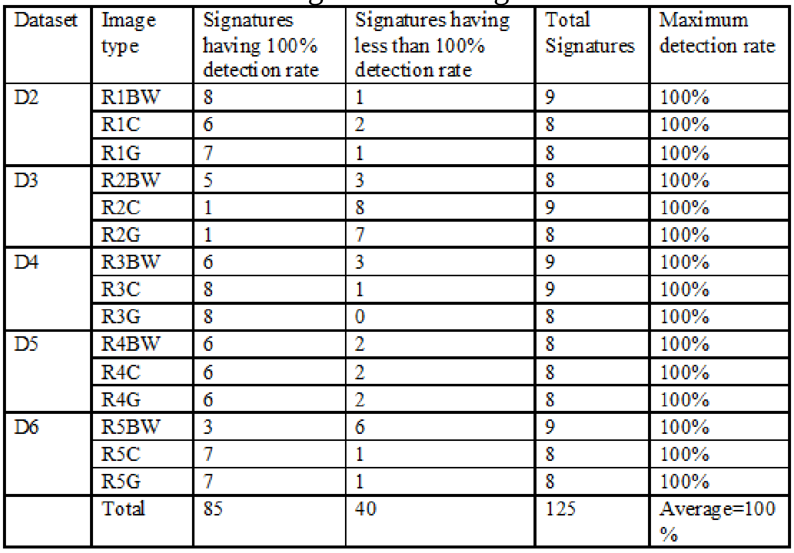
Extracted signatures, after pairwise alignment from dataset D1, were tested against the images of dataset D1 using Clamscan as Step 6 of the BIIIA method. Results for their detection rate are presented in Table 5.13 where MS1 symbolizes signature one and MS2 denotes signature two, so on. Pairwise alignments performed by employing SWA and seven signatures were extracted in Step 5 for test**.**

From Table 5.13, it is clear that, out of seven signatures, three have a 100% detection rate, two have 83.33%, one has a 66.66% detection rate, and one has a 0% detection rate. From these results, we conclude that we successfully identified all images belonging to dataset D1 (i.e. NWND images) by using extracted signatures with the help of our novel syntactic approach for image identification, i.e., BIIIA.

Results for the non-watermarked and degraded (NWD) images are presented in Table 5.14. For dataset D2, R1BW is a combination of two words: the first one R1 stands for Round 1 of MPS; and the second one BW refers the images are scanned in black and white mode. Similarly R1C denotes Round 1 of MPS where images are scanned in color mode; R1G represents Round 1of MPS and images are scanned in greyscale mode and so on; for datasets D3, R2BW, R2C, R2G, etc. Clamscan was used for testing datasets D2, D3, D4, D5 and D6 images. Fifteen of pairwise alignments were performed by employing SWA algorithm; in total, 125 signatures were extracted in Step 5 for test**.**

Table 5.14, the first column has dataset number, the second column contains the images for which detection was performed. The third column keeps the number of signatures having 100% detection rate. The fourth column determines the signatures having less than 100% detection rate; the fifth column shows the total number of signature extracted. Lastly, the sixth column will have the maximum rate of detection.

Table 5.14 The detection results of non-watermarked and degraded image by signatures testing



Dataset	Image type	Signatures having 100% detection rate	Signatures having less than 100% detection rate	Total Signatures	Maximum detection rate
D2	R1BW	8	1	9	100%
	R1C	6	2	8	100%
	R1G	7	1	8	100%
D3	R2BW	5	3	8	100%
	R2C	1	8	9	100%
	R2G	1	7	8	100%
D4	R3BW	6	3	9	100%
	R3C	8	1	9	100%
	R3G	8	0	8	100%
D5	R4BW	6	2	8	100%
	R4C	6	2	8	100%
	R4G	6	2	8	100%
D6	R5BW	3	6	9	100%
	R5C	7	1	8	100%
	R5G	7	1	8	100%
	Total	85	40	125	Average=100 %

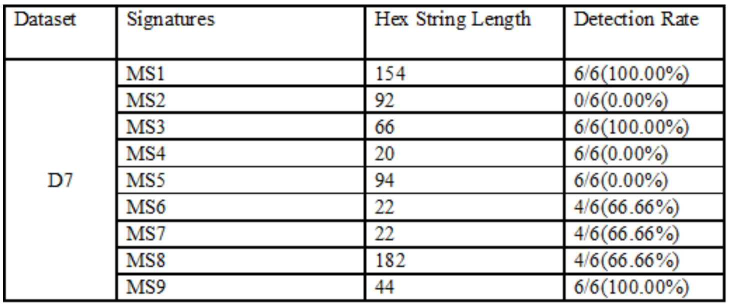
Dataset D2 contains three kinds of images: R1BW, R1C and R1G. Signatures of R1BW images were tested against R1BW images using Clamscan. Table 5.14 shows that 100% detection rate was achieved with eight out of nine signatures individually on R1BW images of dataset D2.

Similarly, with R1C images, eight signatures were tested against R1C images and results indicate a 100 % detection rate for six signatures out of eight. The same followed for R1G images: they were tested against R1G images and a 100% detection rate was achieved for seven signatures out of eight.

The same procedure was followed for testing dataset D3 images: R2BW, R2C, and R2G non-watermarked and degraded (NWD) images. The results show a 100% detection rate for each group of images. Table 5.14 shows that signature test results have a 100% detection rate for the datasets D4 (R3BW, R3C, and R3G), D5 (R4BW, R4C, and R4G) and D6 (R5BW, R5C, and R5G) images without any exception. Out of 125 signatures tested, 85 signatures had a 100% detection rate, and 40 signatures have less than 100% detection rate. In short, the proposed approach had an average detection rate of 100% for identifying non-watermarked images due to degradation from MPS in different scanning modes. We conclude from the above discussion and results that our novel syntactic approach successfully identified the non-watermarked and degraded images in various scanning modes with a 100% detection rate. The above analysis for NWND and NWD images partially answers our research sub-question 2 (c). We identify the non-watermarked and degraded images by using MPS, non-degraded and non-watermarked images by using the proposed syntactic approach.

Datasets D7 to D12 contain watermarked and degraded images after MPS. Extracted signatures from dataset D7 were tested against the images of dataset D7 using Clamscan. Results are shown in Table 5.15 where MS1 symbolizes signature one; MS2 stands for signature two, so on for MS3, MS4, etc.

Table 5.15 The detection results of watermarked and non-degraded images by testing signatures



Dataset	Signatures	Hex String Length	Detection Rate
D7	MS1	154	6/6(100.00%)
	MS2	92	0/6(0.00%)
	MS3	66	6/6(100.00%)
	MS4	20	6/6(0.00%)
	MS5	94	6/6(0.00%)
	MS6	22	4/6(66.66%)
	MS7	22	4/6(66.66%)
	MS8	182	4/6(66.66%)
	MS9	44	6/6(100.00%)

Table 5.15 shows that, out of nine signatures, five had a 100% detection rate, three had 66.66%, and the last one had a 0% detection rate. Results show that we

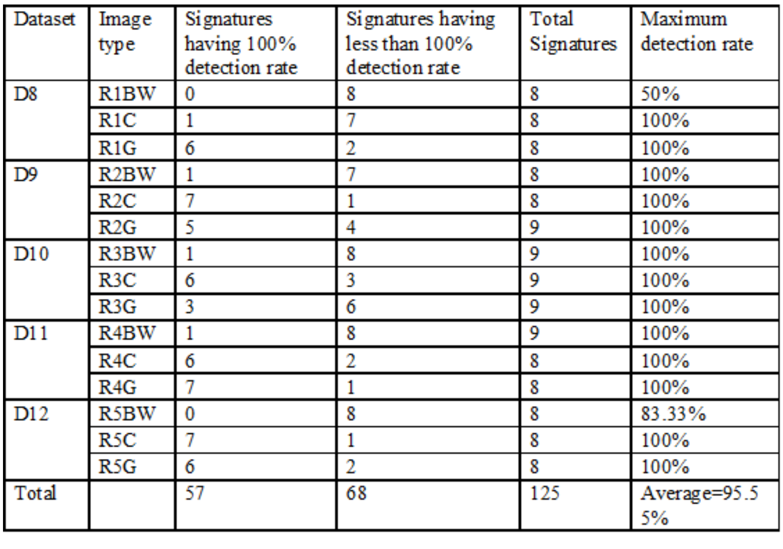
successfully identified all images belonging to dataset D7 (i.e., WND images) by using the proposed syntactic approach with the extracted signatures.

Results of the extracted signature testing for watermarked and degraded images are presented in Table 5.16. For dataset D8, R1BW is a combination of two words: the first R1 represents Round 1 of MPS; the second BW designates that the images are scanned in black and white mode. Similarly, R1C corresponds to Round 1 of MPS where the images are scanned in color mode; the R1G refers to Round 1 of MPS and images are scanned in greyscale mode, so on for dataset D9 (R2BW, R2C, R2G), D10 (R3BW, R3C, R3G), D11 (R4BW, R4C, R4G) and D11 (R5BW, R5C, R5G) images.

For test datasets D8, D9, D10, D11 and D12 images, Clamscan was used. Fifteen pairwise alignments were performed by employing SWA; 125 signatures were extracted in Step 5 for test**.**

Dataset D8 contains three types of the watermarked and degraded (WD) images: R1BW, R1C and R1G images. Signatures of R1BW images were tested against R1BW images using Clamscan. From Table 5.16, it is clear that the maxima of 50% detection rate was achieved after testing eight signatures individually on R1BW images. Similarly, on R1C images, eight signatures were tested against R1C images and results indicate a 100 % detection rate. In the same way, R1G images were tested against R1G images and a 100% detection rate was achieved. The same procedure was followed for testing dataset D9 images (R2BW, R2C, and R2G images); the results show a 100% detection rate for each group of images.

Table 5.16 The results of watermarked and degraded images by testing signatures



Dataset	Image type	Signatures having 100% detection rate	Signatures having less than 100% detection rate	Total Signatures	Maximum detection rate
D8	R1BW	0	8	8	50%
	R1C	1	7	8	100%
	R1G	6	2	8	100%
D9	R2BW	1	7	8	100%
	R2C	7	1	8	100%
	R2G	5	4	9	100%
D10	R3BW	1	8	9	100%
	R3C	6	3	9	100%
	R3G	3	6	9	100%
D11	R4BW	1	8	9	100%
	R4C	6	2	8	100%
	R4G	7	1	8	100%
D12	R5BW	0	8	8	83.33%
	R5C	7	1	8	100%
	R5G	6	2	8	100%
Total		57	68	125	Average=95.55%

For D10 dataset (R3BW, R3C, and R3G) testing results shows a 100% detection rate. For D11 (R4BW, R4C, and R4G) and D12 (R5BW, R5C, and R5G) images signatures were tested without exception from Table 5.16. It is clear that the signatures have a 100% detection rate except for R5BW where the maximum detection rate was around 83.33%. Out of 125 signatures tested, 57 signatures had a 100% detection rate and 68 signatures had less than 100% detection rate.

In short, the proposed approach has an average detection rate of 95.55% for identifying watermarked images after degraded by MPS in multiple scanning modes. We conclude from the above discussion and results that our novel syntactic BIIIA approach successfully identified the watermarked and degraded images in various scanning modes with 100% detection rate except for two cases, R1BW and R5BW, where 50% and 83.3% detection rates, respectively, were achieved.

The above analysis for WND and WD images partially answers our research sub-question 2 (c). We can identify the watermarked and degraded images after MPS as well as non-degraded and watermarked images by using the proposed syntactic approach.

The performance of our identification of degraded images for either watermarked or non-watermarked was not compared with any of other methods, because of its novelty. In traditional, the image identification is lack of the syntactic string-based approach. Our approach is more robust as it can be utilized for different kinds of images, i.e., either for watermarked / non-watermarked images, tampered or degraded by MPS.

5.6 Discussions

The ultimate purpose of the syntactic method for robust and automatic extraction of signatures is to discover the possible ‘grammar’ of an image from a relatively small number of test sets. The proposed syntactic BIIIA approach for the automatic generation of specific dataset image signatures was tested by using two experiments: Experiment 1 based on I1, I2, I3 and I4 datasets using the initial BIIIA method, shown as Tables 5.6 to Table 5.7; Experiment 2 based on D1 to D12 datasets using the final BIIIA method, shown on the Tables 5.13 to Table 5.16. The experiments answer Question 2 as a whole; it is possible to extract syntactic patterns to identify watermarked (W) / non-watermarked (NW) and their degraded variants images by using biological representation and bioinformatics alignment algorithms. Additionally, with sub-question 2(a), the answer was supported by Experiment 1, i.e., DNA is suitable for encoding in biology based on the BIIIA, shown as in Tables 5.6 and Table 5.7; for sub-question 2 (b), the answer was also justified by using Experiment 1, i.e., SWA with ID matrix is the most suitable bioinformatics sequence alignment algorithm for the BIIIA (see Tables 5.6 to 5.7). Experiment 2 answered sub-question 2 (c); it is possible to identify a watermarked/non-watermarked and degraded images from the best biology-based encoding (i.e., DNA) and the best sequence alignment algorithm (i.e., SWA with ID matrix) from sub-questions 2 (a) and (b), shown in Table 5.13 to Table 5.16.

The proposed BIIIA method detected all images of the same dataset for all degraded and original variants of the watermarked / non-watermarked images datasets, i.e., from I1 to I4 datasets and from D1 to D12 datasets images. Our analysis shows in Table 5.6 to Table 5.7 for I1 to I4 datasets, Table 5.13 to Table 5.16 for D1 to D12 datasets, the current image identification BIIIA successfully and consistently identified all watermarked / non-watermarked and degraded datasets after MPS.

These results led to final verification of the hypothesis proposed in Section 5.2, i.e., the research hypothesis for watermarked/non-watermarked and degraded images is tested correct; it is possible to identify syntactic structures or patterns using bioinformatics tools and techniques that help to determine whether a degraded / non-degraded image contains a type of watermark or without a watermark that helps to identify the images of their expected categories.

The BIIIA method has some significant concerns about whether it will identify watermarked and degraded variants; secondly, whether degradation of images will be identified. The proposed work unveils the requirement for novel software technology that identifies the degraded images by using different kinds of degradation with various watermarks using one approach effectively. It indicates the need for developing specific approaches identifying watermarks and degradation. The future possibilities of this research are to implement a software system that will successfully identify NWD, NWND, WD and WND images and extend it for identification of different kinds of watermarked images (i.e., media / non-media watermarked), after degradation.

Our interest is focusing on identification of watermarked / non-watermarked and degraded / non-degraded images by using a biological representation of the image, bioinformatics alignment algorithm and pattern matching; it does not consider rapid evolution of other forms of watermarks, such as media / non-media watermarks and divergent varieties of watermarking algorithms. Furthermore, we did not examine the other types of degradation of watermarked / non-watermarked images for the purpose of identification.

Formation of such big dataset by using divergent degradation methods to create a different kind of watermarked with distinct watermarking approaches, will allow us to verify the robustness of the proposed approach. The proposed approach is to extract vital details of an image (i.e., watermarked / non-watermarked and degraded images). The approach summarised in this thesis may be appropriate to different watermarking algorithms.

5.7 Summary

A novel BIIIA method based on biologically-based encoding of images, sequence alignment, and pattern matching was proposed in this chapter. The image analysis based on pattern matching using the biology-based encoding of images and sequence alignment (i.e., bioinformatics-inspired) was employed for the first time to deal with the identification of watermarked / non-watermarked and degraded images and their original variants. To find the most suitable encoding and sequence alignment algorithms for the BIIIA method, we repeated biology-based encoding with DNA and the same for sequence alignment with the NWA and the SWA. DNA encoding of images in biology and the SWA with the ID matrix is the most suitable one for the BIIIA method.

This research work claims a notable advancement in the area of bioinformatics-inspired image analysis, one crucial area for future work is to see the effect of different gap open and gap extended penalties for the BIIIA method. The second is to check whether BIIIA works with different watermarking algorithms and different types of media / non-media watermarks.

To conclude this chapter, we have shown how biology-based encoding and sequence alignment algorithms help to resolve the research question of the watermarked/non-watermarked and degraded image identification problem. We have proven that our

BIIIA method works well for the identification of watermarked/non-watermarked and degraded images. In order to explore a research method for a bioinformatics-inspired grouping, the next chapter will expand this idea.

Chapter 6

BIIGA: Bioinformatics Inspired Image Grouping Approach

The aim of this chapter is to answer the third question. In Section 6.1*, we introduce the background of different approaches for image phylogeny (i.e. grouping) of degraded / non-degraded images after MPS by using data mining and machine learning. In Section* 6.2*, we describe the hypothesis and research questions addressed in this chapter. The overview and steps of BIIGA are explained in Section* 6.3 *and Section* 6.4*, respectively. In Section* 6.5*, phylogenetic tree and statistical analysis are explained. A discussion of the results of the phylogenetic tree and BIIGA analysis are presented in Section* 6.6*. In Section* 6.7*, the chapter is summarised.*

6.1 Background

Phylogeny explains how a genetically-connected set of organisms evolve with time. In other words, it tells us relationships between a collections of biological things (genes, organs, proteins, etc.) that have advanced from a common forefather. A phylogenetic tree is a tree-like diagram for representing the evolutionary relationship between the different biological species. These trees demonstrate the evolutionary connections among different species or elements—their phylogeny—in light of the resemblance and dissimilarity in the physical or hereditary attributes.

Multimedia phylogeny is used to develop phylogenetic trees of images, videos and audios to find the history of evolution in these digital entities, while grouping relevant and irrelevent entities. For example, audio phylogeny (Nucci, Tagliasacchi, & Tubaro, 2013), Image Phylogeny Trees (IPTs) (Dias, Rocha, & Goldenstein, 2012; Dias, Rocha, & Goldenstein, 2010; Dias, Goldenstein, & Rocha, 2013), video phylogeny (Dias, Rocha, & Goldenstein, 2011), image phylogeny forests (Dias, Goldenstein, & Rocha, 2013; Costa, Oikawa, Dias, Goldenstein, & Rocha, 2014), large scale scenarios (Dias, Goldenstein, & Rocha, 2013) and multiple parenting relationships (Oliveira, et al., 2014). This research focuses on the image phylogenies. Image phylogeny explains how we can find parent-child relationships among near duplicate images. Near duplicate images are transformed copies of an image that conserve its semantics. In our case, we use the term "near duplicate images" for watermarked/non-watermarked images generated from MPS.

In this thesis, the main aim is to redesign the image phylogeny tree, considering the MPS degraded images in different scanning modes (i.e., grayscale, colour, black and white) by using bioinformatics concepts, i.e., MSA and tools. Image phylogeny approaches disussed in the literature were based on the idea of manifold and spectral clustering, dimensionality reduction, viewpoint localisation, heuristics-based solution-oriented Kruskal algorithms, optimum branching or automatic optimum branching, etc. None of the above approaches uses bioinformatics concepts like MSA to develop an image phylogenetic tree. This inspired us to extend the work of the image phylogenetic tree by using bioinformatics concepts.

Data mining, machine learning and phylogenetic approaches were tested for analysing image phylogeny to generate results that group images in two categories as shown below:

a. MPS degraded / non-degraded images grouping

- Between NWD images by using NWND after MPS
- Between WD images by using MPS and WND images

b. Watermarked and non-watermarked images grouping.

In the next paragraphs, we will check whether the data mining and machine learning approaches classify the images as the above categories.

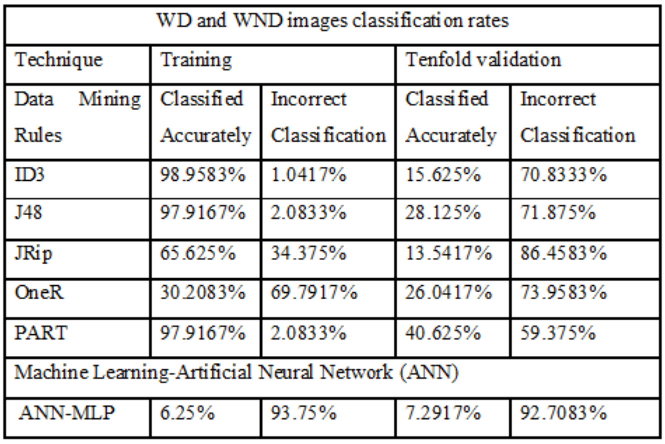
A. Data mining and machine learning for degraded / non-degraded images grouping

Initial experiments for image phylogeny development related to degraded images grouping were performed by employing Weka3.6 and the neural network to classify the degraded / non-degraded images from aligned biologically-encoded images in DNA sequences. Data mining rules were ID3, J48, JRip, OneR and PART and a simple neural network was used, i.e., a multi-layer perceptron with zero nodes and a training time of 10.

a. Between NWD images after MPS and NWND images

Table 6.1 shows that during training, the data mining rule ID3 had the maximum correct classification rate of 98.9583%, after tenfold validation, it decreased to 15.625% which indicates ID3 is not suitable for the classification purpose of WD and WND images. Similarly, for all the other data mining rules, the classification rate decreased after tenfold validation, i.e., for J48 from 97.9167% to 28.125%, JRip from 65.625% to 13.5417%, OneR from 30.2083% to 26.0417%, PART from 97.9167% to 40.625%. These tell us that none of the tested data-mining rules is suitable for classification of the WND and WD images. For the multi-layer perceptron (MLP), the accurate classification rate during training was 6.25%; after the tenfold validation, it increased to 7.2917%, these results also indicate that MLP is also not a good choice for classification purposes. That motivated us to use an alternative choice for the classification of WND and WD images.

Table 6.1 The classification results of watermarked and degraded/non-degraded images by using data mining and MLP

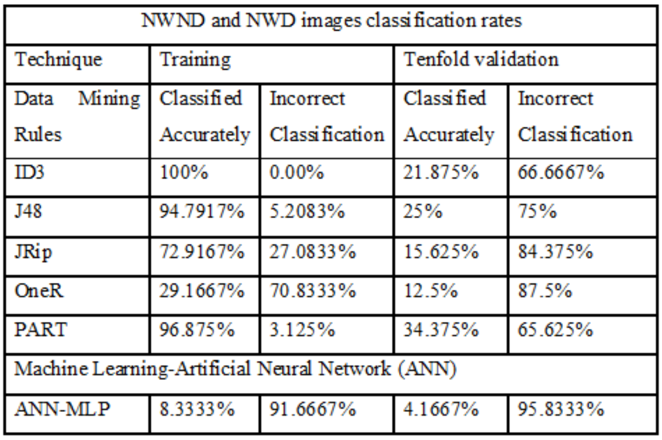


WD and WND images classification rates				
Technique	Training		Tenfold validation	
Data Mining Rules	Classified Accurately	Incorrect Classification	Classified Accurately	Incorrect Classification
ID3	98.9583%	1.0417%	15.625%	70.8333%
J48	97.9167%	2.0833%	28.125%	71.875%
JRip	65.625%	34.375%	13.5417%	86.4583%
OneR	30.2083%	69.7917%	26.0417%	73.9583%
PART	97.9167%	2.0833%	40.625%	59.375%
Machine Learning-Artificial Neural Network (ANN)				
ANN-MLP	6.25%	93.75%	7.2917%	92.7083%

b. Between WD images from MPS and WND images

From Table 6.2, it is clear that, during training, the data mining rule ID3 had the maximum correct classification rate which is 100%; after tenfold validation, it decreased to 21.875% which indicates ID3 is not suitable for the classification of WD and WND images.

Table 6.2 The classification results of non-watermarked and degraded/non-degraded images by using data mining and MLP



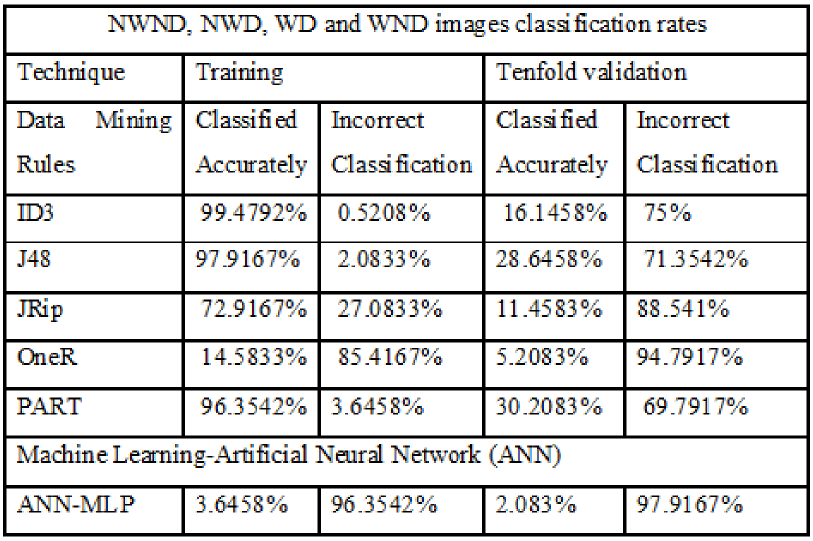
NWND and NWD images classification rates				
Technique	Training		Tenfold validation	
Data Mining Rules	Classified Accurately	Incorrect Classification	Classified Accurately	Incorrect Classification
ID3	100%	0.00%	21.875%	66.6667%
J48	94.7917%	5.2083%	25%	75%
JRip	72.9167%	27.0833%	15.625%	84.375%
OneR	29.1667%	70.8333%	12.5%	87.5%
PART	96.875%	3.125%	34.375%	65.625%
Machine Learning-Artificial Neural Network (ANN)				
ANN-MLP	8.3333%	91.6667%	4.1667%	95.8333%

Similarly, for all the other data mining rules, the classification rate decreases after tenfold validation, i.e. for J48 from 94.7917% to 25%, for JRip from 72.9167%% to 15.625%, for OneR from 29.1667% to 12.5%, for PART from 96.875% to 34.375%. These results tell us that none of the tested data-mining rules is suitable for classification of the NWND and NWD images. For the MLP, the accurate classification rate during training was 8.3333%; after the tenfold validation, it decreased to 4.1667%, these results also indicate that MLP is not a good choice for classification purposes. That again motivated us to use an alternative choice for the classification of NWND and NWD images.

B. *Data mining and Machine Learning for grouping Watermarked and non-watermarked images*

Finally, all DNA sequences of watermarked / non-watermarked and degraded / non-degraded images were aligned to check whether machine learning and data mining approaches would group *watermarked / non-watermarked* images correctly.

Table 6.3 The classification results of watermarked/non-watermarked and degraded/non-degraded images by using data mining and MLP



NWND, NWD, WD and WND images classification rates				
Technique	Training		Tenfold validation	
Data Mining Rules	Classified Accurately	Incorrect Classification	Classified Accurately	Incorrect Classification
ID3	99.4792%	0.5208%	16.1458%	75%
J48	97.9167%	2.0833%	28.6458%	71.3542%
JRip	72.9167%	27.0833%	11.4583%	88.541%
OneR	14.5833%	85.4167%	5.2083%	94.7917%
PART	96.3542%	3.6458%	30.2083%	69.7917%
Machine Learning-Artificial Neural Network (ANN)				
ANN-MLP	3.6458%	96.3542%	2.083%	97.9167%

Table 6.3 shows that, during training, the data mining rule ID3 had the maximum correct classification rate 99.4792%; after tenfold validation, it decreased to 16.1458% which indicates ID3 is not suitable for the classification of watermarked / non-watermarked images. Similarly, for all the other data mining rules, the classification rate decreased after tenfold validation, i.e., for J48 from 97.9167% to 28.6458%, for JRip from 72.9167% to 11.4583%, for OneR from 14.5833% to 5.2083%, for PART from 96.3542% to 30.2083%. These results tell us that none of the tested data mining rules is suitable for classification of the degraded and variants of watermarked / non-watermarked mages. For the multi-layer perceptron (MLP), the accurate classification rate during training was 3.6458%; after the tenfold validation, it decreased to 2.083%, these results indicate that MLP is also not an ideal choice for classification purposes. This indicated that the tested approaches of data mining and machine learning are not good for classification of watermarked / non-watermarked images. In the literature,

phylogenetic trees have been a proven technique for grouping of the species, which motivated us to check the concept phylogenetic tree in bioinformatics as an alternative for WD, WND, NWND or NWD images in all the three cases.

The core idea behind this approach is that the images can mutate as living beings (animals, plants, etc.) evolved in biology. The evolution process of images is performed by MPS degradation, which is considered relevant to the mutation of living beings in this thesis. To test, whether a phylogenetic tree can group the watermarked / non-watermarked and non-degraded images and their degraded variants after MPS, by using our authentically novel approach for the automatic grouping to their relevant group of the images, three phylogenetic trees were generated:

- Phylogenetic Tree 1 for grouping degraded and non-degraded images between NWD images and NWND images.
- Phylogenetic Tree 2 for grouping the degraded / non-degraded images between WD images and WND images for grouping degraded / non-degraded images.
- Phylogenetic Tree 3 for grouping watermarked / non-watermarked images between non-watermarked / watermarked images.

Phylogenetic and molecular evolutionary analysis was conducted by using MEGA version 7 (MEGA7) (Kumar, Stecher, & Tamura 2015) for watermarked/non-watermarked images. MEGA7 is an open source tool, by which three different phylogenetic trees were generated by using different datasets with a maximum likelihood approach.

6.2 Hypothesis Test and Research Questions

The research hypothesis is that for watermarked / non-watermarked and degraded images; it is possible to identify syntactic structures or patterns using bioinformatics-based tools and techniques that determine whether a degraded / non-degraded image contains a type of watermark without a watermark or has specific degradation that helps to group images in the expected categories. If this research hypothesis does not apply to image grouping, it is highly unlike that syntactic structures, extracted by using bioinformatics tools, will be used for grouping images. The third research question is to be answered in this chapter as stated below:

Q.3. Is it possible to extract syntactic patterns or signatures for grouping the watermarked / non-watermarked images before and after MPS degradations by using *biology-based representation, bioinformatics alignment algorithms and phylogenetic trees?*

Sub-question 1. Is it possible to group NWD images and NWND images by using phylogenetic tree analysis?

Sub-question 2. Is it possible to group WD images and WND images by using phylogenetic tree analysis?

Sub-question 3. Is it possible to group watermarked / non- watermarked images from a mix of NWD, NWND, WD, and WND images by using the phylogenetic tree?

6.3 BIIGA : Overview

For successful implementation of the research experiments, the dataset, test environment conditions and a method are required**.**

Dataset

In this chapter, we focus on grouping the NWND / NWD, WND / WD, WD/ WND, NWND/NWD images. However, at this stage, we are not clear what the implication is for using biologically-based image encoding, multiple sequence alignment and phylogenetic tree for grouping the images. This rationale limits us to one watermarking algorithm and one watermark from a potentially large number of watermarking algorithms and the watermarks.

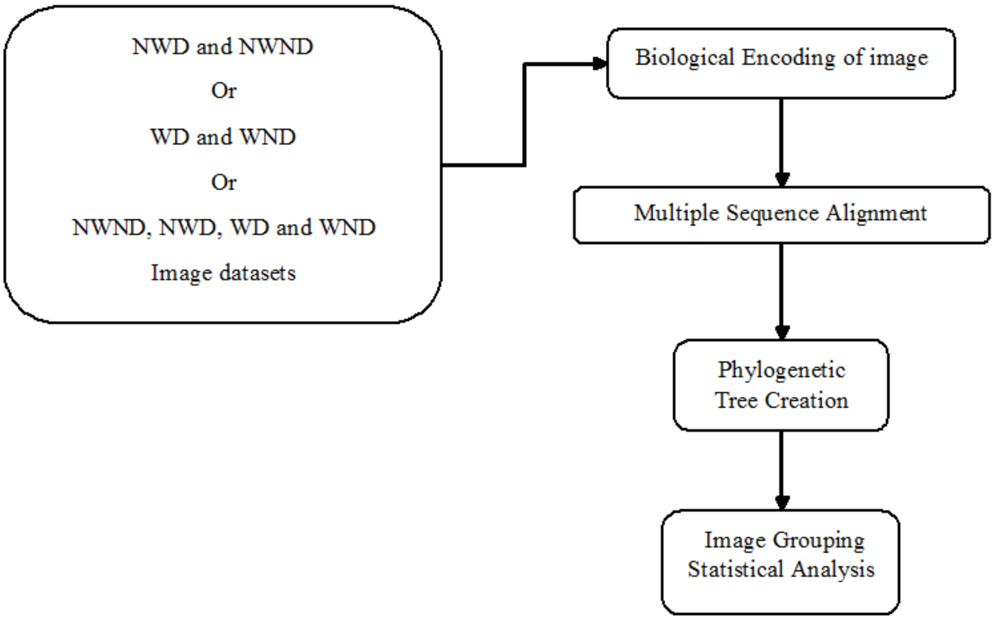

NWD and NWND
Or
WD and WND
Or
NWND, NWD, WD and WND
Image datasets
Biological Encoding of image
Multiple Sequence Alignment
Phylogenetic Tree Creation
Image Grouping Statistical Analysis

Figure 6.1 BIIGA method: overview

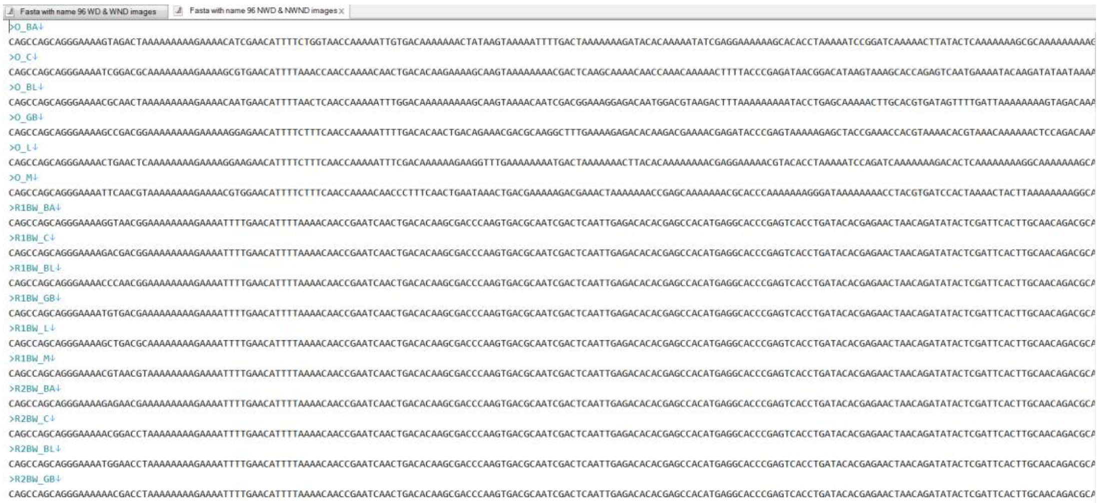
Eight image datasets were used, i.e., G1 to G8, where G1 to G4 comprise non-watermarked images (G1) and their MPS degraded copies in three scanning modes: black and white (BW), colour (C) and greyscale ((G). G5 to G8 are made up of the watermarked images (G5) and the degraded copies in three scanning modes: black and white (G6), colour (G7) and greyscale (G8). More details about the dataset preparation were explained in Section 3.8.3 for BIIGA.

Novel Method BIIGA and Test environment conditions

The BIIGA method consists of three main steps. biology-based encoding of images, multiple sequence alignment and phylogenetic tree creation as shown in Figure 6.1. The test environment conditions along with the steps of BIIGA method are explained in the next few paragraphs.

Biology-based encoding of images

Images of the all datasets, i.e., from G1 to G8, were encoded biologically using the same biologically-based image to DNA encoding utilised in Chapter 5 for the BIIIA system. Figure 6.2 shows that these DNA sequences of the biologically-encoded images were used as the input to MAFT for multiple sequence alignment.


Fasta with name 96 WD & WND images
Fasta with name 96 NWD & NWND images

```
>O_BA↓
CAGCCAGCAGGGAAAAGTAGACTAAAAAAAAAGAAAACATCGAACATTTTCTGGTAACCAAAAATTGTGACAAAAAAACTATAAGTAAAAATTTTGACTAAAAAAAGATACACAAAAATATCGAGGAAAAAAGCACACCTAAAAATCCGGATCAAAAACTTATACTCAAAAAAAGCGCAAAAAAAAG
>O_C↓
CAGCCAGCAGGGAAAATCGGACGCAAAAAAAAGAAAAGCGTGAACATTTTAAACCAACCAAAACAACTGACACAAGAAAAGCAAGTAAAAAAAACGACTCAAGCAAAACAACCAAACAAAAACTTTTACCCGAGATAACGGACATAAGTAAAGCACCAGAGTCAATGAAAATACAAGATATAATAAAA
>O_BL↓
CAGCCAGCAGGGAAAACGCAACTAAAAAAAAAGAAAACAATGAACATTTTAACTCAACCAAAAATTTGGACAAAAAAAAAGCAAGTAAAACAATCGACGGAAAGGAGACAATGGACGTAAGACTTTAAAAAAAAATACCTGAGCAAAAACTTGCACGTGATAGTTTTGATTAAAAAAAAGTAGACAAA
>O_GB↓
CAGCCAGCAGGGAAAAGCCGACGGAAAAAAAAGAAAAAGGAGAACATTTTCTTTCAACCAAAAATTTTGACACAACTGACAGAAACGACGCAAGGCTTTGAAAAGAGACACAAGACGAAAACGAGATACCCGAGTAAAAAGAGCTACCGAAACCACGTAAAACACGTAAACAAAAAACTCCAGACAAA
>O_L↓
CAGCCAGCAGGGAAAACTGAACTCAAAAAAAAGAAAAGGAAGAACATTTTCTTTCAACCAAAAATTTCGACAAAAAAGAAGGTTTGAAAAAAAATGACTAAAAAAACTTACACAAAAAAAACGAGGAAAAACGTACACCTAAAAATCCAGATCAAAAAAAGACACTCAAAAAAAAGGCAAAAAAAGCA
>O_M↓
CAGCCAGCAGGGAAAATTCAACGTAAAAAAAAGAAAACGTGGAACATTTTCTTTCAACCAAAACAACCCTTTCAACTGAATAAACTGACGAAAAAGACGAAACTAAAAAAACCGAGCAAAAAAACGCACCCAAAAAAAGGGATAAAAAAAACCTACGTGATCCACTAAAACTACTTAAAAAAAAGGCA
>R1BW_BA↓
CAGCCAGCAGGGAAAAGGTAACGGAAAAAAAAGAAAATTTTGAACATTTTAAAACAACCGAATCAACTGACACAAGCGACCCAAGTGACGCAATCGACTCAATTGAGACACACGAGCCACATGAGGCACCCGAGTCACCTGATACACGAGAACTAACAGATATACTCGATTCACTTGCAACAGACGCA
>R1BW_C↓
CAGCCAGCAGGGAAAAGACGACGGAAAAAAAAGAAAATTTTGAACATTTTAAAACAACCGAATCAACTGACACAAGCGACCCAAGTGACGCAATCGACTCAATTGAGACACACGAGCCACATGAGGCACCCGAGTCACCTGATACACGAGAACTAACAGATATACTCGATTCACTTGCAACAGACGCA
>R1BW_BL↓
CAGCCAGCAGGGAAAACCCAACGGAAAAAAAAGAAAATTTTGAACATTTTAAAACAACCGAATCAACTGACACAAGCGACCCAAGTGACGCAATCGACTCAATTGAGACACACGAGCCACATGAGGCACCCGAGTCACCTGATACACGAGAACTAACAGATATACTCGATTCACTTGCAACAGACGCA
>R1BW_GB↓
CAGCCAGCAGGGAAAATGTGACGAAAAAAAAAGAAAATTTTGAACATTTTAAAACAACCGAATCAACTGACACAAGCGACCCAAGTGACGCAATCGACTCAATTGAGACACACGAGCCACATGAGGCACCCGAGTCACCTGATACACGAGAACTAACAGATATACTCGATTCACTTGCAACAGACGCA
>R1BW_L↓
CAGCCAGCAGGGAAAAGCTGACGCAAAAAAAAGAAAATTTTGAACATTTTAAAACAACCGAATCAACTGACACAAGCGACCCAAGTGACGCAATCGACTCAATTGAGACACACGAGCCACATGAGGCACCCGAGTCACCTGATACACGAGAACTAACAGATATACTCGATTCACTTGCAACAGACGCA
>R1BW_M↓
CAGCCAGCAGGGAAAACGTAACGTAAAAAAAAGAAAATTTTGAACATTTTAAAACAACCGAATCAACTGACACAAGCGACCCAAGTGACGCAATCGACTCAATTGAGACACACGAGCCACATGAGGCACCCGAGTCACCTGATACACGAGAACTAACAGATATACTCGATTCACTTGCAACAGACGCA
>R2BW_BA↓
CAGCCAGCAGGGAAAAGAGAACGAAAAAAAAAGAAAATTTTGAACATTTTAAAACAACCGAATCAACTGACACAAGCGACCCAAGTGACGCAATCGACTCAATTGAGACACACGAGCCACATGAGGCACCCGAGTCACCTGATACACGAGAACTAACAGATATACTCGATTCACTTGCAACAGACGCA
>R2BW_C↓
CAGCCAGCAGGGAAAAACGGACCTAAAAAAAAGAAAATTTTGAACATTTTAAAACAACCGAATCAACTGACACAAGCGACCCAAGTGACGCAATCGACTCAATTGAGACACACGAGCCACATGAGGCACCCGAGTCACCTGATACACGAGAACTAACAGATATACTCGATTCACTTGCAACAGACGCA
>R2BW_BL↓
CAGCCAGCAGGGAAAATGGAACCTAAAAAAAAGAAAATTTTGAACATTTTAAAACAACCGAATCAACTGACACAAGCGACCCAAGTGACGCAATCGACTCAATTGAGACACACGAGCCACATGAGGCACCCGAGTCACCTGATACACGAGAACTAACAGATATACTCGATTCACTTGCAACAGACGCA
>R2BW_GB↓
CAGCCAGCAGGGAAAAAACGACCTAAAAAAAAGAAAATTTTGAACATTTTAAAACAACCGAATCAACTGACACAAGCGACCCAAGTGACGCAATCGACTCAATTGAGACACACGAGCCACATGAGGCACCCGAGTCACCTGATACACGAGAACTAACAGATATACTCGATTCACTTGCAACAGACGCA
```

Figure 6.2 Screeenshot of the biologically-encoded images

Multiple Sequence Alignment

Bioinformatics-based multiple sequence alignment is used for aligning all the extracted DNA in the three cases: (a) 96 extracted DNA from NWND and NWD images (b) 96 extracted DNA from WND and WD images (c) 192 extracted DNA from NWND, NWD, WND and WD images.

Phylogenetic Tree Creation and Statistical Analysis

Phylogenetic trees were generated by using MEGA7 with a maximum likelihood approach. The statistical analysis was performed on all groups of phylogenetic trees by calculating true positive, true negative, false positive, false negative, sensitivity, negative predictive value, precision and specificity as described in Section 2.2.4. The next section deals with a detailed explanation of the steps of BIIGA.

6.4 Steps of BIIGA

In Step 1, the datasets G1 to G8 including the watermarked / non-watermarked and degraded images were created. Step 2 explains how we converted the images into an acceptable form of sequence alignment. Step 3 explains the image representation method in DNA sequence. Step 4 explains multiple sequence alignment of DNA sequences. Step 5 explains the procedure of generating a phylogenetic tree. Finally, Step 6 deals with statistical analysis of the generated phylogenetic tree. The steps used in the BIIGA method are shown in Figure 6.3.

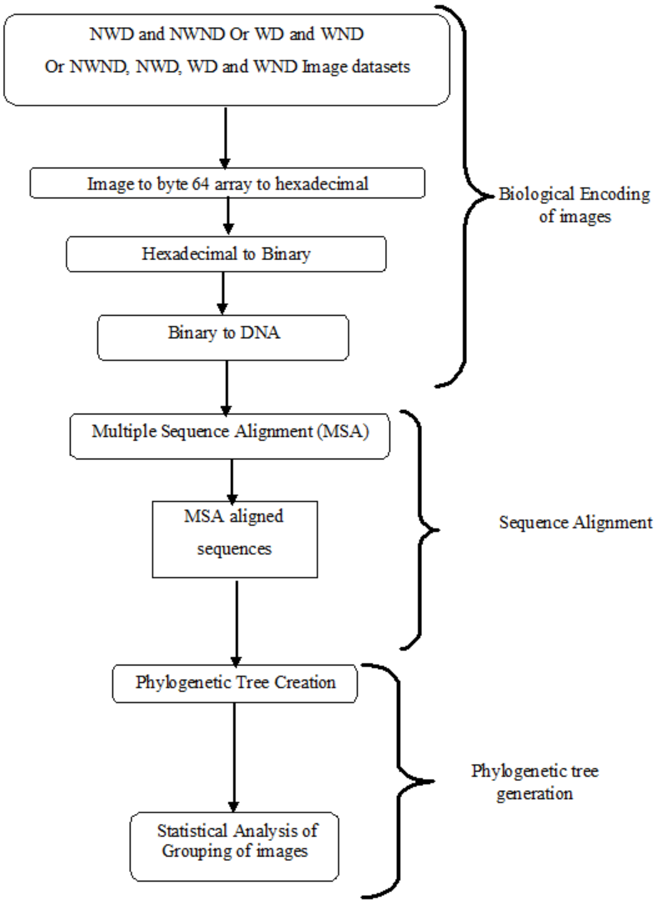

NWD and NWND Or WD and WND
Or NWND, NWD, WD and WND Image datasets
Image to byte 64 array to hexadecimal
Hexadecimal to Binary
Binary to DNA
Biological Encoding of images
Multiple Sequence Alignment (MSA)
MSA aligned sequences
Sequence Alignment
Phylogenetic Tree Creation
Statistical Analysis of Grouping of images
Phylogenetic tree generation

Figure 6.3 BIIGA method

A. Biology-based encoding of images into DNA

Step 1. Dataset Creation. Datasets created in Section 3.8.4 and near duplicate image datasets for BIIGA were used in this chapter to create phylogenetic trees. The three different combinations of datasets were used to create three phylogenetic trees as described below:

- Datasets G1 to G4: NWND and NWD image datasets. By using these datasets, the first phylogenetic tree was created, tested and examined, it would successfully group NWND and NWD images.
- Datasets G5 to G8: WND and WD image datasets. By utilising this dataset, the second phylogenetic tree was generated to test and analyse that it would conveniently group WND and WD images.
- Datasets G1 to G8: NWND, NWD, WND and WD image datasets. By exploiting these datasets, the third phylogenetic tree was developed to test and investigate that it would successfully group watermarked / non-watermarked images.

Step 2. Hexadecimal conversion. In this step, datasets G1 to G8 were converted into base64 byte array and then to hexadecimal.

Step 3. Hexadecimal to DNA. In this step, the extracted hexadecimal data from the dataset G1 to G8 in Step 2 were converted to the binary code and then into the DNA. The DNA-based encoding scheme was as same as in Chapter 5. The rules for the transformation of the hexadecimal into the binary code were followed from the Naidu and Narayanan approach (Naidu & Narayanan, 2016) :‘0’ → ‘0000’; ‘1’ → ‘0001’; ‘2’ → ‘0010’; ‘3’ → ‘0011’; ‘4’ → ‘0100’; ‘5’ → ‘0101’; ‘6’ → ‘0110’; ‘7’ → ‘0111’; ‘8’ → ‘1000’; ‘9’ → ‘1001’; ‘a’ → ‘1010’; ‘b’ → ‘1011’; ‘c’ → ‘1100’; ‘d’ → ‘1101’; ‘e’ → ‘1110’; and ‘f’ → ‘1111’ .

Rules for the conversion of a binary code to DNA were followed from the Naidu and Narayanan approach (Naidu & Narayanan, 2016): ‘00’ → ‘A’; ‘11’ → ‘T’; ‘10’ → ‘G’; and ‘01’ → ‘C’.

B. Sequence Alignment

Step 4. Multiple sequence alignment. The multiple sequence alignment program MAFT (Katoh & Standley, 2016) will align sequences of the biologically-encoded image datasets in the following three situations.

- 96 extracted DNA sequences of NWND and NWD images from G1 to G4 datasets.
- From G5 to G8 datasets, 96 extracted DNA sequences of WND and WD images.
- 192 extracted DNA sequences of NWND, NWD, WND and WD images from G1 to G8 datasets.

C. Phylogenetic Tree Creation and Statistical Analysis

Step 5. Phylogenetic tree creation. a phylogenetic tree is estimated from the multiple aligned sequences obtained in Step 4. Molecular phylogenetic trees are in statistical ways of understanding the grouping and classifying images during the evolution process. There are different ways nowadays to generate a phylogenetic tree, each has its strengths and weaknesses in Tables 2.1, 2.2, 2.3 and 2.4. In Chapter 2, the approaches for phylogenetic tree construction were compared in Table 2.4; it was found that the phenetic approach with the maximum likelihood was most suitable for our research work on grouping the watermarked/non-watermarked and degraded images. The settings for phylogenetic tree creation are as follows:

Approach: Phenetic approach

Tool: MEGA 7 (see Section 3.8.7)

Biological character for image encoding: DNA

Data as input for phylogenetic tree: Character-based method, DNA of the MSA aligned sequence, obtained from Step 4.

Clustering algorithm: a maximum likelihood approach based on the Tamura-Nei model (Tamura & Nei, 1993)

Computational method for finding optimal trees: Heuristic algorithms

Output: Phylogenetic tree

Step 6. Statistical analysis. Tree evaluation must be performed in such a way that it clearly conveys the relevant information to others. The first evaluation in phylogenetic tree creation is bootstrap analysis as the number of groups increases; the bootstrap value starts decreasing and becomes meaningless as discussed in Section 2.2.4; we have four different groups to analyse. Due to this limitation, we are not evaluating phylogenetic trees with phylogenetic analysis, though in the MEGA7 setting, we chose 700 value for bootstrap analysis to get a more accurate tree.

Statistical analysis was performed for actual validation of the correct grouping by using the process of phylogenetic tree reconstruction. Four groups were created for the first two phylogenetic trees: Group 1 to Group 4, i.e., Group 1 for black and white (BW) scanned images, Group 2 for NWND / WND images, Group 3 for colour (C) scanned images and Group 4 for greyscale (G) scanned images. As the third phylogenetic tree was to group watermarked / non-watermarked images separately, only two groups were tested: Group 5 for non-watermarked images, Group 6 for watermarked images. Statistical analysis was employed on all groups by calculating true positive, true negative, false positive, false negative, sensitivity, negative predictive value, precision and specificity. These groups for the different phylogenetic trees are summarised below:

a. Phylogenetic Tree 1 (90 NWND and 6 NWD images from G1 to G4):

Group 1: 30 NWD images scanned in black and white (BW) mode (dataset G2).

Group 2: 6 NWND images (Dataset G1).

Group 3: 30 NWD images scanned in colour (C) mode (Dataset G3).

Group 4: 30 NWD images scanned in greyscale (G) mode (Dataset G4).

b. Phylogenetic Tree 2 (90 WD and 6 WND images from G5 to G8):

Group 1: 30 WD images scanned in black and white (BW) mode (Dataset G6).

Group 2: 6 WND images (Dataset G5).

Group 3: 30 WD images scanned in colour (C) mode (Dataset G7).

Group 4: 30 WD images scanned in greyscale (G) mode (Dataset G8).

c. Phylogenetic Tree 3 (6 NWND, 90 NWD, 90 WD and 6 WND images in total 192 images from G1 to G8):

Group 5: 90 NWD and 6 NWND images (Dataset G1 to G4).

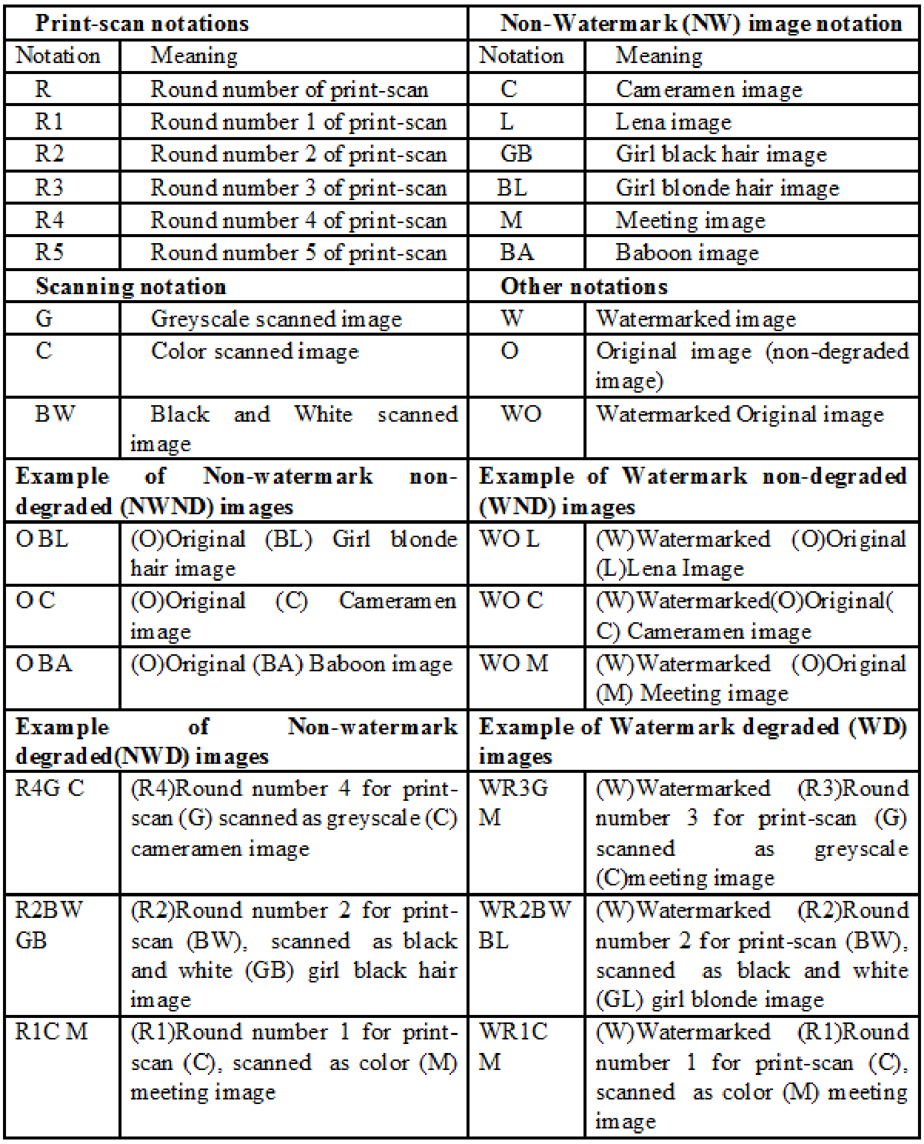
Group 6: 90 WD and 6 WND images (Dataset G5 to G8).

In the next section, there is a detailed explanation of the results along with the analysis of the three phylogenetic trees.

6.5 Results

The results were divided into three parts: the first part (i.e., Result 1) of the results will represent the NWND/NWD images using phylogenetic Tree 1 that groups the images into four groups, i.e., one for non-degraded and three for MPS degraded images, Group 1 for NWD images scanned in BW mode, Group 2 for original NWND images, Group 3 for NWD images scanned in colour mode and Group 4 for NWD images scanned in greyscale mode.

Table 6.4 Notations used in phylogenetic tree



Print-scan notations		**Non-Watermark (NW) image notation**	
Notation	Meaning	Notation	Meaning
R	Round number of print-scan	C	Cameramen image
R1	Round number 1 of print-scan	L	Lena image
R2	Round number 2 of print-scan	GB	Girl black hair image
R3	Round number 3 of print-scan	BL	Girl blonde hair image
R4	Round number 4 of print-scan	M	Meeting image
R5	Round number 5 of print-scan	BA	Baboon image
Scanning notation		**Other notations**	
G	Greyscale scanned image	W	Watermarked image
C	Color scanned image	O	Original image (non-degraded image)
BW	Black and White scanned image	WO	Watermarked Original image
Example of Non-watermark non-degraded (NWND) images		**Example of Watermark non-degraded (WND) images**	
OBL	(O)Original (BL) Girl blonde hair image	WO L	(W)Watermarked (O)Original (L)Lena Image
OC	(O)Original (C) Cameramen image	WO C	(W)Watermarked(O)Original(C) Cameramen image
OBA	(O)Original (BA) Baboon image	WO M	(W)Watermarked (O)Original (M) Meeting image
Example of Non-watermark degraded(NWD) images		**Example of Watermark degraded (WD) images**	
R4G C	(R4)Round number 4 for print-scan (G) scanned as greyscale (C) cameramen image	WR3G M	(W)Watermarked (R3)Round number 3 for print-scan (G) scanned as greyscale (C)meeting image
R2BW GB	(R2)Round number 2 for print-scan (BW), scanned as black and white (GB) girl black hair image	WR2BW BL	(W)Watermarked (R2)Round number 2 for print-scan (BW), scanned as black and white (GL) girl blonde image
R1C M	(R1)Round number 1 for print-scan (C), scanned as color (M) meeting image	WR1C M	(W)Watermarked (R1)Round number 1 for print-scan (C), scanned as color (M) meeting image

The second part of the results (result 2) shows the WND/WD images using phylogenetic Tree 2 that groups the images into four groups, i.e., one for non-degraded and three for MPS degraded images, Group 1 for WD images scanned in BW mode, Group 2 for WND images, Group 3 for WD images scanned in colour scanned mode, and Group 4 for WD images scanned in greyscale mode for investigation. Lastly, the results of the third phylogenetic tree were represented as Result 3 for grouping watermarked / non-watermarked images with group 5 and group 6 for examination.

All three phylogenetic trees were represented in two shapes (i.e., circular and rectangular) for better understanding and visualisation. The notations used in the phylogenetic tree creation are shown in Table 6.4.

The term clade is repeatedly used in the next few paragraphs that express a grouping of items (i.e., in phylogeny "items" represents the species and in this thesis, it refers to the images) that have a common forefather from whom these items are descended. A clade may have a few items or thousands of items. In this thesis, a clade will represent a group of particular categories of images in a phylogenetic tree that was generated and analysed for grouping degraded images after MPS. The BW clade of the phylogenetic tree implies it is a group of images containing the MPS degraded W/NW images scanned in the BW mode.

Similarly, for the phylogenetic tree, a colour and greyscale image clade contains the degraded images scanned in the colour and greyscale mode. The clade of the original images will have the original W or NW images.

Phylogenetic Tree 1. It is generated by using four datasets G1 to G4 that comprise of non-degraded and degraded images as shown in Figure 6.4 and Figure 6.5.

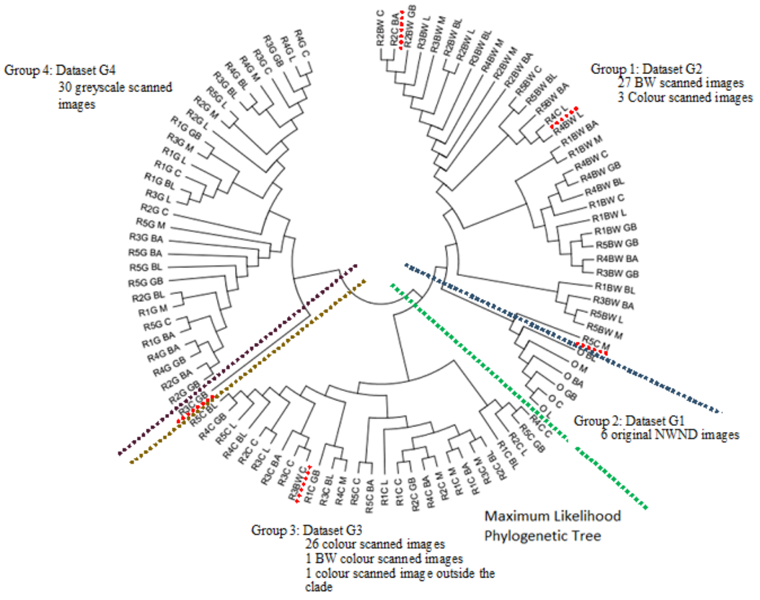


Figure 6.4 Circular shape of phylogenetic tree for non-watermarked and degraded/non-degraded images

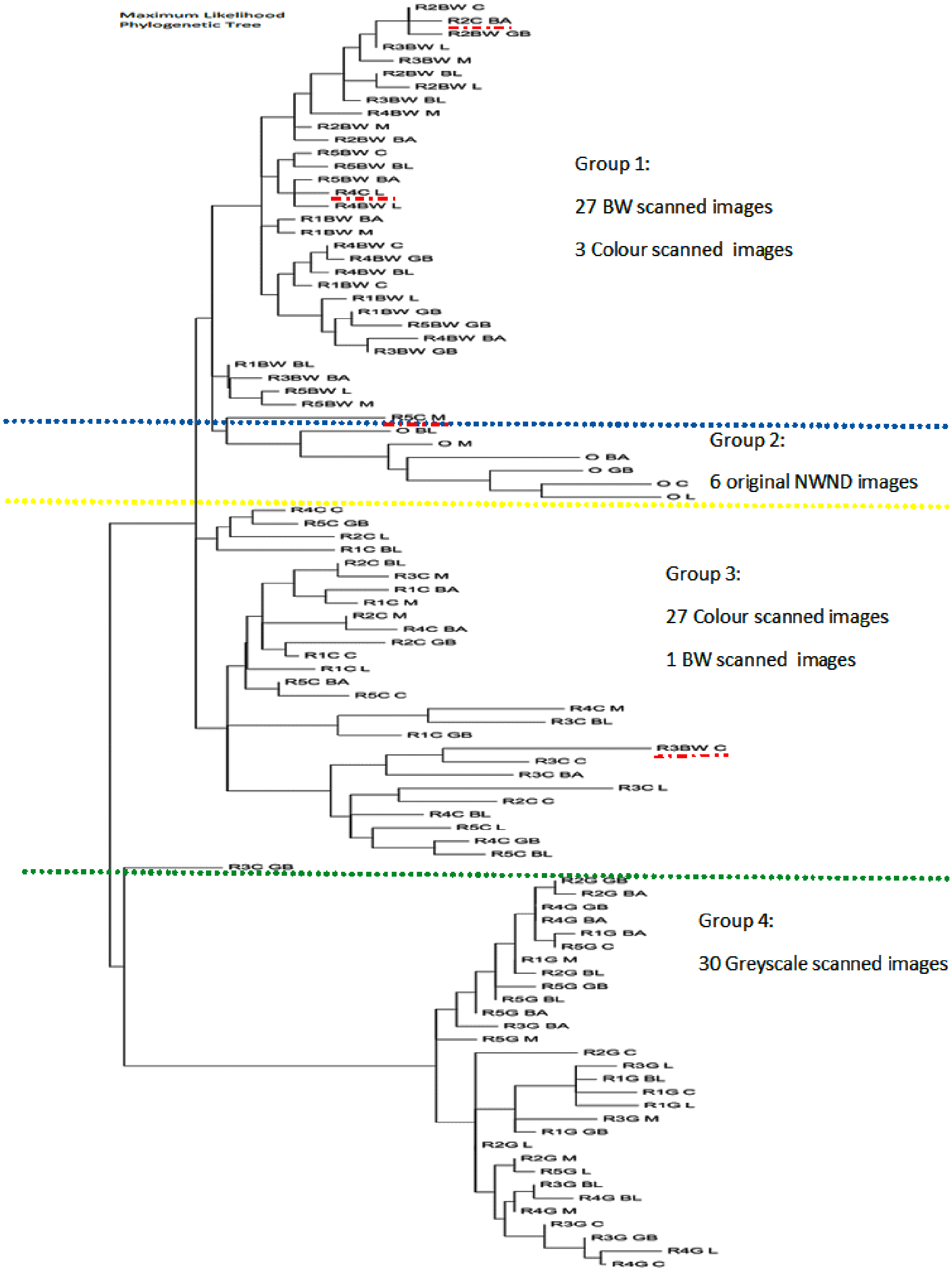


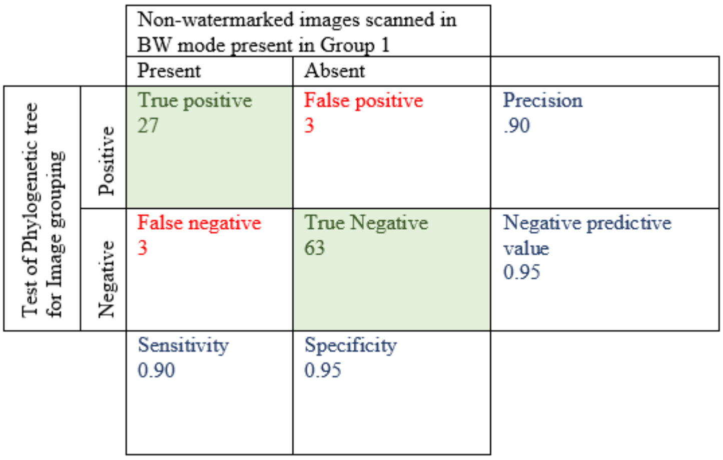
Figure 6.5 Phylogenetic tree in rectangular shape for watermarked and degraded/non-degraded images

For Groups 1 to 4, true positive, true negative, false negative, false positive, sensitivity, specificity, negative predictive value and precision were calculated using Figures 6.4 or Figure 6.5.

Group 1. Analysis was conducted for verifying the accurate grouping of Dataset G2 in Group 1. G2 dataset consists of 30 NWD images scanned in BW mode. That is out of 96 images (i.e., from datasets G1 to G4), 30 belong to Group 1 and 66 are owned by other groups.

True positive means the number of correct groupings of dataset G2 images in Group 1. True negative value suggests the correct rejection of images from Datasets G1 to G4 during phylogenetic tree creation. From Table 6.5, it is clear that true positive value is 27 and true negative is 63. This indicates that 27 images belong to Group 1 clade and are mapped correctly in the right clade as shown in Figure 6.5. Furthermore, true negative value suggests that 63 images do not belong to Group 1 and are not mapped in that clade.

Table 6.5 Measuring the results of grouping non-watermarked & degraded images scanned in binary mode



		Non-watermarked images scanned in BW mode present in Group 1		
		Present	Absent	
Test of Phylogenetic tree for Image grouping	Positive	True positive 27	False positive 3	Precision .90
	Negative	False negative 3	True Negative 63	Negative predictive value 0.95
		Sensitivity 0.90	Specificity 0.95	

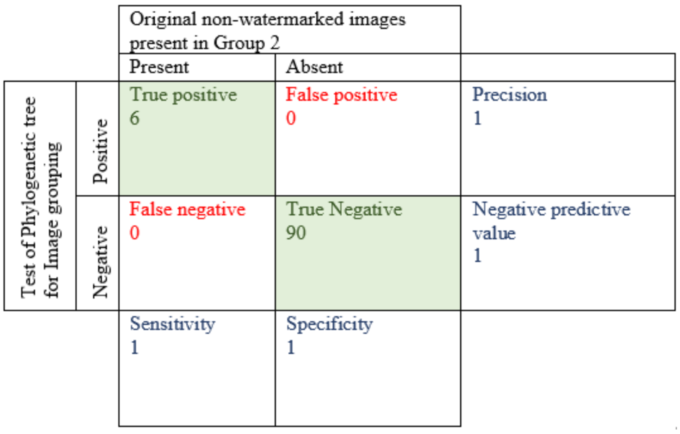
From Table 6.5, the false positive and negative value is 3. False positive indicates the number of images incorrectly grouped in Group 1. It implies that 3 images scanned in colour mode are wrongly mapped in Group 1 clade as represented in Figure 6.5. False negative denotes the number of G2 dataset images incorrectly rejected for grouping in Group 1. It conveys that 3 images belong to Group 1 and are not mapped in the correct clade.

Table 6.5 shows few false negatives and few false positives which indicates the generated phylogenetic Tree 1 is very good at grouping the dataset G2 images in Group 1, i.e., the precision is equal to 90%. It correctly grouped 90% of the images (sensitivity). However, as a grouping test, a higher negative predictive value will reassure that an image is not grouped in the wrong clade (NPV = 95%); it correctly grouped 95% of the non-watermarked & BW scanned images (specificity). This investigation established the great potential of the phylogenetic Tree 1 in grouping dataset G2 images.

Group 2. The investigation was performed for checking the correct grouping of dataset G1 images in Group 2. Dataset G1 has 6 non-watermarked images. That is out of 96 images (i.e., from Datasets G1 to G4), 6 are owned by Group 2 and 90 are associated with other groups.

True positive refers to the count of correct grouping and true negative for right declining of non-watermarked images in Group 2. The true positive and true negative are 6 and 90 can be seen from Table 6.6. It represents that 6 images, belonging to the Group 2 clade, are mapped correctly in the original image as displayed in Figure 6.5. In addition, a true negative value indicates that 90 images are correctly rejected.

Table 6.6 Measuring grouping results of non-watermarked images



		Original non-watermarked images present in Group 2		
		Present	Absent	
Test of Phylogenetic tree for Image grouping	Positive	True positive 6	False positive 0	Precision 1
	Negative	False negative 0	True Negative 90	Negative predictive value 1
		Sensitivity 1	Specificity 1	

From Table 6.6, it is clear that both the false positive and negative values are zero. False positive indicates incorrect image grouping in Group 2. It implies that none of the images is wrongly mapped in the Group 2 clade as shown in Figure 6.5. Similarly, a zero false negative value denotes no incorrect rejection for grouping in the correct clade.

Table 6.6 has zero false negatives and false positives which indicates 100% correct grouping (the sensitivity) with 100% precision in Group 2. Despite this, for a grouping test, 100% negative predictive value (NPV) confirms that none of the images is grouped in the wrong clade and 100% of the images are correctly grouped in group-2 (the specificity). That indicates that the phylogenetic tree performed excellently in grouping images from G1 dataset in Group 2.

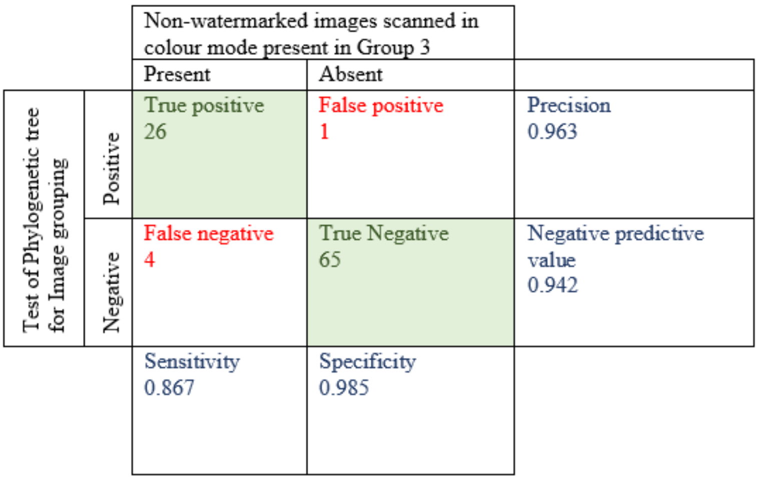
Group-3. The evaluation is performed for examining the correct image grouping in G3 dataset.

Dataset G3 consists of 30 NWD images scanned in colour mode. Out of 96 images (i.e., from datasets G1 to G4), 30 are associated with Group 3 and 66 belong to the rest of the groups.

True positive and true negative represent the number of images grouped in correct groups and right rejection during the grouping of dataset G3 images in Group 3. Moreover, from Table 6.7, the true positive and true negative observed values are 26 and 65 respectively. This indicates that 26 images are mapped correctly in Group 3 and 65 images are rejected correctly during the phylogenetic Tree 1 creation as displayed in Figure 6.5.

Table 6.7 shows that false positive and negative values are 1 and 4, respectively. This indicates that 1 image scanned in BW mode is grouped incorrectly in Group 3 and 4 images are incorrectly rejected for grouping in Group 3 as appear in Figure 6.5. Additionally, one image scanned in colour mode acts as an outlier and it does not belong to any group as shown in Figure 6.5

Table 6.7 Grouping results of non-watermarked and degraded images scanned in colour mode



		Non-watermarked images scanned in colour mode present in Group 3		
		Present	Absent	
Test of Phylogenetic tree for Image grouping	Positive	True positive 26	False positive 1	Precision 0.963
	Negative	False negative 4	True Negative 65	Negative predictive value 0.942
		Sensitivity 0.867	Specificity 0.985	

The generated phylogenetic tree is grouped correctly 86.7% (sensitivity) in Group 3 with 96.3% precision. 94.2% NPV affirms that very few images are grouped in the wrong clade and 98.5% of the images are correctly grouped in Group 3 (the specificity). This shows that the phylogenetic tree performed very well in grouping images of G3 dataset.

Group 4. The investigation was conducted for investigating correct image grouping from dataset G4.

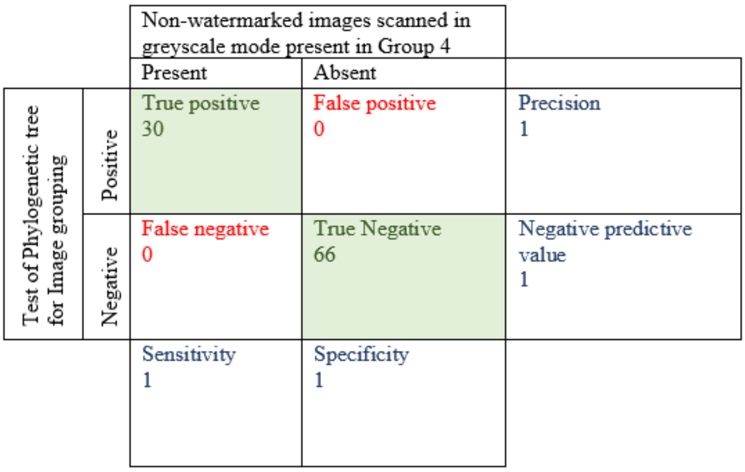
G4 dataset has 30 images scanned in greyscale mode. From 96 images (i.e., from datasets G1 to G4) 30 are owned by Group 4 and 66 are associated with other groups.

Grouping NW images scanned in greyscale mode in Group 4 were measured by true positive and true negative for correctly grouping and correct rejection. From Table 6.7, it is clear that the true positive and the true negative values are 30 and 66 respectively. This shows that 30 images (i.e., all the images of dataset G4) are grouped correctly in Group 4 and 66 images are rejected correctly during phylogenetic tree creation as exhibited in Figure 6.5.

Moreover, from Table 6.7, the false positive and negative values are zero. This represents that no images is grouped incorrectly and wrongly rejected for grouping NW images scanned in greyscale mode in Group-4 as shown in Figure 6.5.

Phylogenetic Tree 1 correctly grouped 100% images (sensitivity) in Group 4 with 100% precision. None of the images is grouped in the wrong clade supported by 100% NPV; 100% non-watermarked images, scanned in greyscale mode, are correctly grouped in Group 4 (specificity). This indicates the excellent achievement of the phylogenetic Tree 1 in grouping images from G4 dataset.

Table 6.8 Measuring the grouping results of non-watermarked images scanned in greyscale mode



		Non-watermarked images scanned in greyscale mode present in Group 4		
		Present	Absent	
Test of Phylogenetic tree for Image grouping	Positive	True positive 30	False positive 0	Precision 1
	Negative	False negative 0	True Negative 66	Negative predictive value 1
		Sensitivity 1	Specificity 1	

Result 2. The generated six WND images and ninety WD images from Datasets G5 to G8 shown as a circular representation in Figure 6.6 and as a rectangular representation in Figure 6.7.

For Groups 1 to 4, true positive, true negative, false negative, false positive, sensitivity, specificity, negative predictive value and precision were calculated using Figure 6.6 or Figure 6.7.

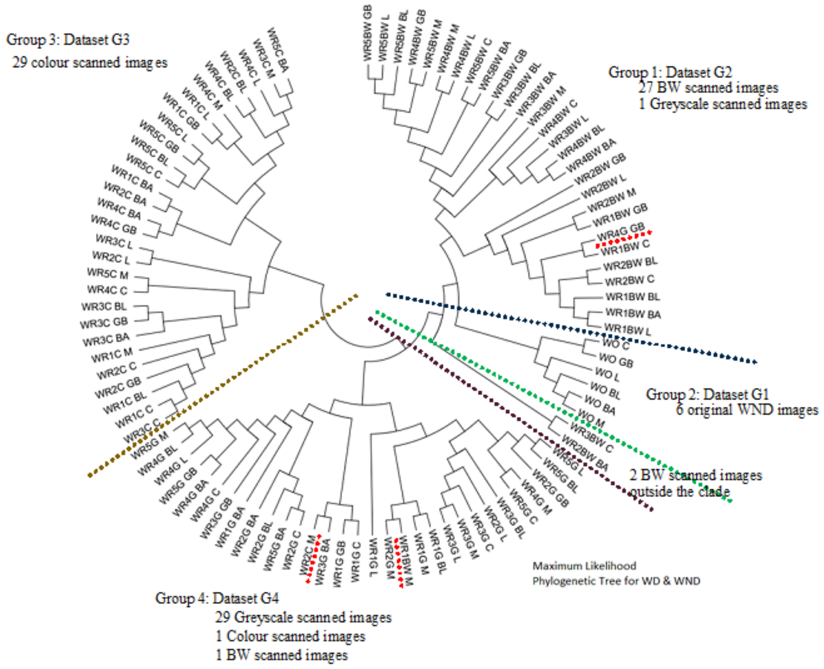


Figure 6.6 Circular shape of phylogenetic tree for degraded/non-degraded watermarked images

Group 1. The analysis was performed for examining the correct grouping of Dataset G6 images in Group 1. G6 Dataset was made up of 30 images scanned in greyscale mode. 30 images belong to group-1 and 66 images are owned by other groups (i.e., from Datasets G5 to G8).

From Table 6.9, it is clear that the value of true positive is 27 and true negative is 65. True negative value suggests that 27 images are grouped correctly in Group 1 as shown in Figure 6.6. Furthermore, true negative value indicates that 65 images were correctly rejected during the second phylogenetic tree.

It is clear from the results in Table 6.5 that the false positive is 1 and the false negative value is 3. This implies that one image scanned in greyscale mode does not belong to this clade, which is wrongly mapped in Group 1 clade as demonstrated in Figure 6.6. False negative denotes that 3 images are not mapped in the correct clade (incorrect rejection).

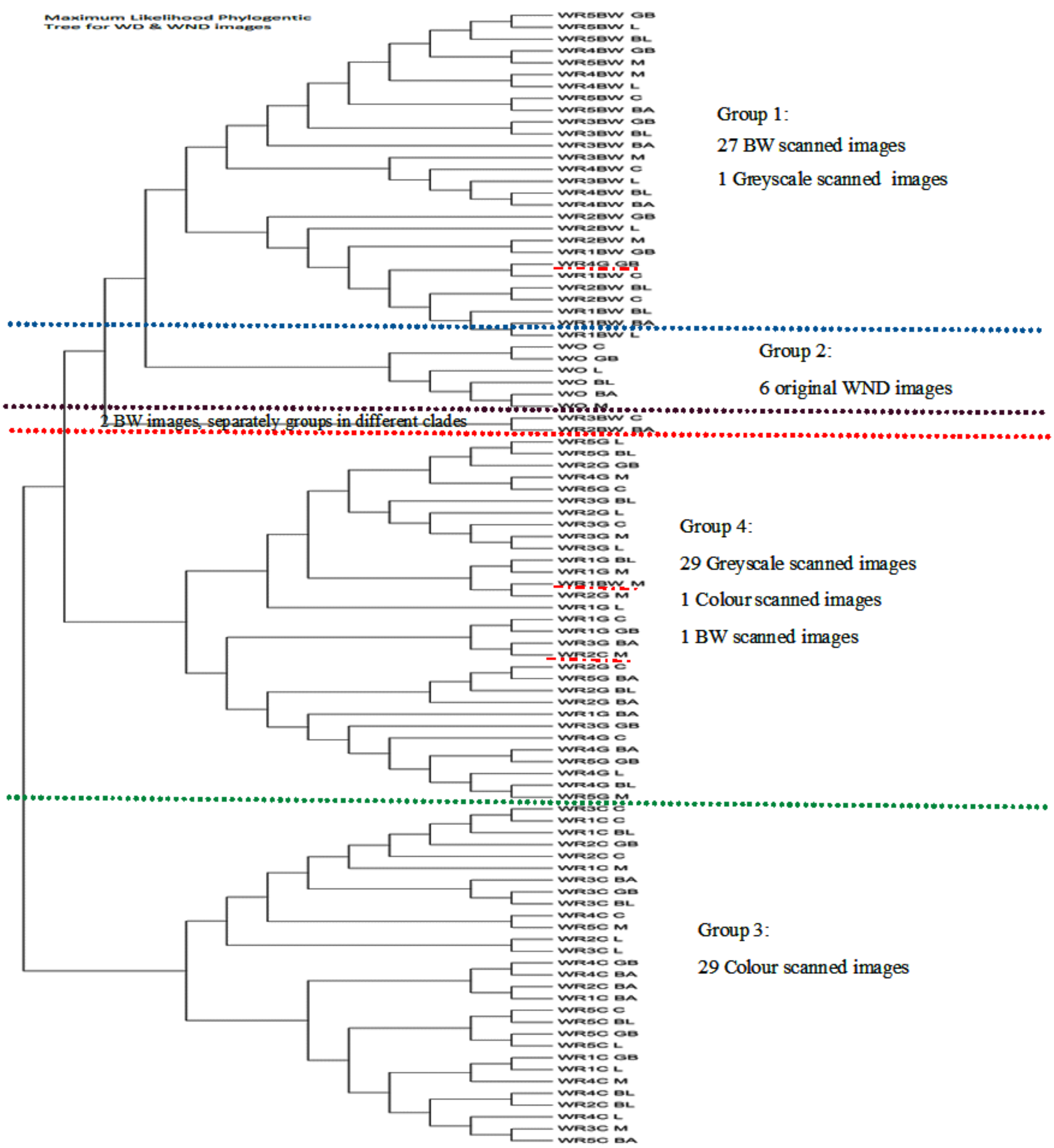
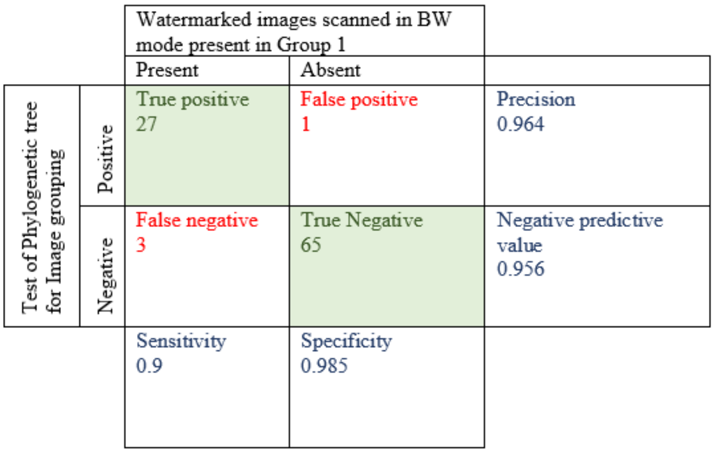


Figure 6.7 Phylogenetic tree in rectangular shape for watermarked/non-watermarked and degraded/non-degraded images

Table 6.9 provides the grouping percentages by using phylogenetic Tree 2. Correct image grouping in Group 1 is 90% (sensitivity) with 96.4% precision. Additionally, 95.6% of images are not grouped in the wrong clade (NPV) and 98.5% watermarked images, scanned in BW mode, are correctly grouped in group-1 (specificity). These results affirm the excellent performance of phylogenetic Tree 2 in grouping G6 dataset images.

Table 6.9 Measuring the grouping results of watermarked and degraded images scanned in binary mode

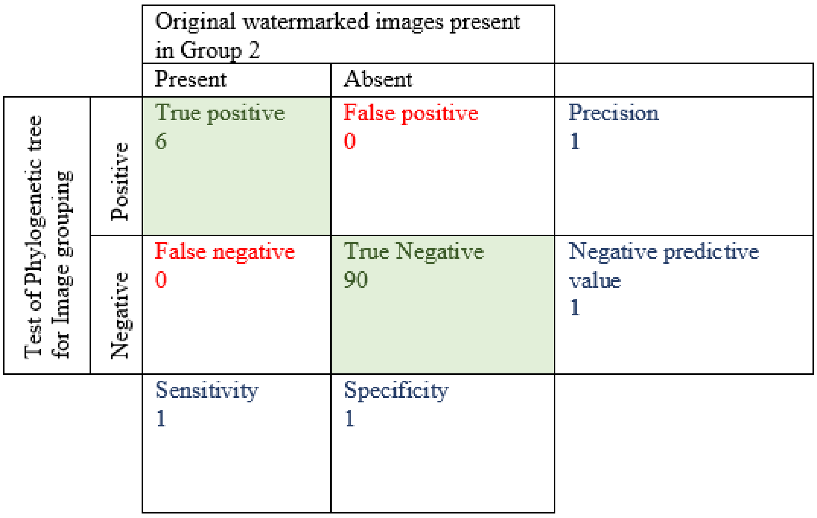


		Watermarked images scanned in BW mode present in Group 1		
		Present	Absent	
Test of Phylogenetic tree for Image grouping	Positive	True positive 27	False positive 1	Precision 0.964
	Negative	False negative 3	True Negative 65	Negative predictive value 0.956
		Sensitivity 0.9	Specificity 0.985	

Group-2. The correct image grouping from dataset G6 is examined in Group 1. G6 dataset is made up of 6 images. That implies 6 images are associated with Group 2 and 90 images belong to other groups out of 96 images (i.e., from datasets G5 to G8).

The true positive and the true negative are 6 and 90 from Table 6.10. True positive indicates that 6 images belong to the Group 2 clade and are mapped correctly in the original image clade as shown in Figure 6.7. In addition, the true negative value represents that 90 images are rejected correctly.

Table 6.10 Measuring the grouping results of watermarked images



		Original watermarked images present in Group 2		
		Present	Absent	
Test of Phylogenetic tree for Image grouping	Positive	True positive 6	False positive 0	Precision 1
	Negative	False negative 0	True Negative 90	Negative predictive value 1
		Sensitivity 1	Specificity 1	

Table 6.10 has zero value for both false positive and negative. False positive indicates images are incorrectly grouped in Group 2. It shows that none of the images is incorrectly mapped in the group-2 clade as presented in Figure 6.7. Similarly, zero false negative value denotes no wrong rejection for grouping in correct clades.

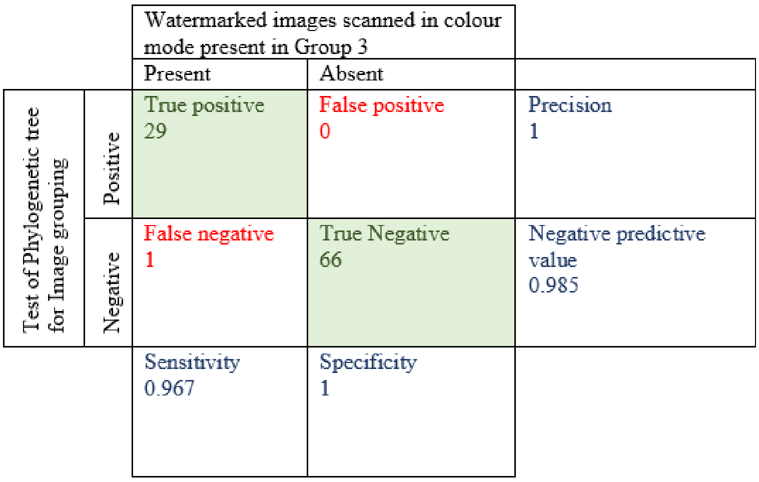
Additionally, in Table 6.10, other values indicate 100% correct grouping of images (the sensitivity) with 100% precision in Group 2. Furthermore, a 100% negative predictive value (NPV) justifies that none of the images is grouped in the wrong clade and 100% of the images are correctly grouped in Group 2 (the specificity). These results verify the magnificent performance of phylogenetic Tree 2 in Grouping G1 dataset images in Group 2.

Group 3. The analysis is performed for verifying the correct image grouping from dataset G7 in Group 3.

Dataset G7 has 30 WD images scanned in colour mode. That is out of 96 images (i.e., from datasets G5 to G8), 30 belong to Group 3 and 66 are owned by other groups.

Table 6.11 provides true positive and true negative values of 29 and 66, respectively. True positive value denotes that 29 images are grouped correctly in Group 3 as shown in Figure 6.7. True negative value signifies that 66 images are rejected correctly to group G7 images in Group 3.

Table 6.11 Measuring the results of grouping watermarked images scanned in colour mode



		Watermarked images scanned in colour mode present in Group 3		
		Present	Absent	
Test of Phylogenetic tree for Image grouping	Positive	True positive 29	False positive 0	Precision 1
	Negative	False negative 1	True Negative 66	Negative predictive value 0.985
		Sensitivity 0.967	Specificity 1	

From Table 6.11, the false positive and negative values are 0 and 1, respectively. This represents that none of the images is grouped incorrectly and 1 image is incorrectly rejected for grouping in Group 3 as displayed in Figure 6.5.

Phylogenetic tree 2, grouped 96.7% images correctly (sensitivity) in Group 3 having a 100% precision. High NPV value, i.e., 98.5% confirms that less images are grouped in the incorrect clade and 100% of the images are correctly mapped in Group 3 (the specificity). This resultant analysis asserts the outstanding performance of the phylogenetic Tree 2 in grouping of images from G7 dataset in group-3.

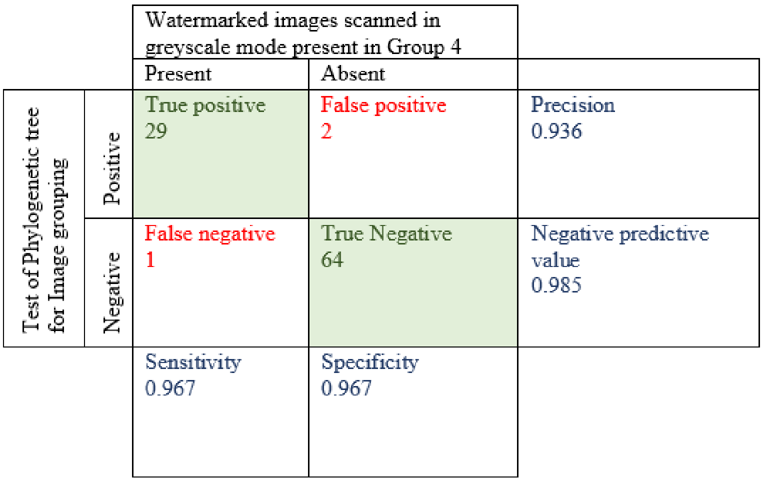
Group 4. The investigation was performed for checking the correct image grouping for dataset G8 in group-4.

G8 dataset has 30 WD images scanned in greyscale mode. That suggests that 30 images are associated with Group 4 and the remaining 66 images are owned by other groups, out of 96 images (i.e., from datasets G5 to G8).

Table 6.12 supplies the true positive and true negative values of 29 and 64 respectively. True positive value explains that 29 images are grouped correctly in Group 4 as displayed in Figure 6.7. True negative suggests that 64 images should be rejected correctly during the grouping process of the phylogenetic Tree 2 for Group 4.

Furthermore, from Table 6.12, the false positive is 2. This shows that two images are grouped incorrectly in Group 4, i.e., one image is scanned in colour mode and the other one is scanned in BW mode as reported in Figure 6.7. False negative is 1, which indicates that only one image was wrongly rejected for grouping in Group 4

Table 6.12 Evaluating results of the watermarked images scanned in greyscale mode



		Watermarked images scanned in greyscale mode present in Group 4		
		Present	Absent	
Test of Phylogenetic tree for Image grouping	Positive	True positive 29	False positive 2	Precision 0.936
	Negative	False negative 1	True Negative 64	Negative predictive value 0.985
		Sensitivity 0.967	Specificity 0.967	

In addition, from Table 6.12, it is clear that phylogenetic Tree 2 grouped correctly 96.7% images (sensitivity) in Group 4 with 93.6% precision. A very few images are grouped in the incorrect clade that is supported by 98.5% NPV; 96.7% watermarked images scanned in greyscale mode are correctly grouped in Group 4 (specificity). This analysis confirms the magnificent performance of the phylogenetic Tree 2, for grouping images from G8 dataset in Group 4.

Result 3. The third phylogenetic tree (i.e. phylogenetic tree 3) for six NWND images, ninety NWD images from Datasets G1 to G4, six WND and ninety WD images from Datasets G1 to G8 shown in Figure 6.8 for circular representation, Figure 6.9 for rectangular representation.

For Groups 5 and 6, true positive, true negative, false negative, false positive, sensitivity, specificity, negative predictive value and precision are calculated using Figure 6.8 or Figure 6.9.

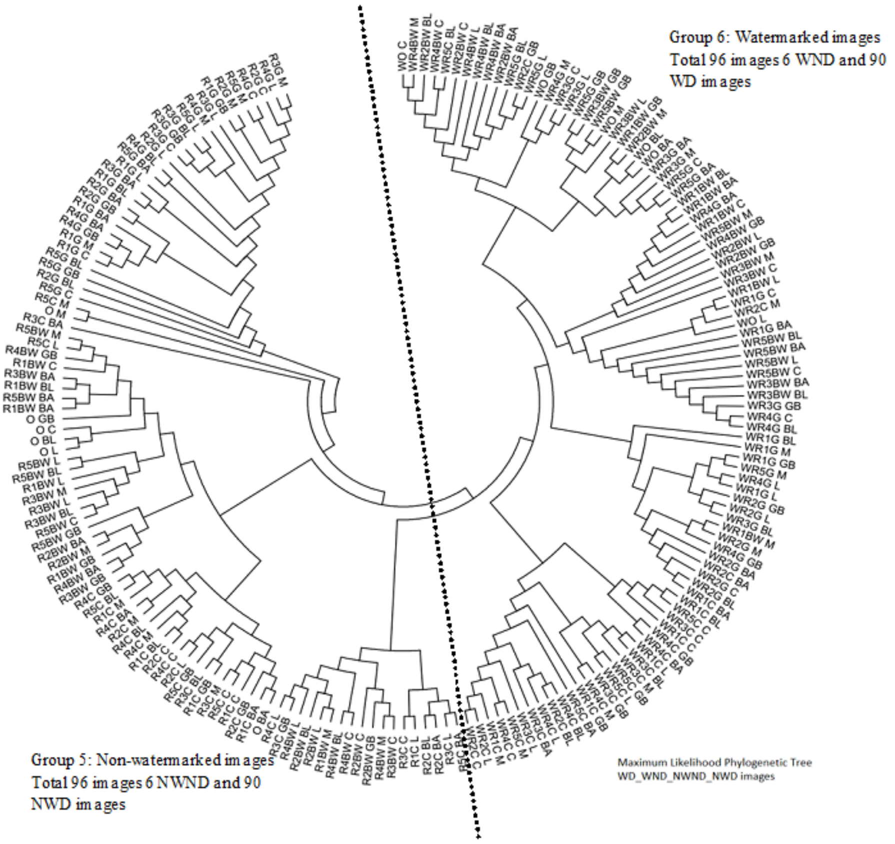
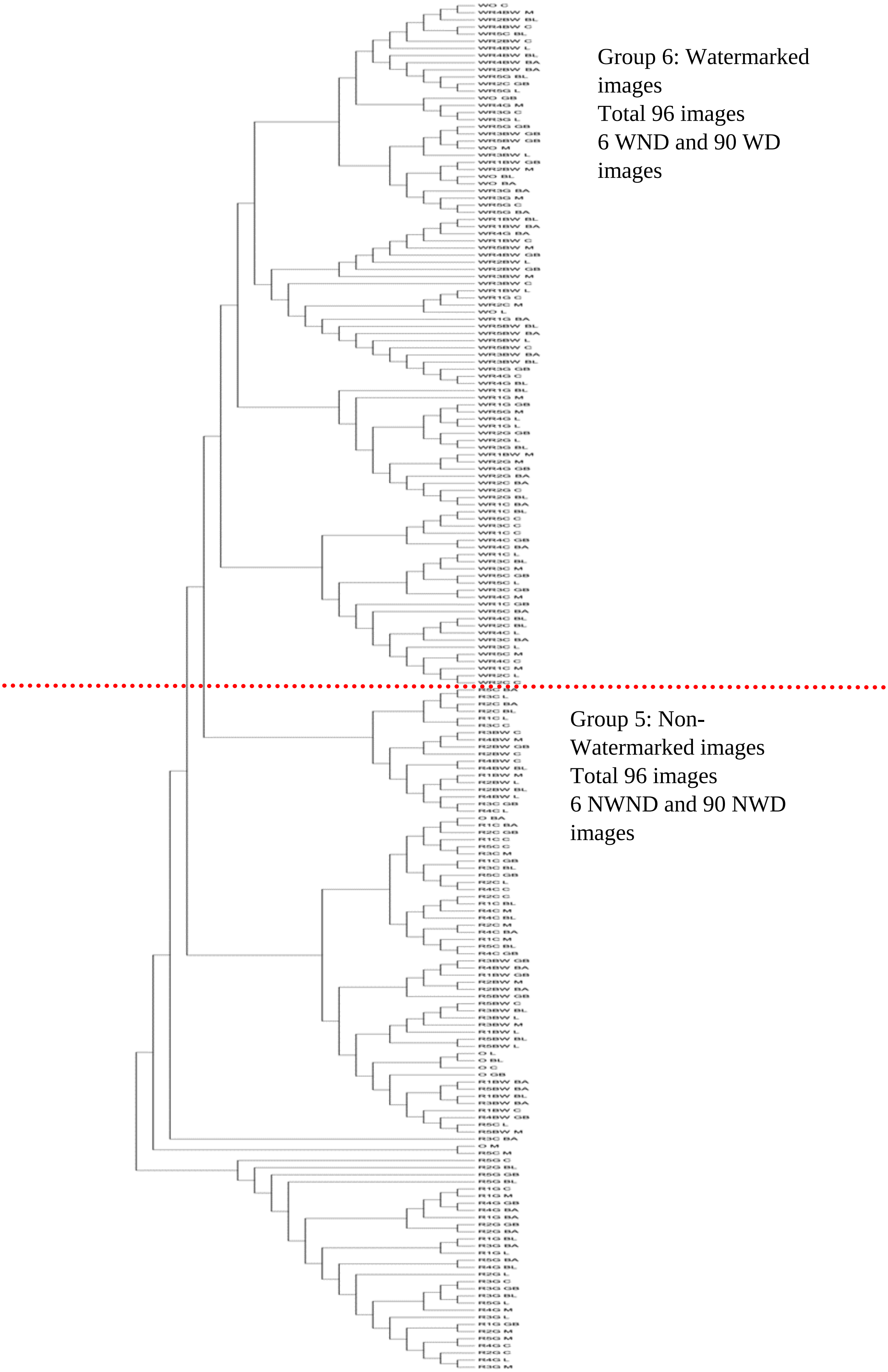


Figure 6.8. Circular shape of phylogenetic tree for watermarked/non-watermarked and degraded/non-degraded images

Group 5. The examination was conducted for checking the correct image grouping from Datasets G1 to G4 in Group 5.

Datasets G1 to G4 contain 96 non-watermarked and degraded images. This indicates that out of 192 images from Datasets G1 to G8, 96 images belong to Group 5 and the rest of the 96 images are associated with the watermarked images.



Group 6: Watermarked images
Total 96 images
6 WND and 90 WD images

Group 5: Non-Watermarked images
Total 96 images
6 NWND and 90 NWD images

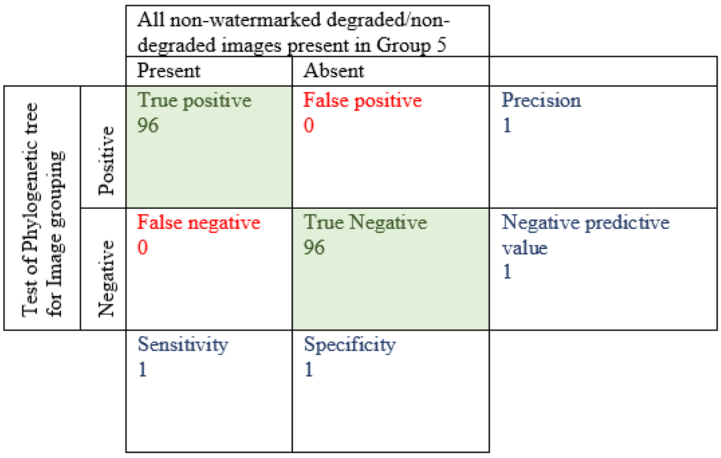
Figure 6.9 Phylogenetic tree in rectangular shape for watermarked/non-watermarked and degraded/non-degraded images

The true positive and the true negative value is 96 from Table 6.13. True positive value specifies that all 96 non-watermarked and degraded/non-degraded images are grouped correctly in Group-5, as displayed in Figure 6.8. True negative expresses that 96 images are rejected correctly during the grouping proccess of the phylogenetic Tree 3 for Group 5.

Moreover, from Table 6.13, it is clear that the false positive and negative is zero. This shows that none of the images is grouped incorrectly in Group 5 and wrongly rejected by phylogenetic Tree 3, as shown in Figure 6.8.

Additionally, Table 6.13 provides that phylogenetic Tree 3, grouped 100% correctly non-watermarked and degraded/non-degraded images (sensitivity) in Group 5 with 100% precision. 100% NPV asserts that none of the images is grouped in the incorrect clade and 100% non-watermarked images are correctly grouped in Group 5 (specificity). This evaluation assures the marvellous achievement of phylogenetic tree 3 for grouping images from G1 to G4 dataset in Group 5.

Table 6.13 The evaluations of grouping results of the non-watermarked images



		All non-watermarked degraded/non-degraded images present in Group 5		
		Present	Absent	
Test of Phylogenetic tree for Image grouping	Positive	True positive 96	False positive 0	Precision 1
	Negative	False negative 0	True Negative 96	Negative predictive value 1
		Sensitivity 1	Specificity 1	

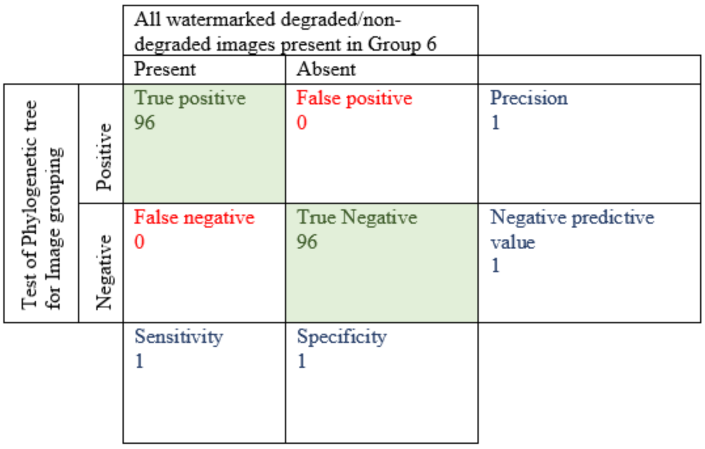
Group 6. The examination was accomplished for verifying the correct image grouping from Datasets G5 to G8 in Group 6.

Datasets G5 to G8 incorporate 96 watermarked and degraded/non-degraded images. It represents that 96 images will be associated with Group 6 and the remaining 96 images are owned by the non-watermarked image group out of 192 images from Datasets G1 to G8.

From the Table 6.14, the true positive and true negative value is 96. This indicates that all watermarked and degraded/non-degraded images are mapped correctly in Group 6, all other images are rejected correctly as displayed in Figure 6.9.

The zero value for the false positive and negative is shown in Table 6.14. This implies that none of the images is wrongly mapped and rejected incorrectly in Group 6 by using phylogenetic Tree 3, as shown in Figure 6.9.

Table 6.14 Evaluation of grouping results of watermarked images



		All watermarked degraded/non-degraded images present in Group 6		
		Present	Absent	
Test of Phylogenetic tree for Image grouping	Positive	True positive 96	False positive 0	Precision 1
	Negative	False negative 0	True Negative 96	Negative predictive value 1
		Sensitivity 1	Specificity 1	

Phylogenetic Tree 3, grouped 100% of watermarked and degraded/non-degraded images (sensitivity) correctly in Group 6 having 100% precision. NPV value is 100% that reasserts that none of the images is mapped in the wrong clade and 100% watermarked images are correctly grouped in Group 6 (specificity). This investigation established the exceptional capability of the phylogenetic Tree 3, for grouping images from G5 to G8 dataset in Group 6.

6.6 Discussions

The proposed syntactic BIIGA method for grouping images and specific dataset was tested by generating three phylogenetic trees: phylogenetic Tree 1 on Datasets G1, G2, G3 and G4 (see Tables 6.5 to 6.8), phylogenetic Tree 2 on Datasets G5 to G8 (see Tables 6.9 to 6.12) and phylogenetic Tree 3 on Datasets G1 to G8 (see Tables 6.13 and 6.14) using the BIIGA method. The experiments verify Question 3 as a whole; it is possible to extract syntactic patterns or signatures for grouping the watermarked (W) / non-watermarked (NW) images due to MPS by using biological representation, bioinformatics alignment algorithms and phylogenetic trees.

Additionally, with sub-question 3(a), the answer was supported by the first phylogenetic tree, it is possible to group NWD images and NWND images by using phylogenetic tree analysis (see Tables 6.5 to 6.8 and Figures 6.4 or 6.5) for sub-question 3 (b). The second phylogenetic tree justified that it is possible to group WD images and WND images by using phylogenetic tree (see Tables 6.9 to 6.12 and

Figures 6.6 or 6.7). Finally, sub-question 3(c) was answered by the third phylogenetic tree, it is possible to group watermarked /non-watermarked images from a mix of NWD, NWND, WD, and WND images by using phylogenetic tree analysis (see Tables 6.13 to 6.14 and Figures 6.8 or 6.9).

The proposed BIIGA approach grouped images in the expected categories: Datasets G1 to G4 and G5 to G8 in non-degraded / degraded Group 1, colour scanned mode in Group 3 and greyscale scanned images in Group 4. Furthermore, images from Datasets G1 to G8 were grouped in non-watermarked / watermarked images. Our analysis shows in Tables 6.5 to 6.8 for G1 to G4 datasets, Tables 6.9 to 6.12 for G5 to G8 datasets, Tables 6.13 to 6.14 for G1 to G8. The current image grouping, i.e. BIIGA successfully and consistently grouped images in required categories: (a) non-degraded and degraded images (b) watermarked / non-watermarked images. These results led to final verification of the hypothesis proposed in Section 6.2. i.e., the research hypothesis for watermarked/non-watermarked and degraded images, it is possible to identify syntactic structures, namely, patterns by using bioinformatics-based tools and techniques that help to determine whether a degraded / non-degraded image contains a type of watermark or specific degradation that group images in the expected categories.

The BIIGA method has significant concerns about whether it will group watermarked and degraded variants, whether it will group degraded images generated other than MPS. The proposed work unveils the requirement for novel software that can group degraded images after MPS with watermarks effectively. It indicates the need for developing specific approaches and grouping watermarked and degraded images. The future possibilities of this research are to implement a software system that will successfully group NWD, NWND, WD and WND images and then extend it for grouping different kinds of watermarked images (i.e., media / non-media watermarked), after degradation due to other than MPS.

6.7 Summary

Image grouping has a long history in data mining and machine learning, many methods exist, only a few of them have been introduced in this thesis. Nearly do all previous approaches extract features from the images, convert them into proper form, then use either data mining or neural networks to group or classify in the expected categories.

In this chapter, we have proposed a novel approach, i.e., BIIGA for automatically grouping degraded / non-degraded images and watermarked / non-watermarked images using bioinformatics-inspired techniques. We take into account the multiple sequence alignment role in grouping the images by using biologically-encoded images in DNA. For the first time, these bioinformatics-based tools and techniques were applied to the watermarked / non-watermarked and degraded images after MPS for the grouping. This indicates that bioinformatics-based tools and techniques were overlooked for the purposes of pattern matching in image grouping.

To conclude this part of the tools in bioinformatics, we raised the question of how bioinformatics tools could be used for pattern matching in image analysis at the start of this thesis; then, we implemented BIIIA, after that, BIIGA through bioinformatics-inspired image analysis by using pattern matching for the first time. The previous chapter and this chapter focused on inspiration from tools and techniques in bioinformatics for pattern matching that leads to a novel solution of image identification and grouping. Thus, the gap of bioinformatics-inspired image analysis has been filled. Also, we claim that bioinformatics-inspired BIIIA and BIIGA work reliably for watermarked/non-watermarked and degraded images. In the next chapter, the conclusion and future work will be stated.

Chapter 7

Conclusion and Future Work

This is the final chapter of this thesis. In Section 7.1, *we will deal with the evaluation of the research methodology employed in this thesis. In Section* 7.2*, we will explain the main contribution of this research work. Finally, in Section* 7.3*, this thesis will be closed with our visions of future research.*

7.1 Research Methodology Evaluation

In this thesis, we investigated bioinformatics-inspired image analysis as a way to resolve the issue of identifying and grouping the degraded and watermarked images as well as evaluation of image degradation. While most of the existing methods of image identification and grouping are not based on the biology-based encoding without extracting the features of the image; we have shown how our evaluation of watermarked and degraded is fulfilled. Encoding whole images biologically and applying alignment algorithms are effective in pattern matching for identifying and grouping the images degraded by MPS.

In our BIIIA and BIIGA algorithms, DNA was considered for biology-based encoding of images, sequence alignment was performed on these biologically-encoded images (i.e., local, global, and multiple sequence alignment) for pattern matching that helps in identifying and grouping the degraded images from MPS. Moreover, the BIIIA method was inspired from the idea of image analysis; the BIIGA method from the idea of image phylogeny and multiple sequence alignment is used to determine the evolutionary history of the species. To the best of our knowledge, the extracted syntactic patterns using bioinformatics tools and techniques was applied to deal with the image identification and grouping.

BIIIA and BIIGA algorithms both are served to resolve the issue of the image identification and grouping. In particular, BIIIA algorithm solved the image identification problem through not extracting image features, biology-based encoding of the DNA and SWA algorithm in pattern matching. For the BIIGA method, the image grouping problem was resolved through not extracting image features, biology-based encoding of images into the DNA, multiple sequence alignment, and phylogenetic trees. Therefore, the research question: "is it possible to use bioinformatics-based tools and techniques for pattern matching in image analysis?" has been answered.

7.2 Summary of Contributions

This PhD thesis has shown how evaluation of the watermarked and degraded images as well as bioinformatics-inspired image analysis can be adopted in identifying and grouping the images. Pattern matching is the critical aspect for image identification and grouping but there was a gap: how bioinformatics tools and techniques could be employed to pattern matching. We have filled this gap by introducing bioinformatics-inspired image analysis for degraded images from MPS and exploring a method between *pattern matching* and *bioinformatics tools*. The idea using bioinformatics tools for pattern matching with the issue “watermarked/non-watermarked and degraded/non-degraded image identification and grouping” helped in the image analysis of the degraded images from MPS. BIIIA permits identification of the image, i.e., how we can identify the watermarked / non-watermarked and degraded images; how we can allow ourselves further to study on image analysis that led to BIIGA algorithm for grouping images, allow us to separate the degraded from non-degraded images as well as the watermarked from non-watermarked images.

The purpose of this thesis has been to develop a novel image identification algorithm: BIIIA as well as a novel image grouping algorithm: BIIGA for image identification and grouping. The contributions are summarized below:

(i) Evaluation of image degradation by using the eight image metrics, i.e., CC, Bias, ERGAS, RMSE, RASE, Q, SSIM, and DSSIM.

(ii) BIIIA algorithm

- Successfully detected watermarked / non-watermarked images by using a biological representation of the image, sequence alignment algorithms, and pattern matching.
- SWA is better than NWA for biologically-represented images.
- DNA-based representation of images is found suitable for BIIIA.

(iii) BIIGA contributions

- Successfully use phylogenetic trees for grouping original / degraded images as well as watermarked / non-watermarked images.
- Successfully group watermarked / non-watermarked images by using a phylogenetic tree.

7.3 Our Vision

In this section, we will explain the future directions in identifying and grouping the degraded images generated from MPS with their original variants. However, the issue of evaluation, grouping and identification of watermarked / non-watermarked images is very challenging due to lack of data of the degraded images because of MPS. Although this thesis has proposed novel methods for MPS degradation, evaluation, identification and grouping for degraded and watermarked / non-watermarked images, there are some open research problems and limitations that require to be further examined and solved in future.

7.3.1 Limitations

Our interest, including BIIIA and BIIGA algorithms, was in the identification of watermarked images by using a biological representation of the image, bioinformatics alignment algorithms and pattern matching; but we did not take into account the rapid evolution of other forms of watermarks, such as non-multimedia watermarks and those in the cloud. Furthermore, we did not consider the identification problem of degraded and watermarked images. Establishing such a big dataset with watermarked and degraded / non-degraded images will allow us to check the robustness of the proposed approach. We assume that our approach is to extract the essential aspects of an image, the approach narrated in this thesis may apply to other types of watermarked images.

7.3.2 Future researh

The overview of developing BIIIA and BIIGA algorithms was to show that bioinformatics-based tools and techniques can be used for pattern matching in image analysis. A big part of the inspiration for this research project was to integrate the tools and techniques in bioinformatics for image analysis. Correspondingly, our research direction is to implement a software that will automatically identify watermarked images and then extend it for identifying the watermarks after MPS. The possible techniques of this thesis will be applied to:

(1) Digital Forensics

- Source camera identification. In digital forensics, one of the most interesting problem for a given digital image is to identify the camera model that was used to click the image. The state-of-the-art technologies for source camera identification will depend on the sensor-based noise residues, image features on spatial / frequency domain, etc. The literature used different methods for source camera identification that includes enhancing sensor pattern noise (Li, 2010), deep learning (Freire-Obregon, Narducci, Barra, & Castrillon-Santana, 2017), support vector machine (Wang, Kong, & You, 2009), intrinsic lens radial distortion (Choi, Lam, & Wong, 2006), a set of image features with classifier (Kharrazi, Sencar, & Memon, 2004).

- Identification of pirated videos. Forensic police want to know that how many and which videos from Netflix or Amazon prime videos are pirated. This piracy identification can be completed by employing the BIIIA. It will extract the syntactic patterns from combination of Netflix and Amazon videos. These patterns will help to filter out the videos that belongs to Netflix and Amazon video out of billions of videos. Further, individual pattern from Netflix or Amazon video helps to identify the videos from a mix of Netflix and Amazon video.

(2) Digital Rights Management

- Different watermarking algorithms. The proposed BIIIA approach can be employed to different types of watermarking algorithms for further research.
- Variety of watermark. A huge amount of media/non-media watermarks are available that can be embedded in the images, BIIIA and BIIGA algorithms will be employed for further research.
- Filtering of images ownership. Suppose a company is looking for images watermarked using its logo and classified it as highly confidential. By employing BIIIA approach, we can extract a pattern from these images. From a mix of billions of images, it can give us a number that how much or which one is embedded with the logo with highly confidentiality. Secondly, by applying BIIGA algorithm, we may generate a phylogenetic tree that will group the images in their respective category.
- Videos ownership identification. BIIIA algorithm can be extended for identifying the video ownership by proposing another novel approach, named as “bioinformatics inspired video identification approach (BIVIA)”. This approach can be realised by converting a watermarked video frame into DNA and extracting the pattern for video that will be used for identification of the ownership of video.

(3) Future research in BIIIA and BIIGA

- Colour images as the test dataset. Our approach tested only on the greyscale image – based standard test datasets; the colour images can be considered as new dataset for further experimentation on BIIIA and BIIGA.

- Protein modelling of images. DNA-encoded images can be mapped into proteins for further research in image analysis. That will open a new research chapter that will fill up another gap between the bioinformatics and image analysis.

The proposed research in image analysis will be of extremely exciting and novel to the image analyst and forensic experts because of its high traffic volume of digital images; the Internet availability increases the vulnerability of digital piracy and counterfeiting of images. To the best of my knowledge, the proposed BIIIA and BIIGA in this thesis are unique so far; the only research reported in the world recently analyse images using the state-of-the-art of bioinformatics-based sequence alignment for pattern matching.

Future research includes developing an image analysis software that can automatically identify the ownerships of images and generate a phylogenetic tree that will group the images in the expected category. There may be chances for attracting collaboration with multinational companies like Netflix, Microsoft, etc.

7.3.3 Reflection

The core idea for BIIGA and BIIIA approaches is that the images are able to be mutated as living beings (e.g., animals, plants, etc.) in biology. The evolution of an image is thought of as degradation from MPS. The print-and-scan degradation of images is considered as a mutating agent in image grouping. In living organisms, a mutating agent changes the genes of animals or plants. In the same way, the MPS operations modify the pixels of these images. This thesis successfully identified and grouped the images on the basis of these mutations. MPS operations mutate these images by using three scanning modes: black and white (BW), colour and greyscale. Successfully grouping these mutated images into the expected categories is performed by using phylogenetic trees. Successful identification was achieved for these mutated images.

References

Abramoff, M. D., Magalhaes, P. J., & Ram, S. J. (2004). Image Processing with ImageJ. *Biophotonics International, 11*(4), 36-42.

Adleman, L. (1994). Molecular Computation of Solutions to Combinatorial Problems. *Science, 266*(5187), 1021-1024.

Adler, R. L., Kitchens, B. P., Martens, M., Tresser, C. P., & Wu, C. W. (2003). The Mathematics of Halftoning. *IBM Journal of Research and Development, 47*(1), 5-15.

Alex, U., Oswaldo, T., Antonio, J., Cornejo, G., & Perkins, J. R. (2016). Review: High-performance Computing to Detect Epistasis in Genome Scale Data Sets. *Briefings in Bioinformatics, 17*(3), 368-379.

Altschul, S. F., Warren, G., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic Local Alignment Search Tool. *Journal of Molecular Biology, 215*(3), 403-410.

Bacon, D. J., & Anderson, W. F. (1986). Multiple Sequence Alignment. *Journal of Molecular Biology, 191*(2), 153-161.

Baird, H., & Chaudhuri, B. (2007). The State of the Art of Document Image Degradation Modelling. *Springer London: Digital Document Processing*, 261-279.

Bicego, M., & Lovato, P. (2012). 2D Shape Recognition Using Biological Sequence Alignment Tools. *International Conference on Pattern Recognition (ICPR2012),* (pp. 1359-1362).

Bicego, M., & Lovato, P. (2016). A Bioinformatics Approach to 2D Shape Classification. *Computer Vision and Image Understanding, 145*, 59-69.

Bicego, M., Danese, S., Melzi, S., & Castellani, U. (2015). A Bioinformatics Approach to 3D Shape Matching. *European Conference on Computer Vision,* (pp. 313-325). Zurich, Switzerland.

Boc, A., Diallo, A. B., & Makarenkov, V. (2012). T-REX: a Web Server for Inferring, Validating and Visualizing Phylogenetic Trees and Networks. *Nucleic Acids Research*, *40*(W1), W573-W579.

Bornholt, J., Lopez, R., Carmean, D., Ceze, L., Seelig, G., & Strauss, K. (2016). A DNA-Based Archival Storage System. *International Conference on Architectural Support for Programming Languages and Operating Systems,* (pp. 637-649). Atlanta, Georgia, USA.

Bovik, A. C. (2009). *The Essential Guide to Image Processing.* Academic Press .

Brassil, J., Low, S., Maxemchuk, N., & O'Gorman, L. (1994). Electronic Marking and Identification Techniques to Discourage Document Copying. *Infocom.*

Bronstein, A. M., Bronstein, M. M., & Kimmel, R. (2010). The Video Genome. *CoRR, abs/1003.5320*.

Brown, N. P., Leroy, C., & Sander, C. (1998). MView: A Web Compatible Database Search or Multiple Alignment Viewer. *Bioinformatics, 14*(4), 380.

Brown, T. (2002). Molecular Phylogenetics. In *Genomes. 2nd edition.* Wiley-Liss: Oxford.

Bulan, O., Mao, J., & Sharma, G. (2009). Geometric Distortion Signatures for Printer Identification. *IEEE International Conference on Acoustics, Speech and Signal Processing,* (pp. 1401-1404). Taipei.

Cadik, M., Herzog, R., Mantiuk, R., Myszkowski, K., & Seidel, H.-P. (2012). New Measurements Reveal Weaknesses of Image Quality Metrics in Evaluating Graphics Artifacts. *ACM Transactions on Graphics (TOG), 31*(6).

Caronni, G. (1995). Assuring Ownership Rights for Digital Images. *Reliable IT Systems.*

Carpenter, A. E., Jones, T. R., Lamprecht, M. R., Clarke, C., Kang, I. H., Friman, O., . . . Sabatini, D. M. (2006). CellProfiler: Image Analysis Software for Identifying and Quantifying Cell Phenotypes. *Genome Biology, 7*(10), R100.

Cavalli-Sforza, L. L., & Edwards, A. W. (1967). Phylogenetic Analysis: Models and Estimation Procedures. *Evolution, 21*(3), 550-570.

Chambah, M., Ouni, S., Herbin, M., & Zagrouba, E. (2009). Toward an Automatic Subjective Image Quality Assessment System. In *SPIE, IS&T, Electronic Imaging 09, Image Quality and System Performance* (p. 12). San Jose, USA.

Chaumont, F. D., Dallongeville, S., Chenouard, N., Herve, N., Pop, S., Provoost, T., . . . Olivo-Marin, J. C. (2012). Icy: an Open Bioimage Informatics Platform for Extended Reproducible Research. *Nature Methods, 9*(7), 690-696.

Chen, J.-J., & Wang, J.-F. (2016). WEBSAP: Study and Build of Web-based Sequence Alignment Platform. *International Conference on Machine Learning and Cybernetics (ICML),* (pp. 331-336).

Chen, L., Feris, R., & Turk, M. (2008). Efficient Partial Shape Matching using Smith-Waterman Algorithm. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops,* (pp. 1-6).

Chen, Y.-J., Dalchau, N., Srinivas, N., Phillips, A., Cardelli, L., Soloveichik, D., & Seelig, G. (2013). Programmable Chemical Controllers made from DNA. *Nature Nanotechnology, 8*(10), 755-762.

Chetouani, A., Beghdadi, A., & Deriche, M. (2010). Statistical Modeling of Image Degradation Based on Quality Metrics. *International Conference on Pattern Recognition,* (pp. 714-717).

Chiang, P. -J., Khanna, N., Mikkilineni, A. K., Segovia, M. V., Allebach, J. P., Chiu, G. T., & Delp, E. J. (2010). Printer and Scanner Forensics: Models and Methods. *Intelligent Multimedia Analysis for Security Applications Springer Berlin Heidelberg: Berlin, Heidelberg, 282,* 145-187.

Chiang, P.-J., Khanna, N., Mikkilineni, A. K., Segovia, M. V., Suh, S., Allebach, J. P., Delp, E. J. (2009). Printer and Scanner Forensics. *IEEE Signal Processing Magazine, 26*(2), 72-83.

Choi, K. S., Lam, E. Y., & Wong, K. K. (2006). Automatic source camera identification using the intrinsic lens radial distortion. *Optics Express, 14*(24), 11551-11565.

Chowdhury, B., & Garai, G. (2014). A Genetic Approach with Controlled Crossover and Guided Mutation for Biological Sequence Alignment. *Fourth International Conference of Emerging Applications of Information Technology,* (pp. 307-312). Kolkata.

Church, G. M., Gao, Y., & Kosuri, S. (2012). Next-Generation Digital Information Storage in DNA. *Science, 337*(6102), 1628.

Clelland, C. T., Risca, V., & Bancroft, C. (1999). Hiding Messages in DNA Microdots. *Nature, 399*(6736), 533-534.

Cohen, J. (2004). Bioinformatics—An Introduction for Computer Scientists. *ACM Computing Surveys, 36*(2), 122-158.

Collins, T. J. (2007). ImageJ for Microscopy. *BioTechniques, 43*(S1), S25-S30.

Costa, F. D., Oikawa, M. A., Dias, Z., Goldenstein, S., & Rocha, A. R. (2014). Image Phylogeny Forests Reconstruction. *IEEE Transactions on Information Forensics and Security, 9*(10), 1533-1546.

Cox, I. J., Miller, M. L., & Bloom, J. A. (2002). *Digital Watermarking.* San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Cox, I., & Miller., M. (2002). The First 50 Years of Electronic Watermarking. *EURASIP Journal on Advances in Signal Processing, 2,* 1-7.

Cox, I., Miller, M., Bloom, J., & Honsinger, C. (2002). Digital Watermarking. *Journal of Electronic Imaging, 11*(3).

Cox, I., Miller, M., Bloom, J., Fridrich, J., & Kalker, T. (2008). *Digital Watermarking and Steganography.* San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Cox, J. P. (2001). Long-Term Data Storage in DNA. *Trends in Biotechnology, 19*(7), 247-250.

Darwin, C. (1859). *On the Origin of the Species by Means of Natural Selection: or, The Preservation of Favoured Races in the Struggle for Life.* London: J. Murray.

Davis, J. (1996). Microvenus. *Art Journal, 55*(1), 70-74.

Dias, Z., Goldenstein, S., & Rocha, A. (2013). Exploring Heuristic and Optimum Branching Algorithms for Image Phylogeny. *Journal of Visual Communication and Image Representation, 24*(7), 1124-1134.

Dias, Z., Goldenstein, S., & Rocha, A. (2013). Large-Scale Image Phylogeny: Tracing Image Ancestral Relationships. *IEEE MultiMedia, 20*(3), 58-70.

Dias, Z., Goldenstein, S., & Rocha, A. (2013). Toward Image Phylogeny Forests: Automatically Recovering Semantically Similar Image Relationships. *Forensic Science International, 231*(1-3), 178-189.

Dias, Z., Rocha, A., & Goldenstein, S. (2010). First Steps Toward Image Phylogeny. *IEEE International Workshop on Information Forensics and Security*, (pp. 1-6). Seattle, WA.

Dias, Z., Rocha, A., & Goldenstein, S. (2011). Video Phylogeny: Recovering Near-Duplicate Video Relationships. *IEEE International Workshop on Information Forensics and Security*, (pp. 1-6).

Dias, Z., Rocha, A., & Goldenstein, S. (2012). Image Phylogeny by Minimal Spanning Trees. *IEEE Transactions on Information Forensics and Security, 7*(2), 774-788.

Dinh, A., Brill, D., Li, Y., & He, W. (2016). Malware Sequence Alignment. *IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom) (BDCloud-SocialCom-SustainCom),* (pp. 613-617). Atlanta, GA.

Do, C. B., Mahabhashyam, M. S., Brudno, M., & Batzoglou, S. (2005). ProbCons: Probabilistic Consistency-based Multiple Sequence Alignment. *Genome Research, 15*(2), 330-340.

Dorigo, M., Birattari, M., & T. Stutzle. (2006). Ant Colony Optimization: Artificial Ants as a Computational Intelligence Technique. *IEEE, Computational Intelligence Magazine, 1*, 28-39.

Drozda, P., Gorecki, P., Sopyla, K., & Artiemjew, P. (2013). Visual Words Sequence Alignment for Image Classification. *IEEE International Conference on Cognitive Informatics and Cognitive Computing*, (pp. 397-402). New York, NY.

Drozda, P., Sopyla, K., & Gorecki, P. (2014). Different Orderings and Visual Sequence Alignment Algorithms for Image Classification. *International Conference on Artificial Intelligence and Soft Computing (ICAISC 2014),* (pp. 693-702). Zakopane, Poland.

Dufour, A., Shinin, V., Tajbakhsh, S., Guillen-Aghion, N., Olivo-Marin, J., & Zimmer, C. (2005). Segmenting and Tracking Fluorescent Cells in Dynamic 3-D microscopy with Coupled Active Surfaces. *IEEE Transactions on Image Processing, 14*(9), 1396-1410.

Durbin, R. M., Eddy, S. R., Krogh, A., & Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids (1st ed.).* Cambridge: Cambridge University Press.

Edgar, R. C. (2004). MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput. *Nucleic Acids Research, 32*(5), 1792-1797.

Edwards, S. V. (2009). Is a New and General Theory of Molecular Systematics Emerging? *Evolution, 63*(1), 1-19.

Eliceiri, K. W., Berthold, M. R., Goldberg, I. G., Ibanez, L., Manjunath, B. S., Martone, M. E., . . . Carpenter, A. E. (2012). Biological Imaging Software Tools. *Nature Methods, 9,* 697-710.

Escobar, I., Hidalgo, N., Inostroza-Ponta, M., Marín, M., Rosas, E., & Dorn, M. (2016). Evaluation of a Combined Energy Fitness Function for a Distributed Memetic Algorithm to Tackle the 3D Protein Structure Prediction Problem. *International Conference of the Chilean Computer Science Society (SCCC),* (pp. 1-10).

Eskicioglu, A. M., & Fisher, P. S. (1995). Image Quality Measures and Their Performance. *IEEE Transactions on Communications, 43*(12), 2959-2865.

Falk, T. H., Guo, Y., & Chan, W.-Y. (2007). Improving Robustness of Image Quality Measurement with Degradation Classification and Machine Learning. *Asilomar Conference on Signals, Systems and Computers,* (pp. 503-507).

Felsenstein, J. (1981). Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach. *Journal of Molecular Evolution, 17*(6), 368-376.

Felsenstein, J. (1988). Phylogenies from Molecular Sequences: Inference and Reliability. *Annual Review of Genetics, 22*(1), 521-565.

Freddolino, P. L., Park, S., Roux, B., & Schulten, K. (2009). Force Field Bias in Protein Folding Simulations. *Biophysical Journal, 96*(9), 3772-3780.

Freire-Obregon, D., Narducci, F., Barra, S., & Castrillon-Santana, M. (2017). Deep learning for source camera identification on mobile devices. eprint arXiv:1710.01257.

Gajalakshmi, G., & Srikanth, G. U. (2016). A Survey on the Utilization of Ant Colony Optimization (ACO) Algorithm in WSN. *International Conference on Information Communication and Embedded Systems (ICICES).* Chennai.

Gallager, R. G. (2001). Claude E. Shannon: a Retrospective on His Life, Work, and Impact. *IEEE Transactions On Information Theory, 47*(7), 2681-2695.

Gao, X., Lu, W., Tao, D., & Li, X. (2009). Image Quality Assessment Based on Multiscale Geometric Analysis. *IEEE Transactions on Image Processing, 18*(7), 1409-1423.

Garhwal, A. S., & Yan, W. Q. (2015). Evaluations of Image Degradation from Multiple Scan-Print. *International Journal of Digital Crime and Forensics (IJDCF), 7*(4), 55-65.

Gaubatz, M. D., & Simske, S. J. (2009). Printer-Scanner Identification via Analysis of Structured Security Deterrents. *IEEE International Workshop on Information Forensics and Security (WIFS),* (pp. 151-155). London, UK.

Gayathri, B. K., & Raajan, P. (2016). A Survey of Breast Cancer Detection Based on Image Segmentation Techniques. *International Conference on Computing Technologies and Intelligent Data Engineering,* (pp. 1-5).

George, M., Church, Y. G., & Kosuri, S. (2012). Next-Generation Digital Information Storage in DNA. *Science, 337*(6102), 1628.

Gibson, D. G., Glass, J. I., Lartigue, C., Noskov, V. N., Chuang, R.-Y., Algire, M. A., Venter, J. C. (2010). Creation of a Bacterial Cell Controlled by a Chemically Synthesized Genome. *Science, 329*(5987), 52-56.

Goldman, N., Bertone, P., Chen, S., Dessimoz, C., LeProust, E. M., Sipos, B., & Birney, E. (2013). Towards Practical, High-Capacity, Low-Maintenance Information Storage in Synthesized DNA. *Nature, 494*(7435), 77-80.

Gotoh, O. (1982). An Improved Algorithm for Matching Biological Sequences . *Journal of Molecular Biology, 162*, 705-708.

Grass, R. N., Heckel, R., Puddu, M., Paunescu, D., & Stark, W. J. (2015). Robust Chemical Preservation of Digital Information on DNA in Silica with Error-Correcting Codes. *Angewandte Chemie International Edition, 54*(8), 2552-2555.

Gray, R. D., Drummond, A. J., & Greenhill, S. J. (2009). Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement. *Science, 323*(5913), 479-483.

Hagen, J. (2000). The Origins of Bioinformatics. *Nat Rev Genet, 1*, 231-236.

Hase, H. (2011). Quality Evaluation of Character Image Database and Its Application. *International Conference on Document Analysis and Recognition*, (pp. 1414-1418). Beijing, China.

Heng, L., & Durbin, R. (2011). Inference of Human Population History from Individual Whole-Genome Sequences. *Nature, 475*(7357), 493-496.

Henikoff, S., & Henikoff, J. G. (1992). Amino Acid Substitution Matrices from Protein Blocks. *Proceedings of the National Academy of Sciences of the United States of America, 89*(22), 10915–10919.

Hennig, W. (1966). *Phylogenetic Systematics*. Urbana, IL: University of Illinois Press.

Hillis, D. M. (1997). Biology Recapitulates Phylogeny. *Science, 276*(5310), 218-219.

Hogeweg, P. (1978). Simulating the Growth of Cellular Forms. *Simulation, 31*, 90-96.

Hogeweg, P. (2011). The Roots of Bioinformatics in Theoretical Biology. *PLoS Computational Biology, 7*(3).

Hogeweg, P., & Hesper, B. (1978). Interactive instruction on population interactions. *Comput Biol Med, 8,* 319-327.

Homola, T., Dohnal, V., & Zezula, P. (2011). Searching for Sub-images Using Sequence Alignment. *IEEE International Symposium on Multimedia*, (pp. 61-68).

Jian, M. (2011). Reconstructing the History of Large-Scale Genomic Changes: Biological Questions and Computational Challenges. *Journal of Computational Biology, 18*(7), 879-893.

Jiawei Han, Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques, Third Edition (The Morgan Kaufmann Series in Data Management Systems) 3rd edition.* Netherland: Morgan Kaufmann.

Joly, A., Buisson, O., & Frelicot, C. (2007). Content-Based Copy Retrieval Using Distortion-Based Probabilistic Similarity Search. *IEEE Transactions on Multimedia, 9*(2), 293-306.

Kalker, T. (2001). Considerations on Watermarking Security. *IEEE Workshop on Multimedia Signal Processing*, (pp. 201-206).

Kang, X., Huang, J., & Zeng, W. (2010). Efficient General Print-Scanning Resilient Data Hiding Based on Uniform Log-Polar Mapping. *IEEE Transactions on Information Forensics and Security, 5*(1), 1-12.

Kankaanpaa, P., Paavolainen, L., Tiitta, S., Karjalainen, M., Paivarinne, J., Nieminen, J., . . . White, D. J. (2012). BioImageXD: An Open, General-Purpose and High-Throughput Image-Processing Platform. *Nature Methods, 9*(7), 683-689.

Kanungo, T., Haralick, R. M., Baird, H. S., Stuezle, W., & Madigan, D. (2000). A Statistical, Nonparametric Methodology for Document Degradation Model Validation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(11), 1209-1223.

Katoh, K., & Frith, M. C. (2012). Adding Unaligned Sequences into an Existing Alignment using MAFFT and LAS. *Bioinformatics, 28*(23), 3144-3146.

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution, 30*(4), 772-780.

Katoh, K., & Standley, D. M. (2016). A simple method to control over-alignment in the MAFFT multiple sequence alignment program. *Bioinformatics, 32*(13), 1933-1942. doi:https://doi.org/10.1093/bioinformatics/btw108

Katoh, K., & Standley, D. M. (2016). A Simple Method to Control Over-Alignment in the MAFFT Multiple Sequence Alignment Program. *Bioinformatics, 32*(13), 1933-1942.

Katoh, K., & Toh, H. (2007). PartTree: an Algorithm to Build an Approximate Tree from a Large Number of Unaligned Sequences. *Bioinformatics, 23*(3), 372-374.

Katoh, K., & Toh, H. (2008). Improved Accuracy of Multiple ncRNA Alignment by Incorporating Structural Information into a MAFFT-based Framework. *BMC Bioinformatics, 9*(1), 212.

Katoh, K., & Toh, H. (2008). Recent Developments in the MAFFT Multiple Sequence Alignment Program. *Briefing in Bioinformatics, 9*(4), 286-298.

Katoh, K., & Toh, H. (2010). Parallelization of the MAFFT Multiple Sequence Alignment Program. *Bioinformatics, 26*(15), 1899-1900.

Katoh, K., Asimenos, G., & Toh, H. (2009). Multiple Alignment of DNA Sequences with MAFFT. *Bioinformatics for DNA Sequence Analysis, 537*, 39-64.

Katoh, K., Kuma, K.-I., Toh, H., & Miyata, T. (2005). MAFFT Version 5: Improvement in Accuracy of Multiple Sequence Alignment. *Nucleic Acids Research, 33*(2), 511-518.

Katoh, K., Misawa, K., Kuma, K.-I., & Miyata, T. (2002). MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform. *Nucleic Acids Research, 30*(14), 3059-3066.

Kauffmann, A., Gentleman, R., & Huber, W. (2009). arrayQualityMetrics—a Bioconductor Package for Quality Assessment of Microarray Data. *Bioinformatics, 25*(3), 415-416.

Kaur, H., & Chand, L. (2016). Biological Sequence Alignment using Varied Optimization Algorithms. *International Conference on Inventive Computation Technologies,* (pp. 1-5).

Kaur, H., & Chand, L. (2016). Pairwise Sequence Alignment of Biological Database using Soft Computing Approach. *International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)*, (pp. 72-77).

Khan, A., Saha, G., & Pal, R. K. (2017). Quantum Computing Based Inference of GRNs. *International Conference on Bioinformatics and Biomedical Engineering.* Granada, Spain.

Khanna, N., & Delp, E. J. (2010). Intrinsic Signatures for Scanned Documents Forensics : Effect of Font Shape and Size. *IEEE International Symposium on Circuits and Systems,* (pp. 3060-3063). Paris,France.

Kharrazi, M., Sencar, H., & Memon, N. (2004). Blind source camera identification. *International Conference on Image Processing.* Piscataway, NJ, USA.

Kiah, H. M., Puleo, G. J., & Milenkovic, O. (2015). Codes for DNA Storage Channels. *IEEE Information Theory Workshop (ITW),* (pp. 1-5).

Kim, H. Y., & Mayer, J. (2007). Data Hiding for Binary Documents Robust to Print-Scan, Photocopy and Geometric Distortions. *Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI 2007)*, (pp. 105-112).

Kim, H.-S., Chang, H.-W., Lee, J., & Lee, D. (2010). BASIL: Effective Near-Duplicate Image Detection Using Gene Sequence Alignment. *European Conference on Advances in Information Retrieval.* Milton Keynes, UK.

Kim, H.-S., Chang, H.-W., Liu, H., Lee, J., & Lee, D. (2009). BIM: Image Matching using Biological Gene Sequence Alignment. *IEEE International Conference on Image Processing (ICIP)*, (pp. 205-208).

Kremer, J. R., Mastronarde, D. N., & McIntosh, J. (1996). Computer Visualization of Three-Dimensional Image Data using IMOD. *Journal of Structural Biology, 116*(1), 71-76.

Kremer, J., Stensbo-Smidt, K., Gieseke, F., Pedersen, K. S., & Igel, C. (2017). Big Universe, Big Data: Machine Learning and Image Analysis for Astronomy. *IEEE Intelligent Systems, 32*(2), 16-22.

Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for Bigger Datasets. *Molecular Biology and Evolution, 33*(7), 1870-1874.

Kumar, S., Stecher, G., Peterson, D., & Tamura, K. (2012). MEGA-CC: Computing Core of Molecular Evolutionary Genetics Analysis Program for Automated and Iterative Data Analysis. *Bioinformatics, 28*(20), 2685-2686.

Kumar, S., Tamura, K., & Nei, M. (1994). MEGA: Molecular Evolutionary Genetics Analysis Software for Microcomputers. *Bioinformatics*, *10*(2), 189-191.

Kumar, S., Tamura, K., & Nei, M. (2004). MEGA3: Integrated Software for Molecular Evolutionary Genetics Analysis and Sequence Alignment. *Briefings in Bioinformatics, 5*(2), 150-160.

Kumar, S., Tamura, K., Jakobsen, I. B., & Nei, M. (2001). MEGA2: Molecular Evolutionary Genetics Analysis Software. *Bioinformatics*, *17*(12), 1244-1245.

Kun, X., Luxmoore, A., & Davies, T. (1998). Sewer Pipe Deformation Assessment by Image Analysis of Video Surveys. *Pattern Recognition*, *31*(2), 169-180.

Kuraku, S., Zmasek, C. M., Nishimura, O., & Katoh, K. (2013). aLeaves Facilitates on-demand Exploration of Metazoan Gene Family Trees on MAFFT Sequence Alignment Server with Enhanced Interactivity. *Nucleic Acids Research, 41*(W1), W22-W28.

Lassmann, T., Frings, O., & Sonnhammer, E. L. (2009). Kalign2: High-performance Multiple Alignment of Protein and Nucleotide Sequences Allowing External Features. *Nucleic Acids Research, 37*(3), 858-865.

Lee, H.-L., & Chen, L.-H. (2016). A Novel Printable Watermarking Method in Dithering Halftone Images. *Advances in Multimedia, 2016*, 17.

Lee, J.-H., & Allebach, J. P. (2005). Inkjet Printer Model-based halftoning. *IEEE Transactions on Image Processing, 14*(5), 647-689.

Leier, A., Richter, C., Banzhaf, W., & Rauhe, H. (2000). Cryptography with DNA Binary Strands. *Biosystems, 57*(1), 13-22.

Li, C.-T. (2010). Source Camera Identification Using Enhanced Sensor Pattern Noise. *IEEE Transactions on Information Forensics and Security*, *5*(2), 280-287.

Libeskind-Hadas, R., & Bush, E. (2013). A First Course in Computing with Applications to Biology. *Briefing in bioinformatics, 14*(3), 610-617.

Limbachiya, D., Dhameliya, V., Khakhar, M., & Gupta, M. K. (2015). On Optimal Family of Codes for Archival DNA Storage. *International Workshop on Signal Design and its Applications in Communications.*

Lin, C.-Y., & Chang, S.-F. (1999). Distortion Modeling and Invariant Extraction for Digital Image Print-and-Scan Process. *International Symposium on Multimedia Information Processing (ISMIP 99).* Taipei, Taiwan.

Lin, G., Adiga, U., Olson, K., Guzowski, J., Barnes, C., & Roysam, B. (2003). A Hybrid 3D Watershed Algorithm Incorporating Gradient Cues and Object Models for Automatic Segmentation of Nuclei in Confocal Image Stacks. *Cytometry Part A*, *56*(1), 23-36.

Lizhen, S., Zhong, W., Weikuan, Y., & Meng, X. (2017). A Case Study of Tuning MapReduce for Efficient Bioinformatics in the Cloud. *Parallel Computing, 61*, 83-95.

Lobo, I. (2008). Basic Local Alignment Search Tool (BLAST). *Nature Education, 1*(1), 215.

Loce, R., & Lama, W. (1990). Halftone Banding due to Vibrations in a Xerographic Image Bar Printer. *Journal of Imaging Technology*, *16*(1), 6-11.

Lopes, P. E., Guvench, O., & MacKerell, A. D. (2015). Current Status of Protein Force Fields for Molecular Dynamics. *Methods in Molecular Biology (Clifton, N.J.), 1215*, 47-71.

Lovato, P., & Bicego, M. (2012). 2D Shapes Classification Using BLAST. *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR),* (pp. 273-281). Hiroshima, Japan.

Lovato, P., Milanese, A., Centomo, C., Giorgetti, A., & Bicego, M. (2014). S-BLOSUM: Classification of 2D Shapes with Biological Sequence Alignment. *International Conference on Pattern Recognition*, (pp. 2335-2340).

Loytynoja, A., & Goldman, N. (2010). webPRANK: a Phylogeny-aware Multiple Sequence Aligner with Interactive Alignment Browser. *BMC Bioinformatics, 11*(1), 579.

Luscombe N, M., Greenbaum, D., & Gerstein, M. (2001). What is Bioinformatics? An Introduction and Overview. *IMIA Yearbook 2001*, 83-99.

MacAulay, C., & Palcic, B. (1988). A Comparison of Some Quick and Simple Threshold Selection Methods for Stained Cells. *Analytical and Quantitative Cytology and Histology, 10*(2), 134-138.

Mantas, J. (1987). Methodologies in Pattern Recognition and Image Analysis—A Brief Survey. *Pattern Recognition, 20*(1), 1-6.

Marvel, L. M., Boncelet, C. G., & Retter, C. T. (1999). Spread Spectrum Image Steganography. *IEEE Transactions on Image Processing, 8*(8), 1075-1083.

Maser, P., Thomine, S., Schroeder, J. I., Ward, J. M., Hirschi, K., Sze, H., . . . Guerinot, M. L. (2001). Phylogenetic Relationships within Cation Transporter Families of Arabidopsis. *Plant Physiology, 126*(4), 1646-1667.

Mathkour, H., & Ahmad, M. (2009). Genome Sequence Analysis: A Survey. *Journal of Computer Science, 5*(9), 651-660.

Matsui, K., & Tanaka, K. (1994). Video-steganography. *IMA Intellectual Property Project,, 1*, pp. 187-206.

Michener, C. D., & Sokal, R. R. (1957). A Quantitative Approach to a Problem in Classification. *Evolution, 11*(2), 130-162.

Michiels, B., & Macq, B. (2006). Benchmarking Image Watermarking Algorithms with Openwatermark. *European Signal Processing Conference*, (pp. 1-5). Florence, Italy.

Michler, G. H. (2008). Image Processing and Image Analysis. In *Electron Microscopy of Polymers* (pp. 161-171). Heidelberg: Springer Berlin Heidelberg.

Moghaddam, C. M. (2009). Low Quality Document Image Modeling and Enhancement. *International Journal of Document Analysis Recognition,Berlin, Heidelberg, Springer, 11*, 180-201.

Nagel, H. H. (1978). Analysis Techniques for Image Sequences. *Inter. J. Conf. on Pattern Recognition.* Kyoto.

Nagel, H. H. (1983). Overview on Image Sequence Analysis. *Huang T.S. (eds) Image Sequence Processing and Dynamic Scene Analysis. NATO ASI Series (Series F: Computer and System Sciences), 2*, 2-39.

Nahdliyah, S. D., Fitriyani, N., & Biyanto, T. R. (2017). Optimization of CO2 Contents and Energy Saving on Sweetening Gas Processing Plant using Particle Swarm Optimization (PSO) Algorithm. *International Annual Engineering Seminar (InAES)*, (pp. 250-55). Yogyakarta, Indonesia.

Naidu, V., & Narayanan, A. (2014). Further Experiments in Biocomputational Structural Analysis of Malware. *IEEE International Conference on Natural Computation (ICNC)*, (pp. 605-610).

Naidu, V., & Narayanan, A. (2016). Needleman-Wunsch and Smith-Waterman Algorithms for Identifying Viral Polymorphic Malware Variants. *IEEE International Conference on Dependable, Autonomic and Secure Computing (DASC).* Auckland, New Zealand.

Naidu, V., & Narayanan, A. (2016). Syntactic Approach for Detecting Viral Polymorphic Malware Variants. *Pacific Asia Workshop on Intelligence and Security Informatics.* Auckland, New Zealand.

Naidu, V., & Narayanan, A. (2016). The Effects of Using Different Substitution Matrices in a String-Matching Technique for Identifying Viral Polymorphic Malware Variants. *IEEE Congress on Evolutionary Computation (WCCI - IEEE CEC).* Vancouver, Canada.

Needleman, S. B., & Wunsch, C. D. (1970). A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of two Proteins. *Journal of Molecular Biology, 48*(3), 443-453.

Notredame, C., Higgins, D. G., & Heringa, J. (2000). T-Coffee: a Novel Method for Fast and Accurate Multiple Sequence Alignment. *Journal of Molecular Biology, 302*(1), 205-217.

Nucci, M., Tagliasacchi, M., & Tubaro, S. (2013). A Phylogenetic Analysis of Near-Duplicate Audio Tracks. *IEEE International Workshop on Multimedia Signal Processing*, (pp. 99-104).

Nyeem, H., Boles, W., & Boyd, C. (2014). Digital Image Watermarking : A Formal Model, Fundamental Properties, and Possible Attacks. *EURASIP Journal on Advances in Signal Processing, 2014*(1), 1-21.

Nyeem, H., Boles, W., & Boyd, C. (2015). Watermarking Capacity Control for Dynamic Payload Embedding. *Unger H., Meesad P., Boonkrong S. (eds) Recent Advances in Information and Communication Technology. Advances in Intelligent Systems and Computing. Springer.*

Olabarriaga, S., & Smeulders, A. (2001). Interaction in the Segmentation of Medical Images: A Survey. *Medical Image Analysis, 5*(2), 127-142.

Oliveira, A., Ferrara, P., Rosa, A. D., Piva, A., Barni, M., Goldenstein, S., . . . Rocha, A. (2014). Multiple Parenting Identification in Image Phylogeny. *IEEE International Conference on Image Processing*, (pp. 5347-5351). Paris, France.

Panah, A. S., Schyndel, R. V., Sellis, T., & Bertino, E. (2016). On the Properties of Non-Media Digital Watermarking: A Review of State of the Art Techniques. *IEEE Acess, 4,* 2670-2704.

Parikh, R., Mathai, A., Parikh, S., Sekhar, G. C., & Thomas, R. (2008). Understanding and Using Sensitivity, Specificity and Predictive Values. *Indian Journal of Ophthalmology, 56*(1), 45-50.

Peng, B., Zhang, L., & Zhang, D. (2013). A Survey of Graph Theoretical Approaches to Image Segmentation. *Pattern Recognition, 46*(3), 1020-1038.

Pereira, S., Voloshynovskiy, S., Madueno, M., Marchand-Maillet, S., & Pun, T. (2001). Second Generation Benchmarking and Application Oriented Evaluation. *International Workshop on Information Hiding.* Pittsburgh, USA.

Piper, A., & Safavi-Naini, R. (2009). How to Compare Image Watermarking Algorithms. *Transactions on Data Hiding and Multimedia Security IV, Springer Berlin Heidelberg*, 1-28.

Podilchuk, C. I., & Zeng, W. (1998). Image-adaptive Watermarking Using Visual Models. *IEEE Journal on Selected Areas in Communications, 16*(4), 525-539.

Polyanovsky, V. O., Roytberg, M. A., & Tumanyan, V. G. (2011). Comparative Analysis of the Quality of a Global Algorithm and a Local Algorithm for Alignment of two Sequences. *Algorithms of Molecular Biology, 6*(1), 1-12.

Pour, E. S. (2015). A Survey of Multithreading Image Analysis. *CoRR, abs/1506.04472.*

Pramila, A., Keskinarkaus, A., & Sepp, T. (2008). Multiple Domain Watermarking for Print-Scan and JPEG Resilient Data Hiding. *International Workshop on Digital Watermarking.* Springer-Verlag: Guangzhou, China.

Qian, L., Winfree, E., & Bruck, J. (2011). Neural Network Computation with DNA Strand Displacement Cascades. *Nature, 475*(7356), 368-372.

Rampasek, L., Jimenez, R. M., Luptak, A., Vinar, T., & Brejova, B. (2016). RNA Motif Search with Data-driven Element Ordering. *BMC Bioinformatics, 17*(1), 1-10.

Ranchin, T., & Wald, L. (2000). Fusion of High Spatial and Spectral Resolution Images: the ARSIS Concept and Its Implementation. *Photogrammetric engineering and remote sensing, American Society for Photogrammetry, 66*(1), 49-61.

Rani, R. R., & Ramyachitra, D. (2016). Multiple Sequence Alignment using Multi-Objective Based Bacterial Foraging Optimization Algorithm. *Biosystems, 150*, 177-189.

Ridder, D. d., Ridder, J. d., & Reinders, M. J. (2013). Pattern Recognition in Bioinformatics. *Briefings in Bioinformatics, 14*(5), 633-647.

Riedel, D. E., Venkatesh, S., & Liu, W. (2006). A Smith-Waterman Local Alignment Approach for Spatial Activity Recognition. *IEEE International Conference on Video and Signal Based Surveillance*, (pp. 54-54). Sydney, Australia.

Roberto, V., & Hofer, M. (2009). Theia: Multispectral Image Analysis and Archaeological Survey. *International Conference on Image Analysis and Processing.* Vietri sul Mare, Italy.

Rondelez, Y., & Woods, D. (2016). DNA Computing and Molecular Programming. *International Conference on DNA-Based Computers.* Munich Germany.

Roth, V., & Ommer, B. (2006). Exploiting Low-Level Image Segmentation for Object Recognition. *DAGM Symposium.* Berlin, Germany.

Rubio-Largo, A., Vega-Rodríguez, M. A., & Gonzalez-Alvarez, D. L. (2015). Parallel H4MSA for Multiple Sequence Alignment. *IEEE Trustcom/BigDataSE/ISPA*, (pp. 242-247). Helsinki, Finland.

Rueden, C. T., Schindelin, J., Hiner, M. C., DeZonia, B. E., Walter, A. E., Arena, E. T., & Eliceiri, K. W. (2017). ImageJ2: ImageJ for the next generation of scientific image data.

Ryu, S.-J., Lee, H.-Y., Im, D.-H., Choi, J.-H., & Lee, H.-K. (2010). Electrophotographic Printer Identification by Halftone Texture Analysis. *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, (pp. 1846-1849). Dallas, USA.

Sankoff, D. (1972). Matching Sequences under Deletion/Insertion Constraints. *Proceeding of National Academic Sciences, 69*, pp. 4-6. USA.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., . . . Cardona, A. (2012). Fiji: an Open-source Platform for Biological-image Analysis. *Nature Methods, 9*(7), 676-682.

Schindelin, J., Rueden, C. T., Hiner, M. C., & Eliceiri, K. W. (2015). The ImageJ Ecosystem: An Open Platform for Biomedical Image Analysis. *Molecular Reproduction and Development, 82*(7-8), 518-529.

Schneider, C. A., Rasband, W. S., & Eliceiri, K. W. (2012). NIH Image to ImageJ: 25 Years of Image Analysis. *Nature Methods, 9*(7), 671-675.

Schneider, R. Z., & Fernandes, D. (2003). Entropy Among a Sequency of SAR Images for Change Detection. *IEEE International Geoscience and Remote Sensing Symposium. Proceedings (IEEE Cat. No.03CH37477).*

Schwartzkopf, W., Evans, B., & Bovik, A. (2002). Entropy Estimation for Segmentation of Multi-spectral Chromosome Images. *IEEE Southwest Symposium on Image Analysis and Interpretation*, (pp. 234-237). Sante Fe, USA.

Schyndel, R. V., Tirkel, A., & Osborne, C. F. (1994). A Digital Watermark. *Int. Conf. Image Processing*, (pp. 86-90).

Sellers, P. (1974). On the Theory of Computation of Evolutionary Distances. *SIAM Journal Applied Mathematics, 26*, 787-493.

Shyu, C., Sheneman, L., & Foster, J. A. (2004). Multiple Sequence Alignment with Evolutionary Computation. *Genetic Programming and Evolvable Machines, 5*(2), 121-144.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., & Lopez, R. (2011). Fast, Scalable Generation of High-quality Protein Multiple Sequence Alignments using Clustal Omega. *Molecular Systems Biology, 7*(1), 539.

Sliwoski, G., Kothiwale, S., Meiler, J., & Lowe, E. W. (2014). Computational Methods in Drug Discovery. *Pharmacological Reviews, 66*(1), 334-395.

Smith, T. F., & Waterman, M. S. (1981). Identification of Common Molecular Subsequences. *Journal of Molecular Biology, 147*(1), 195-197.

Soding, J. (2017). Big-data Approaches to Protein Structure Prediction. *Science, 355*(6322), 248-249.

Solachidis, V., Tefas, A., Nikolaidis, N., Tsekeridou, S., Nikolaidis, A., & I.Pitas. (2001). A Benchmarking Protocol for Watermarking Methods. *IEEE International Conference on Image Processing (Cat. No.01CH37205), 3*, pp. 1023-1026. Thessaloniki, Greece.

Solanki, K., Madhow, U., Manjunath, B. S., & Chandrasekaran, S. (2005). Modeling the Print-Scan Process for Resilient Data Hiding. *SPIE Security, Steganography, and Watermarking of Multimedia Contents VII.* San Jose, USA.

Soroushnia, S., Daneshtalab, M., Plosila, J., Pahikkala, T., & Liljeberg, P. (2014). High Performance Pattern Matching on Heterogeneous Platform. *Journal of Integrative Bioinformatics, 11*(3).

Takahashi, K., Yaegashi, S., Kameda, A., & Hagiya, M. (2005). Chain Reaction Systems Based on Loop Dissociation of DNA. *International Conference on DNA Computing*, (pp. 347-358). Canada.

Tamura, K., Dudley, J., Nei, M., & Kumar, S. (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0. *Molecular Biology and Evolution, 24*(1), 1596-1599.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., & Kumar, S. (2011). MEGA5: Molecular Evolutionary Genetic Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Molecular Biology and Evolution, 28*(10), 2731-2739.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Molecular Biology and Evolution, 30*(12), 2725-2729.

Tao, Y., Liu, Y., Friedman, C., & Lussier, Y. A. (2004). Information Visualization Techniques in Bioinformatics during the Postgenomic Era. *Drug Discovery Today., 2*(6), 237-245.

Tefas, A., Nikolaidis, N., & Pitas, I. (2009). *The Essential Guide to Image Processing (Second ed.).* Boston: Academic Press.

Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-specific Gap Penalties and Weight Matrix Choice. *Nucleic Acids Research, 22*(22), 4673-4680.

Toennies, K. D. (2017). *Guide to Medical Image Analysis Methods and Algorithms.* Springer London.

Turner, L. F. (1989). *Patent No. IPN WO 89/08915.*

Vandersypen, L., & Leeuwenhoek, A. V. (2017). Quantum Computing - the Next Challenge in Circuit and System Design. *IEEE International Solid-State Circuits Conference (ISSCC),*, (pp. 24-29). San Francisco, CA.

Villn, R., Voloshynovskiy, S., Koval, O., & Pun, T. (2006). Multilevel 2-D Bar Codes: Toward High-Capacity Storage Modules for Multimedia Security and Management. *IEEE Transactions on Information Forensics and Security, 1*(4), 405-420.

Vincent, R., Nicholas, J., & Philippe, D. J. (2011). Document Recto-Verso Registration Using a Dynamic Time Warping Algorithm. *International Conference on Document Analysis and Recognition,* (pp. 1230-1234). Beijing, China.

Voloshynovskiy, S., Pereira, S., Iquise, V., & Pun, T. (2001). Attack Modelling: Towards a Second Generation Watermarking Benchmark. *Signal Processing, 81*(6), 1177-1214.

Wald, L. (2000). Quality of High Resolution Synthesised Images: Is There a Simple Criterion ? *Fusion of Earth Data: Merging Point Measurements, Raster Maps and Remotely Sensed Images.* Sophia Antipolis, France.

Wang, B., Kong, X., & You, X. (2009). Source Camera Identification Using Support Vector Machines. *Advances in Digital Forensics V.* Orlando, Florida, USA.

Wang, Z. (2011). Applications of Objective Image Quality Assessment Methods. *IEEE Signal Processing Magazine, 28*(6), 137-142.

Wang, Z., & Bovik, A. C. (2002). A Universal Image Quality Index. *IEEE Signal Processing Letters, 9*(3), 81-84.

Wang, Z., & Bovik, A. C. (2009). Mean Squared Error: Love it or Leave it? A New Look at Signal Fidelity Measures. *IEEE Signal Processing Magazine, 26*(1), 98-117.

Wang, Z., Simoncelli, E. P., & Bovik, A. C. (2004). Multi-Scale Structural Similarity For Image Quality Assesement. *IEEE Asilomar Conference on Signals, Systems and Computers.* Pacific Grove, CA.

Waterman, M. S., Smith, T. F., & Beyer, W. A. (1976). Some Biological Sequence Metrics. *Advance Mathmatics, 20*, 367-387.

Weizhong, L., Cowley, A., Uludag, M., Gur, T., McWilliam, H., Squizzato, S., . . . Lopez, R. (2015). The EMBL-EBI Bioinformatics Web and Programmatic Tools Framework. *Nucleic Acids Research, 43*(W1), W580-W584.

Whelan, S., Lio, P., & Goldman, N. (2001). Molecular Phylogenetics: State-of-the-Art Methods for Looking into the Past. *Trends in Genetics, 17*(5), 262-272.

Wilgenbusch, J. C., Huang, W., & Gallivan, K. A. (2017). Visualizing Phylogenetic Tree Landscapes. *BMC Bioinformatics, 18*(1), 1-12.

Wilschut, L., Addink, E., Heesterbeek, J., Dubyanskiy, V., Davis, S., Laudisoit, A., . . . De Jong, S. (2013). Mapping the Distribution of the Main Host for Plague in a Complex Landscape in Kazakhstan: An Object-based Approach using SPOT-5 XS, Landsat 7 ETM+, SRTM and Multiple Random Forests. *International Journal of Applied Earth Observation and Geoinformation, 23*, 81-94.

Wolpert, D. H., & MacReady, W. G. (1996). *No Free Lunch Theorems for Search.* Santa Fe Institute: Technical Report SFI-TR-95-02-010. Sante Fe, USA.

Wong, P. C., Wong, K.-k., & Foote, H. (2003). Organic Data Memory using the DNA Approach. *Communications of the ACM, 46*(1), 95-98.

Wu, M., & Liu, B. (1998). Watermarking for Image Authentication. *International Conference on Image Processing. ICIP98 (Cat. No.98CB36269), 2*, pp. 437-441. Chicago, USA.

Wu, X., & Kan, H. (2015). A Blind Dual Color Images Watermarking Method via SVD and DNA Sequences. *Inscrypt 2015.* Beijing, China.

Wu, Y., Kong, X., You, X., & Guo, Y. (2009). Printer Forensics Based on Page Document's Geometric Distortion. *IEEE International Conference on Image Processing (ICIP),* (pp. 2909-2912). Cairo, Egypt.

Xie, Z., Gao, J., Wu, K., & Zhang, J. (2011). Brief Survey on Image Semantic Analysis and Understanding. *International Conference of Soft Computing and Pattern Recognition (SoCPaR),* (pp. 179-183). Dalian, China.

Yaghmaee, F., & Jamzad, M. (2008). Introducing a New Method for Estimation Image Complexity According To Calculate Watermark Capacity. *International Conference on Intelligent Information Hiding and Multimedia Signal Processing,* (pp. 981-984). Harbin, China.

Yamada, K. D., Tomii, K., & Katoh, K. (2016). Application of the MAFFT Sequence Alignment Program to Large Data—Reexamination of the Usefulness of Chained Guide Trees. *Bioinformatics, 32*(21), 3246-3251.

Yang, Z., & Rannala, B. (2012). Molecular Phylogenetics: Principles and Practice. *Nature Reviews Genetics, 13*(5), 303-314.

Yazdi, S. M., Yuan, Y., Ma, J., Zhao, H., & Milenkovic, O. (2015). A Rewritable, Random-Access DNA-Based Storage System. *Scientific Reports, 5*, 14138.

Ye, P., & Doermann, D. (2013). Document Image Quality Assessment: A Brief Survey. *International Conference on Document Analysis and Recognition*, (pp. 723-727). Washington, DC.

Yim, A. K.-Y., Yu, A. C.-S., Li, J.-W., Wong, A. I.-C., Loo, J. F., Chan, K. M., . . . Chan, T.-F. (2014). The Essential Component in DNA-Based Information Storage System: Robust Error-Tolerating Module. *Frontiers in Bioengineering and Biotechnology, 2*, 49.

Yousef, A. H., Salama, C., Jad, M. Y., El-Gafy, T., Matar, M., & Habashi, S. S. (2016). A GPU Based Genetic Algorithm Solution for the Timetabling Problem. *International Conference on Computer Engineering & Systems (ICCES).* Cairo, Egypt.

Zhang, B., & Rahmat-Samii, Y. (2016). Worst-Case Sensitivity Analysis (WCSA) by Particle Swarm Optimization (PSO): Applications in Realistic Optimal Antenna Designs. *International Conference on Electromagnetics in Advanced Applications (ICEAA),* (pp. 974-977). Cairns, Australia.

Zheng, D., Zhao, J., Tam, W. J., & Speranza, F. (2003). Image Quality Measurement by using Digital Watermarking. *IEEE Internatioal Workshop on Haptic, Audio and Visual Environments and Their Applications*, (pp. 65-70).

Zhou, C. E., Zemla, A. T., & Lam, M. W. (2007, October). *United states Patent No. 20070244652.*

Zhou, Z. (2016). Total Variation-Based Denoising Model for Bioinformatics Images. *International Journal of Bioautomation, 20*(4), 457-470.