Development and implementation of an artificial intelligence system for assessing corrosion damage at stem taper of hip replacement implants: A retrieval study

by

Roohollah Milimonfared

BEng (Hons). Mechanical Engineering

MEng. Advanced Manufacturing and Mechanical Engineering

Thesis

Submitted to Flinders University
for the degree of

Higher Degree Thesis (PhD)
College of Science and Engineering

2019

DECLARATION

“I certify that this thesis:

1. does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any university; and
2. to the best of my knowledge and belief, does not contain any material previously published or written by another person except where due reference is made in the text.”

ACKNOWLEDGEMENTS

This journey would not have been possible without the support of my family, professors, and friends. To my principal supervisor, Dr Reza H. Oskouei, for providing me with the opportunity to join his research team and this project. His willingness and trust to let this work steer in a different direction to explore the application of AI in this area is highly appreciated.

I wish to acknowledge the kind and remarkable support provided by my co-supervisors, Prof Mark Taylor and Prof Lucian B. Solomon. There is no doubt that their assistance with establishing communication links with other research groups in the US and Europe as well as Australian Orthopaedic Association provided me with insightful feedback regarding the quality of the work and its direction.

I wish to thank Prof Jan Philippe Kretzer for hosting my short visit to the Laboratory of Biomechanics and Implant Research at Heidelberg University Hospital (Heidelberg, Germany). It was an insightful experience to learn about the status of retrieval studies in Germany and research conducted on biomechanical aspects of orthopaedic implants.

I want to give special thanks to Dr Harry Hothi and Prof Alister Hart from London Implant Retrieval Centre at University College London (London, UK) for hosting my visit. Joining Prof Hart to observe two retrieval and primary operations at Royal National Orthopaedic Hospital as well as his clinic consultation sessions revealed even more to me the importance of retrieval studies and their role at improving the quality of life for many people.

I owe a debt of gratitude to Dr Paul Johan Høl and his colleagues Prof Ove N. Furnes and Mrs Irene O. Moldestad for hosting my research fellowship at Haukeland University Hospital (Bergen, Norway). Joining a retrieval study of Spectron EF implants was a great opportunity to learn about stem-cement-bone interfaces and their potential complications. Also, thanks to this team, I had the privilege to visit the Norwegian Arthroplasty Register and got familiar with their nationwide collection and storage of primary/revision operation records in Norway.

The kind assistance provided by Mr Abbas Banaei with working out a proper set of lighting-optics-viewing conditions to capture stem taper images is much appreciated. His knowledge and expertise in professional photography removed several obstacles to obtaining images with a consistent and acceptable level of quality.

I must thank the following engineering technical officers from the College of Science and Engineering at Flinders University for their great assistance with accessing and using laboratory equipment and also developing my implant photography set-up:

- Mr Mark Walford
- Mr Timothy Hodge
- Mr Craig Peacock
- Mr Richard Stanley

I want to acknowledge the support through an Australian Government Research Training Program Scholarship for my PhD. I also thank the Department of Orthopaedics and Trauma Service, Royal Adelaide Hospital (Adelaide, Australia), and in particular, Mrs Kerry Costi and Dr Roumen Stamenkov for their assistance in accessing their pool of retrieved implants.

To my parents, thank you for supporting me throughout my life with your thoughts, well-wishes/prayers, phone calls, e-mails, texts, and visits. I always knew that you believed in me and wanted the best for me. Thank you for teaching me that my job in life was to learn, to be happy, and to know and understand myself; only then could I know and understand others.

Finally, I wish to thank my wife, Dr Zeinab Hajiabolhasani, for her support and encouragement throughout my studies. This journey would not have been possible without her emotional support. Thank you for accompanying me during the three months of research fellowship despite the pregnancy.

DEDICATION

To ROHAM

THESIS SUMMARY

Despite the clinical benefits of modularity in total hip replacement (THR) implants, modular interfaces such as head-neck taper junction are susceptible to fretting wear and corrosion due to relative micro-motions at the interface and also the presence of corrosive body fluid. This fact induces a chain of host body responses that may ultimately result in revision surgery to retrieve the failed implant. Through large-scale implant retrieval studies, the damage sustained by the implants is assessed, and possible associations between several implant/patients factors and the extent/location of the damage are investigated.

This PhD study aims to conduct the first large-scale retrieval study in Australia through exploring a database of 2100 operation records available at Royal Adelaide Hospital and a retrieval library of implants with approximately the same number of implants that had been retrieved since 1980s. The database was filtered at multiple occasions to identify implants suitable for this study.

Visual scoring of damage at taper junctions is the sole method to quantify corrosion in large-scale retrieval studies. In this work, an intelligent image analysis-based method was developed and implemented that can objectively assess corrosion at the stem taper of retrieved hip implants, according to the popular Goldberg's scoring method. A Support Vector Machine classifier was used that takes in vectors of global and local textural features and assigns scores to the corresponding images. Bayesian optimisation fine-tuned the hyperparameters of six binary learners of this classifier to minimise the cross-validation error and increase the accuracy level to 85%.

Moreover, the spatial distribution and the severity of corrosion damage onto the surface of the metallic stem tapers were objectively explored. An ordinal logistic regression model was developed to find the odds of receiving a higher score at eight distinct zones of stem tapers. A method to find the order of damage severity across the eight zones was introduced based on an overall test of statistical significance. The findings showed that corrosion at the stem tapers occurred more commonly in the distal region in comparison with the proximal region. Also, the medial distal zone was found to possess the most severe corrosion damage among all the studied eight zones.

In the last phase of the project, several multivariate analyses of patient and implant factors were carried out to identify the challenges regarding the causal-explanatory statistical modelling techniques that are currently used in the literature of retrieval studies. It was elaborated why this group of techniques are not suitable for looking at multiple confounding variables. Predictive analytics was recommended to be utilised in conjunction with the existing methods to enable clinicians to predict the likelihood of implants failure for prospective recipients.

LIST OF PUBLICATIONS

- Milimonfared R, Oskouei RH, Taylor M, Solomon LB. An intelligent system for image-based rating of corrosion severity at stem taper of retrieved hip replacement implants. *Medical Engineering & Physics*, 61 (2018) 13–24.

- Milimonfared R, Oskouei RH, Taylor M, Solomon LB. The distribution and severity of corrosion damage at eight distinct zones of metallic femoral stem implants. *Metals*, 8(10) (2018) 840.

- Milimonfared R, Oskouei RH, Taylor M, Solomon LB. An automated method for characterising corrosion at stem trunnions in retrieved total hip arthroplasty implants using digital image processing and machine learning. ASTM, Symposium on Beyond the Implant: Retrieval Analysis Methods for Implants Surveillance, Toronto, Canada, 9 May 2017.

- Milimonfared R, Oskouei RH, Taylor M, Solomon LB. Fretting wear and corrosion in modular hip joint implants: A retrieval study. Australian Orthopaedic Association (AOA) SA/NT Branch Scientific Meeting, The Lyell McEwin Hospital, SA, 24 Feb 2017.

International Visits (partially funded by Flinders Overseas Travelling Fellowship, Flinders University):

- Biomaterial Laboratory, Haukeland University Hospital, Bergen, Norway, 03 Oct - 17 Nov 2017 (7 Weeks)

- London Implant Retrieval Centre, University College London – Royal National Orthopaedic Hospital, London, England, 20 Nov - 01 Dec 2017 (2 Weeks)

- Laboratory of Biomechanics and Implant Research, Heidelberg University Hospital, Heidelberg, Germany, 28 & 29 September 2017.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

List of Abbreviations

Abbreviation	Description
ABS	Australian Bureau of Statistics
AI	Artificial Intelligence
ALTR	Adverse Local Tissue Reactions
ALVAL	Aseptic Lymphocytic Vasculitis-Associated Lesions
AOANJRR	Australian Orthopedic Association National Joint Replacement Registry
ASM	American Society for Metals
ASTM	American Society for Testing and Materials
BOVW	Bag Of Visual Words
CDSS	Clinical Decision Support Systems
CNN	Convolutional Neural Network
DIP	Digital Image Processing
DTC	Double Tapered Cone
DV	Dependent Variable
DVRT	Differential Variable Reluctance Transducer
ECOC	Error Correcting Output Codes
EMR	Electronic Medical Record
FDA	Food and Drug Administration
FEA	Finite Element Analysis
FT	Fourier Transform
GLCM	Grey Level Co-occurrence Matrix
HP	High Pass
HSV	Hue-Saturation-Value
IV	Independent Variable
LP	Low Pass
NCA	Neighborhood Component Analysis
NITINOL	Nickel-Titanium Naval Ordnance Laboratory
OA	Osteoarthritis
OLR	Ordinal Logistic Regression
OR	Odds Ratio
RA	Rheumatoid Arthritis
RGB	Red-Green-Blue
RGB	Red-Green-Blue
ROI	Region Of Interest
SEM	Scanning Electron Microscope
SIFT	Scale Invariant Feature Transform
SURF	Speeded-Up Robust Features
SVM	Support Vector Machine
THR	Total Hip Replacement
VIF	Variance of Inflation Factor
VML	Volumetric Material Loss

1 BACKGROUND

[Image removed due to copyright restriction]

1.1 Introduction

This chapter provides an overview of various aspects of hip replacement devices ranging from medical to engineering. It elaborates on their postoperative complications and the existing approaches to mitigate them.

Since this thesis is about conducting a retrieval study of failed hip replacement implants, it is essential to have a reasonable level of insight into the medical conditions that lead to hip replacement, the history of these prostheses, their constituent components and biomaterials, the potential postoperative complications, and the approaches to mitigate them.

1.2 Joint Disease

Joint diseases can be categorised into several groups. Two popular ones are known as Rheumatoid Arthritis (RA) and Osteoarthritis (OA). RA is a chronic systemic disease that affects connective tissues, muscle, tendons, and fibrous tissues along with joints. It is characterised by an overactive immune system and joint inflammation. However, OA is a typical joint condition that has degenerative characteristics. The joint deterioration most likely occurs owing to continuous stressing of articular cartilage. It encompasses knees, hips, fingers, and lower spine region.

A study conducted by the U.S. Census Bureau in 2005 reported on the prevalence and most common causes of disability among adults in the U.S., which is listed in Table 1-1. This table is based on responses from an estimated 45.1 million persons (94% of total) reporting a disability.

Table 1-1. Arthritis holds the first rank among the causes of disability in the U.S.

Condition	All Persons		Men		Women	
	Estimated Population	**%**	**Estimated Population**	**%**	**Estimated Population**	**%**
Arthritis or Rheumatism	8,552	19	2,154	11.5	6,398	24.3
Back or Spine Problems	7,589	16.8	3,158	16.9	4,431	16.8
Heart Trouble	2,988	6.6	1,570	8.4	1,418	5.4
Lung or Respiratory Problem	2,224	4.9	925	4.9	1,299	4.9
Mental or Emotional Problem	2,203	4.9	982	5.2	1,222	4.6
Diabetes	2,012	4.5	907	4.8	1,106	4.2

In this table, arthritis holds the first rank among the main causes of disability in this country. Arthritis Foundation considers OA as the most common form of arthritis with 27 million people affected by it. The high prevalence of OA in the world and complicated nature of this disease have been among the motivations for conducting this study. Figure 1-1 illustrates the impact of RA and OA on a bone joint.

[Image removed due to copyright restriction]

Figure 1-1. Schematic comparison between a normal and two diseased joint (from www.thefitindian.com)

1.2.1 Osteoarthritis: A Snapshot

Hip OA is characterised by mobility impairment and joint pain in which joints start to lose their elasticity and become stiff. As displayed in Figure 1-2, when cartilage weakens, tendons and ligaments may start to stretch and bones may rub against each other. Hence, joints lose their ability to act as shock absorbers any more.

[Image removed due to copyright restriction]

Figure 1-2. Impact of OA on hip joint (from www.preferredpaincenter.com/hip-pain.html)

While the causes of this condition are not completely understood, it is believed that the following factors may contribute toward it.

- increasing age
- being overweight
- congenital disorder (inherited defects in the cartilage)
- putting excessive stress on joints through activities that involve the hip

Unlike Rheumatoid Arthritis, OA is considered a silent disease. That is, only after the development of pronounced inflammations, symptoms of pain become noticeable.

1.2.2 Prevalence and Incidence

OA is the most prevalent form of arthritis in Australia. In 2011-12, the Australian Bureau of Statistics (ABS) reported 14.8% of Australians suffering from arthritis, which is equal to around 3.3M. More than half portion (55.9%) of this population had OA.

ABS figures accentuate age and sex as the most prominent factors. While people aged 25 or below constituted only less than 1% of the patients, 52.1% falls into the category that was aged 75 or above. Also, women constituted 59.9% of patients in this category.

It is also worthy of note that joint disease is not the only source that stops joints from functioning as trauma leads to that as well.

1.2.3 Treatment

No cure has been found for OA so far. Only a few treatments exist that help alleviate pain or restore motion. Yet, they cannot prevent, halt, or reverse OA progression. Among them, an increasingly popular approach is hip replacement (Figure 1-3).

[Image removed due to copyright restriction]

Figure 1-3. Replacement of femur head and acetabular by a prosthesis (from http://medicalpicturesinfo.com)

In traumatic situations, based on the definition provided by the Australian Orthopaedic Association National Joint Replacement Registry (AOANJRR), these systems are used when "*a fracture just below the head (ball of the femur) where the head and neck have also become separated and out of position*" is encountered (Figure 1-4).

[Image removed due to copyright restriction]

Figure 1-4. Fractured femur bone (from www.hughston.com)

AOANJRR has introduced three broad categories of hip replacement systems, namely primary total, primary partial, and revision hip replacements.

In the first two groups, the common property is the replacement of the femur head by an artificial head. As partial hip replacement is outside the scope of this study; hence, only THR and revision groups are to be elaborated here. Within THRs, the scope narrows to total conventional systems since NJRR reported that resurfacing has become quite infrequent, and, in 2012, only 1.4% of all primary total hip replacements utilised resurfacing. Figure 1-5 compares schematically the difference between these two THR systems.

[Image removed due to copyright restriction]

Figure 1-5. Conventional vs resurfacing hip replacements (from www.toxicdoselaw.com)

Due to its significantly lower popularity, total resurfacing and thrust plate systems are excluded from this study, and total hip replacements (THRs) is to refer to total conventional hip replacement systems from now on.

1.3 Total Hip Replacements

The utilisation of THRs dates back to 1960s. An orthopaedic surgeon, Sir John Charnley, who was working at Manchester Royal Infirmary established the modern THR by introducing his low friction arthroplasty [1]. Figure 1-6 represents his design that consisted of a metal femoral stem, a polyethylene acetabulum, and acrylic bone cement.

[Image removed due to copyright restriction]

Figure 1-6. Sir. John Charnley and his THR design (from www.almuderis.com.au)

Over time, the increasing number of successful operations resulted in the surgery techniques becoming standardised and the average age of recipients to drop. Currently, THRs are comprised of a head which fits into an artificial socket that replaces the acetabulum. The artificial socket is fixed into the pelvis. Also, there is a stem which is passed down inside the femur to hold the head in place (Figure 1-7).

[Image removed due to copyright restriction]

Figure 1-7. THR in position (from www.telegraph.co.uk)

THRs have a vast variety in terms of their designs, materials, and brands in the market (Figure 1-8). Head and stem in THRs come with monoblock (one piece), and modular (separate) designs with the latter introduced more recently.

Orthopaedic surgeons select THRs for patients while taking into account several patient and implant factors plus their experience to ensure achieving the best functionality with the lowest risk of postoperative complications.

[Image removed due to copyright restriction]

Figure 1-8. Which implant configuration serves a specific patient best? (from http://www.levinlaw.com)

1.3.1 Modular THRs

A prominent feature of THR systems is modularity which dates back to 1971 when Harris [2] introduced the first modular prosthesis to address what he claimed as features that may extend wear, increased stability, and preserving range of motion [3]. His modular system included a metal-backed acetabular component with a high-density polyethylene core twisted into it. Figure 1-9 displays this preliminary design, which was still monolithic where head and neck were attached as a single piece.

[Image removed due to copyright restriction]

Figure 1-9. The first modular design made by Harris [2]

The inner, replaceable, and high-density polyethylene core had a higher thickness in comparison to that period's existing designs which extended the longevity of the device. Also, for the first time, it became possible to replace the polyethylene component with a new one.

Over the years, modular designs went through various changes in terms of their popularity and specifications. In the market, different variants of modular THRs can be found which may use various components such as screws, polyethylene inserts, attachable collars, distal and proximal sleeves. Figure 1-10 shows a modular prosthesis. In this design, the head is connected to the stem by a double-tapered neck (Figure 1-10).

Along with offering plenty of merits, modular designs pose some concerns as well. The primary advantage of using modular systems is that the surgeon can assemble a custom prosthesis to accommodate the anatomical variations among patients more effectively. However, this capability may give rise to new and unpredicted risks. Therefore, the modularity benefits need to be carefully weighed up against their associated risks.

[Image removed due to copyright restriction]

Figure 1-10. A modular double-taper THR prosthesis [4]

1.3.1.1 *Advantages*

A higher degree of freedom in customising THR designs warrants the following clinical benefits.

- Choice of bearing surface after femoral fixation
- Intraoperative leg length adjustment via the head-neck taper
- Intraoperative femoral anteversion adjustment via neck-stem taper
- Simplified revision procedures due to having more component-targeted revisions
- Flexibility in terms of material selection
- Retaining well-fixed acetabular shells
- Reduced revision operation times
- Decreased inventory levels [5]

1.3.1.2 *Disadvantages*

Despite these clinical benefits, increasing the number of modular interfaces presents additional sites for failure. Recent research works support the hypothesis that various components of contemporary modular designs are susceptible to fretting and corrosion damage [6]. Solid and soluble wear debris and corrosion products released from bearing surfaces, taper joints, or cement stem interface elicit some untoward body's responses (Figure 1-11) that will be discussed in section 1.4.1.

Combinations of mechanical, electrochemical, geometrical, material, and chemical conditions have been identified as the underlying factors for such problems [7]. Section 1.4 will dig into corrosion mechanisms plus physical and chemical evaluation of the resulting deposits.

[Image removed due to copyright restriction]

Figure 1-11. Metal debris released from modular THRs has posed concerns (from www.yourlawyer.com)

The complex nature of physical and chemical interactions between a host body and an implant may arise different complications within short or long periods after primary surgery. Based on the intensity of these postoperative complications and patients' circumstances, a surgeon may proceed with replacing THRs totally or partially with a new one by doing revision surgery.

THR systems are under constant pre-screening and review for any higher than expected rate of revision and postoperative complications. Over the past decades, there have been some THR systems which were recalled and withdrawn from the market. In July 2012, Stryker Orthopaedics (Mahwah, NJ) initiated a voluntary recall of two of their relatively new THR systems, ABG II and Rejuvenate (Figure 1-12). This recall was due to concerns raised by early implant failures. The issue was identified to be in association with corrosion at modular neck-stem interface [8].

[Image removed due to copyright restriction]

Figure 1-12. Rejuvenate & ABG II THRs recalled in July 2012 (from www.metalonmetalhipsettlement.com)

Properties of biomaterials used in the fabrication of these prostheses underlie their success or failure. Therefore, the next section provides an overview of the biomaterials that are currently in use and compares them based on their applications and properties.

1.3.2 Biomaterials

In Oxford dictionary, biomaterial is defined as "*A biological or synthetic substance which can be introduced into body tissue as part of an implanted medical device or used to replace an organ, bodily function, etc.*"

Biomaterials are used in a host of situations such as implants, extracorporeal medical devices, dermal and mucosal treatments and devices, drug delivery systems, sensors, and diagnostic assays. Hence, a wide variety of them exists based on three major goals that they can serve. The first goal is to assume the functional properties of a replaced tissue while provoking reasonable (or minimal) levels of deleterious response by a host body. The second goal is associated with biomaterials that are resorbable. Based on the situation at hand, they can be designed so that desirable degradation rates can be achieved. That is, the discrete interface between the surface of an implant and surrounding biological tissues could controllably degrade into soluble and nontoxic products by the host body. The third goal comprises biomaterials that support and stimulate the regeneration of functional tissues. They can be utilised to regenerate organs or tissues lost due to disease or trauma. The scope of this research project only encompasses the first goal.

The biomaterials used in hip prostheses belong to a diverse range that includes metallic alloys, ceramics, polyethylene, composites, hydrogels, engineered natural materials, Pyrolytic Carbon. Due to the overall synthesised nature of them, they mostly serve the first group of the purposes above. The performance of THRs is a function of the properties of their biomaterials. The forthcoming sub-sections introduces the metallic alloys, ceramics, and polyethylene that are used in their fabrication.

The first level of interaction between biomaterials and host body is through their surface which comes in contact with biological species. Hence, the overall responses of the biological environment to biomaterials depend on their compositions, structures, and properties. The following sections discuss briefly these aspects of these three classes of biomaterials [9].

1.3.2.1 *Metallic Biomaterials*

Today, a big portion of the medical device industry relies on implants with one or more metallic parts which has given them a special place in the biomaterials market [10]. Three types of metallic biomaterials, namely stainless steel, Co-based alloys, and Ti-based alloys will be discussed here. Some of their categories, as established by the American Society for Testing and Materials (ASTM), along with their desirable and unfavourable properties, are to be introduced in this section.

Stainless steel is the first (since 1926) metallic alloy that was used in orthopaedic practice. Currently, the most popular form is 316L (ASTM F138 – ISO5832-1). "316" and "L" indicate austenitic and low Carbon content. Low concentration of carbon prevents carbide (Chromium-Carbon) accumulation at the grain boundaries. More recent stainless steel alloys such as 22-13-5 (ASTM F1314), Rex 734 (Ortron 90 – ASTM F1586 – ISO 5832-9), and Biodur 108 (ASTM F2229) exhibit even better characteristics (e.g. more resistance to corrosion) [10]. Stainless steels are comprised of Iron and Carbon, and may typically possess other elements such as Chromium, Nickel, and Molybdenum. Different concentrations of these secondary elements may affect steels' mechanical properties through alteration of their microstructures [11].

Presence of foreign particles known as *inclusions* may adversely affect steels behaviour against corrosion. Initial melting of alloys may give rise to the creation of oxide particles such as alumina or silicates. These particles may become trapped within the material during the subsequent processing. If located on the alloy's surface, they may act as corrosion initiation sites. That justifies the importance of their processing history. ASTM has specified permissible ranges for the size and the number of inclusions [10].

In general, stainless steels have lower strength and resistance to corrosion in comparison to the other implant alloys [10]. On the other hand, they promise desirable ductility (up to threefold elongation at fractures) in comparison to the other types of alloys [11].

Strengthening methods such as cold-working (strain hardening) via plastic deformations combined with annealing is used to reduce slip of dislocation within the crystal structure. Cold-working increases hardness, yield, ultimate tensile, and fatigue strength in comparison to the annealed state. However, it reduces the ductility which ordinarily does not raise major concern in metallic implants. After cold working metal, it is usually heated to a sufficiently higher temperature (around 750°C) where at grain boundaries, new grains (quite different to original ones) begin to form. They grow rapidly until a new undistorted grain structure completely takes over the old distorted one (recrystallisation). Then, the metal is left to cool in the air where the old type of grains begins to reappear and grows till they meet their neighbours. The new grains look similar to the old ones, yet they have reduced sizes and are more uniform, which in turn change the properties of the steel. This process restores softness and ductility and reduces tensile strength.

The recommended grain size for 316L is ASTM #6 or finer (higher) as smaller grain diameters may induce higher levels of yield stress. Grain size is mostly a function of manufacturing history (solidification conditions, cold working, annealing cycles, and recrystallisation) [10]. Manufacturing operations can elongate or distort grains which result in a change in mechanical properties of the metal.

Cobalt-based alloys have been in use since the 1920s. At first, their application was limited to dental implants, yet through time, these alloys found applications in orthopaedic, spinal, and cardiovascular products [12]. Table 1-2 lists some popular series of this alloy that is used for such applications. Apart from their processing histories, F75 and F799 are identical in composition. They both contain 58-70% Cobalt and 26-30% Chromium. F90 and F562 possess slightly less Chromium and Cobalt as well as more Tungsten (F90) and Nickel (F562) [10].

Relatively, large percentages (25-37%) of Nickel in F562 and F563 (Co-Ni-Cr-Mo-W-Fe) provide better corrosion resistance. However, the Nickel released from these alloys has raised concerns in association with toxicity and Adverse Local Tissue Reactions (ALTR). Also, due to their poor frictional (wear) properties, their application in articulations are not recommended [11]. The main attribute of F75 is corrosion resistance in chloride environments due to its bulk and oxide (Cr_2O_3) compositions. Metallurgical processes used for this group of metallic biomaterials are predominantly casting (F75) and in some cases (F1537) wrought [10].

Cobalt-based metals used for fabrication of joint replacement components are the strongest, hardest, and most fatigue resistant alloys. However, finishing treatments such as coating must be executed with relative care to avoid losing these properties [11].

Table 1-2. Cobalt-based alloys used in implants [10]

NAME	Co-Cr-Mo	Co-Cr-Ni-W	Co-28Cr-6Mo	Co-Ni-Cr-Mo
SERIES	ASTM F75	ASTM F90	ASTM F799	ASTM F562

Titanium-based alloys have been in use in the aviation industry since the mid-1940s. These alloys entered orthopaedic practice around the same time. Commercially pure Titanium (cpTi), Ti-6Al-4V (ASTM F136), and Ti-6Al-7Nb are currently in use for fabrication of implants. cpTi with 98-99.6% Titanium is commonly used in dental implants and porous coating (due to higher ductility) of joint components which are made of Ti-6Al-4V [11].

Alloys that are rich in Titanium are covered by TiO_2 layer. This passive oxide film turns Ti-based alloys into superior corrosion resistant biomaterials. Also, this group of biomaterials is relatively bioinert and do not illicit in-vivo allergenic responses. The other attractive feature is their moduli of elasticity which is closer to that of bone in comparison to other alloys. Table 1-3 summarises some mechanical properties of these three groups in comparison to their counterparts [10].

Efforts are being made to achieve closer Young moduli to that of bone in the next generations of Ti-based biomaterials. For instance, Ti-Nb-Ta-Zr (E = 40 GPa) or Tu-35Nb-7Zr-5Ta (E = 55 GPa) are gaining attention owing to their desirable elastic properties. Furthermore, Nitinol (Nickel-Titanium Naval Ordnance Laboratory) which is an equiatomic alloy of Nickel and Titanium is a recent addition to this group. Commercial use of NiTi commenced in the mid-1990s due to its superelastic and shape memory properties. Yet, some drawbacks in association with Nickel ionic release and provoking allergenic responses at high concentration levels have restricted utilisation of this type of biomaterials [10].

Table 1-3. Mechanical properties of some Ti-based alloys used in orthopaedic applications

Alloy	Microstructure	Elastic Modulus (GPa)	Yield Strength (MPa)	Ultimate Tensile Strength (MPa)
cp Ti	α	105	692	785
Ti-6Al-4V	α/β	110	850-900	960-970
Ti-6Al-7Nb	α/β	105	921	1024
Steel 316L	-	205-210	170-750	465-950
Co-Cr-Mo	-	220-230	275-1585	600-1785
Bone	-	10-40	-	-

Titanium-based alloys have the highest corrosion resistance relative to the other two groups. Some of their mechanical properties such as lower flexural rigidity can exceed those of stainless steel and Cobalt-based alloys. The proximity of their torsional and axial stiffness moduli to those of bone provides lower stress shielding. On the other side, their relative softness and poor wear and frictional properties are considered as their major drawbacks. Ti-6Al-4V is more than 15% softer than Co-Cr-Mo and subsequently wear significantly more when deployed in articulation sites [11].

1.3.2.2 *Ceramic Biomaterials*

The second class of biomaterials is Ceramics. Compounds such as Al_2O_3 (Alumina), SiC (Silicon Carbide), MgO (Magnesia), Fe_3O_4 (Magnetite), and ZrO (Zirconia) are usually brittle and corrosion-resistant. They are inorganic non-metallic compositions used in a wide variety of medical applications including orthopaedic prostheses.

For more than 30 years, alumina has been utilised in the fabrication of articulating surfaces of (usually young patients) joints. Properties such as excellent biocompatibility, high strength, and corrosion and wear resistance are among desirable characteristics of this nearly inert crystalline material. They usually have relatively low friction coefficients. Ageing and fatigue studies have shown that it is essential to produce alumina while meeting or exceeding the relevant quality assurance standards (i.e. ISO 6474) to achieve excellent resistance to subcritical crack growth and dynamic and impact fatigue. Also, very small grains (<4 µm) and narrow size distribution facilitate superb tribologic (friction and wear) properties (Ra <0.02 µm). These conditions have led to 10 times lower wear on alumina-alumina articulations in comparison to their metal-polyethylene counterparts. Currently, non-cemented cups and femoral balls made of alumina have generally good long-term results (1 failure in 2000 over ten years), especially in younger recipients [13]. As ceramic particulate debris is chemically stable, it causes no adverse biologic response at high concentrations [11].

Zirconia was also in use for fabrication of hip and knee articulations between 1985 and 2000 when due to a series of implant failures (fracture in 400 femoral heads) within a very short period it was withdrawn from the market. Yet not entirely, as zirconia toughened alumina put forward for orthopaedic components is expected to enhance strength and toughness properties over those of alumina. Toughening is a phase transformation process that increases the resistance to crack propagation [14].

Apart from the intrinsic properties of biomaterials, other properties such as surface characteristics are receiving an increasing level of attention about the biological performance of biomaterials. As it is the surface of an implant that interacts with the biological environment, it can strongly influence the overall material-body response (Figure 1-13). This response can be quite detrimental to the overall performance of a prosthesis which gives rise to the notion of biocompatibility.

[Image removed due to copyright restriction]

Figure 1-13. Biomaterial surface and surrounding tissues[15]

A point to bear in mind is the fact that the properties discussed here are not the sole determinant of successful performance of implants. That is, inadequate attention to them may doom an implant to failure; yet, there exist other considerations such as the mechanical design that may result in their failure as well [10].

1.3.3 Biocompatibility

Williams [16] defined biocompatibility as "*the ability of a material to perform with an appropriate host response in a specific application*". A device is considered biocompatible when it successfully fulfils its intended functionality [17].

A prosthesis and the biological environment of a host impact each other mutually. Biocompatibility is about the impact of the host body on the prosthesis material such as tissue reactions to bone cement, an uncemented titanium stem, or an acetabular cup. These are reactions to specific components of the device (biomaterials) [17]. One preliminary level of this response can take place postoperatively. The surgery leads to injury to surrounding tissues or organs which initiates the host defence system. The response of defence system can be in different forms, namely inflammatory, wound healing, and foreign-body responses.

Unlike biological biomaterials, synthetic biomaterials are less likely to be attacked by the immune system. That arises from having immunologically recognisable biologic motifs on the tissue. Still, non-biologic biomaterials may induce other variants of biological responses such as the clotting of blood and foreign-body reaction that are non-specific and can impair their usefulness. The basis for such reactions is associated with the adsorption of adhesion proteins to the surface of biomaterials.

[Image removed due to copyright restriction]

Figure 1-14. Biocompatibility, toxicity, and allergic reactions are three major criteria for biomaterials (from http://biomatsci.blogspot.com.au)

Figure 1-14 represents how they can be converted into biologically recognisable materials. Yet, the protein adsorption event which occurs on all implanted materials within seconds is out of the scope of this research along with tissue regeneration approach.

Traditionally, the main focus in biomaterial science was on the synthesis, characterisation, and the material-host interaction. However, it has been witnessed that biomaterials that met the mandating standards still induce non-specific biological responses known as a *foreign-body reaction* [18].

The non-specific nature of these responses justifies the current studies being undertaken to work out some subjective measures which characterise them. Establishing quantitative linkage between human and implant factors sheds light on a better selection of implant systems for patients, and consequently lower risk of postoperative complications.

1.3.4 The Mechanical and Electrochemical Behaviour of THRs

As mentioned previously, a prosthesis and the corrosive environment of the host body impact each other mutually. Here, the impact of a prosthesis on a host body will be discussed. The mechanical and electrochemical properties of a device such as fatigue, corrosion resistance, and distribution of the stresses that are transferred to the bone are substantial to warrant a desirable performance. Unlike the preceding section which concerns biological reactions to specific components of a device (biomaterials), here, a device is evaluated in its entirety. This distinction can be seen in the US Food and Drug Administration (FDA) policy where only complete devices instead of materials receive approval [17].

It should be noted that an implant system is to replace a joint. Joints provide relative motion between body parts. For instance, hip joints are to connect femur bone to pelvis. The head of femur bone makes a ball-socket type of junction with the concave surface of the pelvis being known as acetabulum. Hence, as the implant system imitates the same mechanism, the bearing area formed between the head surface and the acetabular shell is to go through multiple loading-unloading cycles over time. The dynamic nature of these mechanical cycles can result in malfunctioning of the implant which appears as wear, fracture, corrosion, dislocation, impingement, and so on.

Figure 1-15 displays wear on the surface of a hard-on-hard articulation site. Metal particles generated at other parts of the system may move to the bearing surface. Then, patch or stripe wear of the bearing surface may take place after subsequent loading cycles.

[Image removed due to copyright restriction]

Figure 1-15. Wear on articulation surfaces[19]

Hence, the mechanical properties of the implants are critical to their performance since the relative motion between the components of the implant systems can give rise to issues which adversely affects both the implant and the nearby tissues.

1.4 **Postoperative Complications**

As explained previously, two major concerns exist about the performance of implant systems: (1) their biomaterials must not influence the surrounding host tissues and fluids adversely; (2) in return, they are not expected to sustain damage from the surrounding tissues and fluids [11]. These systems are operating within an aqueous medium that contains diverse cations (e.g. Na^+, K^+, Ca^{2+}, Mg^{2+}), anions (chloride, phosphate, and bicarbonate ions), organic substances (e.g. proteins and lipids), and dissolved oxygen (in venous blood approximately a quarter of that in air) [20]. Over time, either the system or the surrounding tissues may sustain such serious damage that leaves the orthopaedic surgeon/s no choice but to redo the surgery. More information about these two concerns is provided in sections 1.4.1 and 1.4.2.

1.4.1 Adverse Host Body Reactions

Over time, failure of an implant system may elicit a range of host body responses which can be categorised under the general term of adverse local tissue reaction (ALTR) [20]. After reviewing the current concepts in biological reactivity to metal wear particles, ions and corrosion products, Billi et al. [21] maintained that the mechanisms through which ALTR is induced are still unclear. They maintained the physiochemical properties of these by-products play a significant role in governing their cellular uptake and the succeeding intercellular fate.

- Aseptic Lymphocyte-dominated Vasculitis-Associated Lesions (ALVALs)
- Osteolysis (bone loss) due to particles wearing off the implant surface and releasing into hip joint and surrounding tissues
- Periprosthetic fracture, fractures around joint replacement prostheses
- Hypersensitivity
- Pseudotumor formation owing to increased levels of metal ions
- Metallosis due to deposition and build-up of metal debris in the soft tissues
- Inflammation that reduces blood pH from about 7.4 to 4 or 5

[Image removed due to copyright restriction]

Figure 1-16. Periprosthetic fracture at different areas of femur bone (left) (from www.orthointerview.com) the surrounding black tissue stemmed from implant wear and particle release (right)[22]

There are various mechanisms in use to monitor and identify ALTR. Among them, ions concentration in blood plasma and extracellular fluid is a popular method. Each type of ion has a permissible range of concentration. Consistent monitoring of their levels in THR recipients help identify problems at an early stage. Research works have endeavoured to discern possible correlations between ion levels and THR damage.

evaluated Chromium and Cobalt ion levels in a cohort of asymptomatic patients (n=16) who had implanted ABG II modular hip (Stryker) between July 2007 and November 2011. It was to determine whether the increase in ion levels had been originating from corrosion at the modular neck-stem interface. In doing so, they selected patients who had received either a ceramic-on-ceramic or a ceramic-on-polyethylene interface to remove any chance of ion generation from the articulating interfaces.

The result of their study showed higher levels of serum Cobalt than Chromium. They hypothesised it could be due to the rapid transport of Co from joint and elimination in the urine. Whereas Chromium is stored in the tissues and eliminated gradually. Hence, they concluded that the rise of ion in serum could only be an ALTR predictor with moderate specificity and sensitivity. However, they believed their sample size was too small for such an analysis, and further study is required to confirm this result.

Molloy et al. [23] measured serum Cobalt and Chromium levels in 15 patients. They found a poor correlation between serum metal ion levels and ALTR, and no patient had undergone revision only due to high serum metal ion levels.

1.4.2 Impaired THR Systems

Different components of a THR system may sustain various forms of damage. At taper junctions that is the interest of this research, the damage modes are as follows.

- Fretting wear, a contact damage process resulting from micro-motions of interfacing metals that usually happens in modular junctions
- Corrosion, chemical reaction with the biological environment that results in the formation of compounds such as oxides or hydrated oxides on the surface exposed to air, water, or electrolyte

[Image removed due to copyright restriction]

Figure 1-17. Corrosion on metallic implant surface through random attacks by immune system cells[8]

It should be noted though that these two modes of damage are not always mutually exclusive. These two damage modes may integrate into a destructive process known as mechanically assisted crevice corrosion which will be described shortly. This research looks into the surface damage and does not investigate the resultant surrounding tissue damage.

There are other types of damage involved which are either not relatively significant or within the scope of this project. Still, they are not independent of each other. That is, one may contribute to the initiation or magnification of the other/s through some chain reactions. For instance, fretting may wear off or fracture the oxide layer on a metal that triggers corrosion. Through this process, the debris released from the implant can be in the form of metal ions (Co^{2+}, Cr^{3+}, Mo^{2+}, and Ti^{2+}), fractured oxide, or metallic particulate. These particles may initiate Osteolysis (Figure 1-18) where Osteoclasts (cells responsible for the resorption of the bone matrix during bone remodelling) are activated. Since Osteoclasts are abnormally activated, they make bones too weak to withstand the loads which in turn results in a higher likelihood of fractures and loosening of the femoral stem. Also, the release of metal particles may give rise to Metallosis.

[Image removed due to copyright restriction]

Figure 1-18. Metal debris release is the underlying factor for further complications (from http://www.ltu.se)

This type of collaboration between fretting and corrosion is known as mechanically assisted corrosion. Still, there are other types of corrosion involved in THRs which will be discussed as well.

1.4.2.1 *Fretting*

ASM (American Society for Metals) handbook on fretting and fracture defined fretting as “A special wear process that occurs at the contact area between two materials that are under load and subject to minute relative motion by vibration or some other force”.

In THRs, the male taper of necks fits into the female taper of heads. In THR systems with higher modularity, neck and stem can be separated as well. The neck in this type of systems has double tapers which fit into the female taper of the head and the stem.

At taper junctions, no relative motion between tapers is meant to occur, but in conditions of poor lubrication in interfacing metals, high friction moments can trigger fretting. Also, the constant loading-unloading cycles between bearing surfaces of articulations may fracture and wear off segments of the passive oxide film which initiates corrosion.

Hence, fretting is directly or indirectly involved in biomaterial damage by removing material from the surface or triggering other types of damage respectively. Figure 1-19 displays an SEM image taken of a CoCrMo head taper. The head has been in contact with a threaded trunnion where due to fretting, horizontal bands of material loss appeared on the surface.

[Image removed due to copyright restriction]

Figure 1-19. Fretting altered surface characteristics of a CoCrMo head taper[7]

Among the existing biomaterials, Cobalt-based alloys exhibit higher resistance to fretting in comparison to Titanium-based alloys owing to its higher modulus of elasticity (210 vs 114 GPa) [24].

Fretting is visually associated with perpendicular deformations to the original machining marks [25]. considered fretting as "*scratching perpendicular to machining lines on the taper, and/or wearing away of the machining lines*". described fretting as small scars running perpendicular to the circumferential machine lines of the screw thread.

1.4.2.2 *Corrosion*

Corrosion, in general, is an electrochemical process where the gradual destruction of materials occurs through some chain chemical reactions which induce material deposition or loss [26].

This phenomenon degrades different biomaterials distinctively. It may work locally and evolve into pitting and then cracks, or it may affect arbitrary regions on the surface of prostheses. In metallic biomaterials, degradation is more significant compared to ceramics. In ceramics, corrosion is not directly involved in failure because crack propagation is the main mechanism. Accumulation of corrosive molecules like water at the crack tip and their reactions with the ceramic molecules can lead to rupture of chemical bonds in ceramics [27].

The unique combination of strength and resistance to wear and corrosion has been desirable enough to rank metallic biomaterials as the most popular class of biomaterials. They have been chosen for orthopaedic implants because of their relatively high load-bearing capacity, low cost, and low wear rates. However, their electrochemical behaviour in the body may pose unique and specific concerns [28].

Metallic biomaterials have a high driving force to corrode, and it is their oxide film that acts as a kinetic barrier to corrosion. This micrometre thick film (passive layer) is dense and adheres strongly to the underlying metallic substrate [10]. However, highly repeated contact loading situations at taper-locked or clamped implant interfaces increase the likelihood of surface fretting wear which disrupts the oxide layer [27]. Disruption of the passive layer exposes the underlying metal to body fluids which sets in motion highly energetic reduction-oxidation (redox) reactions. As a result, positive ions are generated in the solution (ionic dissolution) and also it promptly induces establishing a new oxide film. This recurring passivation process is illustrated in Figure 1-20 and is known as mechanically assisted crevice corrosion.

[Image removed due to copyright restriction]

Figure 1-20. Mechanically assisted corrosion at articulation surfaces & taper junctions are dominant[29]

The dynamic nature of loadings leads to iterative passivations. Gradually, as the recurring redox reactions consume the ambient oxygen molecules, the level of negative ions starts to drop while positive ions such as H^+ and M^+ accumulate in the solution and makes the local net charge positive. To balance the charge in the solution, other negative ions (usually Cl^-) flow to the scene and alter the solution chemistry and consequently, behaviour.

For starters, the subsequent oxide layers will be less resistant to corrosion attacks. Also, H^+ and Cl^- ions start to combine and form HCl (Hydrochloric acid), which will increase the solution's acidification (pH level drops to below 5). These two factors intensify the corrosion damage locally, which lead to pitting and crevice types of corrosion which are *autocatalytic* in nature.

As mentioned in section 1.3.1, modular THRs have shown vulnerability to corrosion. Twisting the male taper of the neck into the female taper of the head forms a crevice space between the two opposing surfaces. Through time, the nearby body fluids leak into the crevice and form a crack-like fluid-filled environment. Repassivation within this environment alters chemical properties of the trapped solution more significantly which paves the road for crevice corrosion with a relatively high intensity [30].

Kop et al. [25] defined corrosion visually as surface irregularity with associated black debris, pits, and etch marks. Higgs et al. [6] considered corrosion as "*white haziness (indicative of inter-granular crevice corrosion), discolouration, and/or blackened debris*". Hothi et al. [31] considered corrosion regions as those with discolouration or dullness or with black debris or signs of pitting or etching.

1.5 Mitigating the Complications

Introduction of new THR systems in the market, the high cost of these surgeries, long surgery (upwards of 12 months) waiting times (if not self-insured), and complex nature of postoperative complications raised by non-specific nature of body reactions call for the vigilant selection of patient-specific implant systems with desirable functionalities. These reasons have been motivations for scientists, industrials or clinicians to become engaged with investigating the possible failure mechanisms and monitoring the performance of THRs in the market.

Mitigating the issues associated with THR undesirable postoperative responses can be achieved by undertaking different types of studies. Researchers with different areas of expertise (e.g.

biomaterial science, mechanical engineering, and medical science) endeavour to identify and formulate solutions that lessen the postoperative complications.

Reviewing the literature revealed that the existing studies concerning failure mechanisms of THRs can be categorised into three major groups of retrieval studies, in-vitro laboratory tests, and computational modelling.

1.5.1 Laboratory In-Vitro Experiments

Apart from retrieval studies that investigate damaged implants to identify root cause/s of THR failure, there are two other methodologies that serve the same purpose, yet by using different approaches.

In this type of study, biomaterials behaviours associated with their mechanical, electrochemical, geometrical, and material characteristics in various solution conditions can be examined in a laboratory environment. Unlike retrieval studies where biomaterials are assessed in-vivo, this in-vitro study simulates in-vivo circumstances by deploying testing set-ups.

Affatato et al. [32] have divided these set-ups into two categories, namely wear screening devices and wear joint devices. The first category is to provide information exclusively on the intrinsic features of biomaterials. The biomaterials can be formed into pins, disks, rings, or cylinders (Figure 1-21).

[Image removed due to copyright restriction]

Figure 1-21. Wear screening devices are used to assess intrinsic properties of biomaterials[32]

Hence, they do not represent the geometry of prostheses components accurately which results in reproducing approximate wear mechanisms that occur in-vivo. On the other side, they can handle relatively large numbers of samples within a short time.

Figure 1-22 illustrates a wear joint test set-up based on ISO 7206-4 standard to obtain the load levels required as well as the minimum number of cycles without fracture of the implant [33]. Here, real prostheses can be tested in an environment which resembles in-vivo circumstances fairly well. These set-ups can accommodate complex and dynamic test conditions which has made them a prerequisite for introducing new design and materials combinations before their mass production [32].

One example of these set-ups is the one introduced by to investigate fretting corrosion in metal-on-metal biomaterials. They simulated this type of THR by utilising a disk and pin pair with the pin located above the disk in the set-up.

The top and bottom metal samples were cone-shaped flat pin and a flat circular disk that were brought into direct contact, loaded, and moved in a small cyclic fashion using piezoelectric actuators. The interface was immersed in a phosphate-buffered saline solution.

[Image removed due to copyright restriction]

Figure 1-22. The fatigue test according to ISO 7206-4 simulates the dynamic loading of a hip stem (from www.endolab.org/implant-testing.asp)

High compressive forces (up to 3.5 kN) with a motion range of up to 140 µm were applied by an actuator to move a linear X-Y stage on which the sample chamber (connected to the disk) was rigidly fixed. A similar configuration was used to move a linear Z-stage (creating vertical motion) that was connected to the pin.

During fretting testing, parameters such as normal force, tangential interfacial force, and other moments generated about the interface were tracked by a load cell. Also, a Differential Variable Reluctance Transducer (DVRT) was deployed to characterise the effect of load on the actual motion achieved at the interface. The DVRT's role was to measure the pin-disk displacement at the closest site possible near the interface.

The measurements were used in some equations to derive other variables that relate mechanical factors to fretting and corrosion. Yet, these equations are subject to some assumptions that simplify the system or phenomenon they are addressing. These assumptions eliminate processes with negligible impact on the system.

So, these equations can represent the behaviour of the system partially. For instance, the mechanical energy dissipated in a single cycle of fretting is calculated by assuming no partial or full sticking taking place, and only full slip motion was accounted for. However, in the past, it has been established that titanium alloy when coupled by itself can reveal an adhesive galling nature which results in adhesion between surfaces when fretting takes place.

In the end, the authors have recommended for further development of the model to accommodate voltage, crevice, and solution changes to achieve higher accuracy and realistic outputs.

1.5.2 Computational Modelling (Simulation)

Simulations are based on developing 3D models of THRs and subjecting them to loads in different magnitudes, directions, and frequencies. Mixed alloy material combinations can be created by assigning mechanical properties associated with commercial biomaterials currently used in the fabrication of THRs.

Finite Element Analysis (FEA) as a popular technique allows for assessing the performance of THRs in multitudes of circumstances in terms of geometry, material combination, and load. For instance, Dyrkacz et al. [34] used FEA to discern parameters involved in micro-motion at the head-neck taper interface of modular hip prostheses. Through FEA analysis, head size, assembly force, and taper size were identified with profound influencing on micro-motion.

[Image removed due to copyright restriction]

Figure 1-23. The von Mises stress (MPa) of the femoral head and neck after applying 3300N of compressive force [34]

Figure 1-23 displays their FEA simulation which demonstrates how an inclined compressive force gives rise to von Mises stress in the femoral head and neck. FEA facilitates testing the performance of components of THRs with different geometrical and mechanical properties in any desirable configuration with relatively lower costs, risks, and time.

FEA has some advantages over the other methods. Retrieval studies are only limited to the available explants with arbitrary loading conditions, implantation times, manufacturing processes, and tolerance levels. This fact makes it difficult to isolate factors influencing corrosion damage at any taper or articulation interface. However, FEA is quite flexible in that regard. It enables creating different models and comparing the effects of the involved parameters on the damage individually.

Despite the analysis flexibility offered by 3D models, generating wear and corrosion processes in a specific material requires spending a long amount of time to encode, as the software does not provide pre-existing models for them.

In addition, there are some discrepancies between dimensions of a 3D model and its corresponding component which lead to obtaining models that may not represent the true system accurately. Lowering such errors mandating utilisation of 3D profiling technology with relatively high resolutions and accuracy. These technologies can be quite expensive and not viable for large quantities of components though.

In order to measure simulated damage, simulators measure wear after subjecting a prosthesis component to a high number of gait cycles (physiological profiles) which approximates the cycles a prosthesis undergo over time in the body.

maintained that gait activities do not comprehensively represent the normal activities of a person. Other activities, especially the ones associated with younger generations and sport, are not being accommodated. Besides, they stated the number of considered cycles only represent a few years of life. Hence, they concluded assessing wear after the components have been used in service might offer more accurate outcomes.

To summarise, the assumption made while taking on computation modelling or in-vitro studies can give rise to imprecision and inaccuracy of outcomes. However, as retrieval studies examine the actual damage and performance of THRs under working conditions in the host body, outcomes that are more realistic can be offered by them. Therefore, this research project takes on retrieval study as the methodology.

1.5.3 Retrieval Studies

As the name suggests, this type of research involves analysing the damage sustained by retrieved THRs along with the damage sustained by the host body as the result of the failure.

Investigating the postoperative complications sustained by the host body is the subject of clinical studies of retrieved implants that are usually undertaken by clinicians. Through primary and retrieval surgeries, they may encounter various host body reactions to THR systems. Sharing these observations helps patients, surgeons, and even manufacturers learn about the existing and potential defects and consequently amend the THR systems in the future to mitigate these unfavourable situations (Figure 1-24).

[Image removed due to copyright restriction]

Figure 1-24. THRs are retrieved following the failure and adverse body reactions (from http://istaonline.org)

Alongside conducting clinical investigations, examining retrieved implants and investigating the type and extent of damage they have sustained serve the same goal. As mentioned before, the scope of this research is confined to just targeting the damage on THR explants.

After obtaining appropriate visuals of damaged areas on an implant, observers can choose the suitable areas to assess the damage quantitatively or (semi)qualitatively. Quantifying the damage can be divided into two groups. The first group quantitatively assesses the damage and delves into its mechanism by using advanced analytical techniques that rely on complex apparatus. They can be split into two categories of surface profilometry and spectroscopic techniques. There exists contact surface profilometry (roundness measuring machine, gravimetric wear assessment, and coordinate measuring machine) and non-contact surface profilometry (e.g. RedLux) that quantifies damage as material loss or deposit on the surface. Also, techniques such as X-ray Photoelectron Spectroscopy and Energy Dispersive X-ray can be used to quantify the elements or compounds that sits in the corroded zones at the surface.

Owing to the high precision of measurements that are carried out by this group, they may demand considerable time and financial resources. Therefore, the failure of only a small (below 10) sample size can be investigated by these techniques. On the other hand, the second group is not concerned with details of failure mechanisms and utilises cost-effective methods to (semi)qualitatively quantify the damage which makes conducting large-scale retrieval studies feasible.

(semi)Qualitative assessment is a subjective analysis where the output is an ordinal rather than a ratio metric of the damage. This type of analysis assigns scores (levels) to the damage by visual inspection based on several existing visual criteria. Therefore, this approach is also known as visual scoring. In retrieval studies of modular implants, several scoring methods have been developed to quantify fretting or corrosion damage at various junctions and components of hip replacement devices. Each method may score the component of interest holistically or assign an individual score to various predefined zones on the surface. Also, fretting and corrosion may be scored separately or together.

[Image removed due to copyright restriction]

Figure 1-25. Visual assessment of corrosion scores the damage [35]

A common peer-reviewed method to implement this approach at taper junctions is based on visual scoring (Figure 1-25) of the damage that was introduced in 2002 by Goldberg et al. [24]. It is a four-level scoring method based on some predefined visual definitions in terms of the shape, colour, and reflectivity of the damaged areas. Table 1-4 summarises the visual criteria to distinguish between the damage levels and rank them. As of date, this paper has been cited 160 times, and this scoring method or its modified versions have been used by many large-scale retrieval studies [6, 7, 23-25, 30, 31, 36-39].

Table 1-4. The scoring system developed by Gilbert is used to visually rank corrosion and fretting damage [24]

Score	Corrosion Criteria	Fretting Criteria
1 (None)	No Visible Corrosion	No Visible Fretting
2 (Mild)	<30% Surface Discoloured / Dull	Band(S) for Fretting Scars Across ≤3 Machine Lines
3 (Moderate)	>30% Surface Discoloured / Dull or <10% Containing Black Debris, Pits or Etch Marks	Band(S) Involving >3 Machine Lines on Taper Surface
4 (Severe)	>10% of Surface Containing Black Debris, Pits, or Etch Marks	Several Bands of Fretting Scars Involving Several Machine Lines or Flattened Areas With Nearby Fretting Scars

This research does not utilise the first group of analyses since it is a large-scale retrieval study. It finds the linkage between the observed damage on explants and the factors that may have contributed toward the failure. These factors stem from the recipients (retrospective clinical) or the design of implants as summarised by Table 1-5. The information obtained via this methodology can be used to predict or pre-screen in-vivo or in-vitro performance of implant systems [17].

Table 1-5. Some patient and implant factors that can be correlated with the fretting and corrosion damage

Patient Factors		Implant Factors	
Age	Sex	Manufacturer	Assembly Conditions
BMI	Serum Ion Concentration	Component's Material	Component's Size
Implantation Time	Reason For Revision	Surface Topography	Wear Rate

In the next chapter, a comprehensive literature review of the large-scale retrieval studies is conducted to learn about the scoring methods, the influence of patient and implant properties on fretting and corrosion, and most importantly the common limitations and concerns that they faced.

1.6 **Summary**

This chapter endeavoured to provide an overview of the various aspects of hip replacement prostheses and analytical methods to evaluate their in-vivo and in-vitro performance.

The next chapter will elaborate on the large-scale retrieval studies that have conducted so far to work out the existing methodologies to quantify fretting or corrosion and also find the associations between the damage and patient/implant properties.

2 LITERATURE REVIEW

[Image removed due to copyright restriction]

2.1 Introduction

This chapter reviews the literature of large-scale retrieval studies to investigate the existing methods, the findings, and any potential gap in the literature. To address the identified gaps and concerns in this context, other spheres of literature will be explored to learn how some particular issues have been addressed within other contexts which have encountered similar challenges. Next, the situation that is faced in this particular context is compared with those of the other contexts to highlight the potential limitations that need to be addressed in the methodology.

2.2 Large-Scale Retrieval Studies

Since the early 1990s [40, 41], several large-scale retrieval studies have been conducted to investigate the potential associations between the failure of hip replacement implants and implant/patient properties. They used specific implant properties, patient information, sample size, visual scoring method, and statistical analyses for their investigations.

Goldberg et al. [24] analysed 231 modular hip explants to find out how factors such as material combination, metallurgic condition, flexural rigidity, head and neck moment arm, neck length, and implantation time may affect in-vivo corrosion and fretting of modular taper surfaces. Optical microscopy with magnifications between 3.5X to 40X (based on the required detail level) was utilised for inspecting and scoring. One researcher did the inspection and scoring for consistency reasons. He divided the necks and the heads into four (medial, lateral, posterior, and anterior) quadrants which were further split into two (distal and proximal) zones, and scored them for fretting and corrosion. The scores for each component were combined so that single global scores for fretting and corrosion, the most representative of the damage over an entire component, could be obtained.

Table 2-1. The comparative impact of the investigated factors on the observed damage

Observed Damage		Fretting	Corrosion	Localised Damage		Fretting	Corrosion
	Neck	*	*		Neck	*	*
	Head	*	**		Head	*	***
Longer Implantation Time				Moment Arm			
	Neck	-	up		Neck	-	***
	Head	down	up		Head	***	***
Higher Flexural Rigidity				Neck Length			
	Neck	-	down		Neck	-	-
	Head	down	down		Head	-	-
Material Combination				Regional Damage			
	Mixed	**	**		Proximal	*	*
	Similar	*	*		Distal	*	***

Corrosion scores were higher at the heads with respect to the necks. The scores observed to be correlated. Fretting scores of the heads were higher than those of the necks. The two scores behaved similarly. Corrosion and fretting of mixed alloy samples (both head and neck) were found to be significantly higher than those of the similar alloys. The couples (head and neck) with longer implantation times had higher corrosion scores. While neck fretting seemed to be unaffected by implantation time, head fretting decreased with longer implantation times which is due to etching of fretting scars [24]. The correlations identified between the observed damage and implant factors at head-neck junction are listed in Table 2-1.

Since publishing this scoring method, there has been a fair number of studies that utilised this scoring model to characterise fretting or corrosion at head-neck and other modular or articulation interfaces. Many of these works adopted the same model, and some have introduced an entirely new or a modified version of Goldberg's scoring method which more or less stratify the degree of damage.

This fact was the reason for a group of American researchers to introduce a modified version of the original method with the collaboration of the senior author of that work 11 years later. Higgs et al. [42] conducted a retrieval study of 76 explants. The examined components were comprised of 76 heads. 31 stems (22 modular necks), 10 modular acetabular liners, and 5 corresponding acetabular shells. The interfacing components were fabricated by the same manufacturer (no mixed & matched sets) [42]. This modified scoring method is summarised in Table 2-2.

Table 2-2. Modified version of fretting and corrosion criteria that were established in 2002 [42]

Damage	Score	Criteria
Minimal	1	Fretting on <10% surface and no corrosion damage
Mild	2	Fretting on >10% surface and/or corrosion attack confined to one or more small areas
Moderate	3	Fretting on >30% surface and/or aggressive local corrosion attack with corrosion debris
Severe	4	Fretting over majority (>50%) of mating surface with severe corrosion attack and abundant corrosion debris

Fretting and corrosion were visually identified based on the definitions given by literature. Unlike the original method in which 2 and 4 regions were individually scored, here, each interface was holistically assigned a single score. Each score represents both fretting and corrosion damage due to the synergistic nature of them. Table 2-3 lists the outcomes of the scoring and association between the implants and patient properties [42].

Table 2-3. Summary of the qualitative analysis

Damage vs. Implantation Time	head tapers	modular stems	**Mild Damage %**	head tapers	male neck tapers
	↗	↗		89	91
	taper II	monolithic stems		stem tapers	shells & liners
	↗	X		68	100
Modular Necks	medial	lateral	posterior	anterior	
	*	*	-	-	
Effect of Neck Modularity	heads paired with modular necks		heads paired with monolithic stems		
	**		*		
Observed Damage	Modular Necks				
	head	neck			
	**	*			
	Monolithic Stems				
	head	stem			
	*	*			
material combination	mixed head	similar head			
	***	*			

Kocagoz et al. [7] investigated whether any correlation exists between taper angle clearance and visual fretting-corrosion scores in ceramic (n=50) and CoCrMo (n=50) heads. The values of angle clearance were calculated through roundness measurement (will be introduced later). Then, they applied this four-level scoring model on the images obtained by optical instrumentation.

Failure of 74 implants from two different manufacturers was investigated by Dyrkacz et al. [36] to elucidate any the role of head size and manufacturer in the fretting and corrosion behaviour along bore taper of the head and neck taper of the stem. Taper interfaces were divided into four quadrants with each comprising a superior and an inferior region. Hence, they ended up with eight regions for a neck and eight regions for a head. Each region received two scores, one representing the severity of fretting and corrosion and the other for the amount of damaged area. These two scores were multiplied to form eight regional scores for fretting and corrosion. By adding up these scores, separate scores for the head and neck were obtained. As this process was being performed by three observers, the obtained scores were averaged.

Their study revealed significantly higher scores at explants with the larger head size (36mm) and increased head sizes result in a higher likelihood of fretting and corrosion at head and neck junction. They hypothesised it had been associated with the greater torque they had experienced over their taper interfaces which in turn had resulted in more micro-motions between the heads and the necks. Also, they observed a poor correlation between fretting and corrosion damage versus implantation time [36].

The hypothesis that with head-stem pairs, ceramic-metal tapers (n=50) are less susceptible to corrosion than metal-metal tapers (n=50) was investigated by Kurtz et al. [30]. Ceramic heads were scored based on the observed metal transfer to their bore tapers. Stem tapers with ceramic heads received lower fretting-corrosion scores (Table 2-4). Hence, they maintained that by using ceramic femoral heads, fretting and corrosion at head-neck junctions might be mitigated. Also, they advised visual scoring does not necessarily correlate with the volume of metallic debris generated at a modular interface, and further taper analyses to quantify material loss at ceramic-stem modular connections are required. Yet, they believed their scoring technique was consistent with the approach of other investigations.

Table 2-4. Influence of head size on the damage at neck

Head Material		Fretting	Corrosion
	CoCr	**	**
	Ceramic	*	*

Another study in regard to evaluating the reliability of visual scoring was conducted by Hothi et al. [31] to investigate fretting and corrosion in 150 metal-on-metal bearing of implans. Female tapers of femoral heads with large heads (>36mm) were examined. Each distal and proximal region received fretting and corrosion scores. By combining the two regional scores, two overall scores (fretting and corrosion) were obtained for each taper. They noticed the scores are not normally distributed, and corrosion scores were higher than fretting scores. Also, while a moderate association between Volumetric Material Loss (VML) and corrosion scores was witnessed, the correlation between VML and fretting scores was weak (Table 2-5).

Table 2-5. Comparison between fretting and corrosion damage extent

Observed Damage		Fretting	Corrosion
	Neck	-	-
	Head	*	**
VML		Fretting	Corrosion
	Neck	-	-
	Head	*	**

They maintained it could be due to the central role of galvanic corrosion in comparison to fretting corrosion on material loss, and also the relative difficulty in measuring fretting scores. Therefore, it was recommended to investigate the association between VML and blood metal ion levels.

In another study by the same investigators, Hothi et al. [38] analysed 20 explants that were comprised of 36mm metal on metal Pinnacle heads paired with Corail (n=10) or S-ROM (n=10) stems. They used the Goldberg method to score the head tapers for corrosion based on visual assessment. No significant difference between the corrosion scores at head tapers paired with both types of the stems was witnessed. Also, rougher and shorter stem tapers for Corail designs were observed.

In a novel study by Goyal et al. [37], it was investigated how retaining a corroded stem and just replacing the head in a revision surgery impact the survivorship of the new implant system. As the stems were not accessible, they scored 86 retrieved head tapers (Co-Cr alloy). Each retrieved head was photographed in five different views to cover the entire taper surface. In order to assess the effect of the corrosion at the time of revision on the likelihood of second revision, they defined high and low levels of corrosion when the assigned scores belonged to 3 – 5 and 1 – 2 intervals respectively.

No correlation between the scores and the time in-situ, head type, taper size, and head-diameter (28 and 36mm) was observed. However, patients with moderately higher BMI had greater corrosion scores. In the seven cases with re-revision, no corrosion-related disease had been diagnosed. The extent to which the head taper had been corroded observed to have no effect on the survivorship in both high and low-level heads. Hence, they supported retaining the well-fixed femoral stems with corrosion. Due to the nature of this study, only femoral heads were scored, and these heads were apparently paired with stems that had already sustained damage to some extent.

Molloy et al. [23] reviewed the results in 15 recipients of ABG II (Stryker Orthopaedics, NJ) modular hip systems. The head was BIOLOX delta ceramic (CeramTec, Plochingen, Germany), the neck was a Ti-based alloy, and the stem was fabricated by CoCr. They performed a retrieval study of fretting-corrosion at the modular junctions (head-neck & neck-stem tapers) of 7 recipients who had undergone revision surgery due to ALTR. They examined and rated fretting and corrosion at the male tapers by using a reflected-light stereomicroscope at up to 10X magnification. Fretting and corrosion were observed to have different intensities at these two junctions (Table 2-6). Superior section of neck-stem taper of neck exhibited the most severe corrosion.

Table 2-6. Fretting and corrosion scores witnessed at the male tapers

	Fretting	Corrosion
Head-Neck	mild	no or very little
Neck-Stem	no fretting	moderate to

The influence of taper design and head size on fretting and corrosion at head-neck junction in metal on metal hip arthroplasties with large heads (>= 36mm) was investigated in another study [39]. 40 retrieved heads were subjectively graded for fretting and corrosion using Goldberg method. For each taper, the scores were summed up and averaged between the two observers, so that a single score for fretting and a single score for corrosion at each taper could be obtained. They categorised implants into three taper groups, namely 11/13, 12/14, and type 1. Each taper design had different values for angle, distal diameter, and contact length.

Unlike many past studies, the only observed correlation was between fretting damage and taper geometry. Tapers with thicker and longer contact lengths observed to be more prone to fretting. Hence, they concluded a thinner and shorter taper is more beneficial than a longer and thicker taper. Yet, these three taper designs exhibited no influence on corrosion scores and VML (Table 2-7). Also, parameters such as head size, lateral offset, and implantation time did not correlate with fretting, corrosion, or VML.

Table 2-7. The impact of taper design and location on quantitative and qualitative assessment of damage

		11/13	12/14	Type 1
Contact Length & Thickness	Fretting	***	**	*
	Corrosion	*	*	*
	VML	*	*	*
Regional Damage		Fretting	Corrosion	
	Proximal	-	*	
	Distal	-	**	

They pointed out that the heterogeneous nature of the cohorts and respectively low number of explants has made their dataset underpowered which can lead to overestimating the prevalence of damage at head and neck junctions.

Higgs et al. [6] performed a semi-quantitative evaluation of modular interfaces along with a review of the clinical records. In the first part, they investigated the correlation between modularity and fretting-corrosion damage in 137 metal on metal explants at head-stem and shell-liner interfaces. In their study, patient and implant variables such as the type of paired alloys, head size, medio-lateral offset, and neck moment arm were investigated.

They also found increasing modularity escalated fretting-corrosion damage at bore taper of the head. The evaluations revealed that dissimilar alloy pairing, larger head sizes, increased medio-lateral offsets, longer neck moment arms, and distal neck tapers were associated with increased taper damage at the modular taper interface. Similar to the other investigations, they raised the subjectivity of visual scoring as one of the limitations that may lead to issues in characterising amounts of VML and corrosion debris at taper junctions.

Increased modularity was also investigated in another study to learn more about its role in the failure of THRs [43]. They conducted a retrieval study of 57 explants from seven Double Tapered Cone (DTC) THR designs. Stem trunnion and neck trunnion of the necks were assessed to compare and contrast degradation mechanism/s at the neck-stem junction to its head-neck counterpart in terms of the nature and its contribution toward the failure of the studied THRs. They introduced a modified version of the Goldberg method in which corrosion is graded in five levels, and fretting is assessed using a binary model that states whether fretting is present or not. The observers also commented on the presence of mechanical movement that indicates whether long scratches on the taper were induced by initial application of seating load to the taper or introduced during revision.

In relation to degradation and junction stability, Ti-based modular necks observed having an additional locking mechanism for the neck-stem junction, while the Co-based devices observed relying entirely on mechanical stability through the design of the trunnion. They maintained that the influence of this feature on reducing micro-motion is unknown. No correlation was observed between trunnion machine finish and damage scores. Yet, they recommended further investigation into its possible impact on corrosion susceptibility. They observed higher fretting scores in neck-stem than the head-neck junctions. Therefore, they concluded that increasing modularity by introducing the neck-stem junction may come at a cost. Higher rates of fretting and crevice corrosion witnessed at this additional junction can justify the rise in metallic debris and soluble metallic ions [43]. Table 2-8 summarises the findings of this study.

Table 2-8. Effects of material and modular junction on the degradation of THRs

	Ti-Based	Co-Based
Fretting %	50	90
Corrosion %	30	62
Degradation	Taper I	Taper II
	*	**

With respect to increased modularity, a retrieval study in Australia used Goldberg method for semi-quantitative assessment of corrosion in 16 explants that belonged to double-tapered cone (DTC) Margron necks (Portland Orthopaedics, Matraville, NSW, Australia) that were paired with alumina Biolox heads [25]. They observed no correlation between fretting and corrosion scores and any patient or implant properties. Their study maintained that despite the availability of modern taper designs and corrosion-resistant materials, increasing the modularity can lead to fretting and crevice corrosion. Hence, they recommended further optimisation of the taper design and material.

The influence of three implant characteristics, namely neck-shaft angle, stem size, and overall neck length on fretting and corrosion was the subject of a study by De Martino et al. [44]. They analysed 60 Rejuvenate (Stryker Orthopaedics, NJ) explants. Two orthopaedic surgeons independently assess the presence and severity of damage by a stereomicroscope. The stem tapers were divided into eight zones similar to the Goldberg method. However, head tapers were divided into four quadrants. The visual criteria in Goldberg scoring technique were utilised to score the components. Neck-shaft length and stem size observed to not correlate with fretting and corrosion damage at the necks and stems. Likewise, no association was witnessed between overall neck lengths and damage in the neck components. Also, implantation time had significant positive correlations with the damage on both the neck and stem components. Table 2-9 provides further information about the relative extent of damage on different zones of the explants.

Table 2-9. A comparative list of damage scores at different locations on the explants

NECK Corrosion		anterior	posterior	medial	lateral
	distal	**	**	****	***
	proximal	*	**	**	**
NECK Fretting		anterior	posterior	medial	lateral
	distal	**	**	****	***
	proximal	*	**	**	**
	STEM Corrosion	anteior	posterior	medial	lateral
		***	***	****	***
	STEM Fretting	anteior	posterior	medial	lateral
		**	*	*	**

Early (between 0.8 and 3.1 years) revision of 19 Rejuvenate implants was investigated by Lanting et al. [45] Similar to some other studies, stems were divided into four quadrants before going through the 4-level scoring of fretting and corrosion under stereomicroscopic visualisation. After the visual inspection, as with the anterior and posterior sides, the superior and inferior scores were mutually combined owing to similar degrees of observed damage. Table 2-10 lists the relative extent of damage on these two combined zones.

Table 2-10. Relative comparison of damage in the combined zones on the surface of the necks

	ant/post	sup/inf
Fretting	**	***
Corrosion	*	*

The role of using heads and necks from different manufacturers at corrosion in 151 retrieved implants were investigated by Whittaker et al. [46]. In this study, 51 heads were identified as having stems from different manufacturers. Goldberg scores and VML were measured at this junction, and no significant difference in corrosion between these two groups was observed. The inclusion criteria demanded modular components with Co-Cr metal-on-metal bearings and a 12/14 taper which had been revised after a minimum of 12 months.

Triantafyllopoulos et al. [47] conducted a retrieval study of 154 large diameter metal-on-polyethylene implants to investigate whether head size, implantation time, implantation time, taper design, and alloy combination influence the fretting and corrosion at head-neck taper junction. Two investigators used the Goldberg method and considered the total score as the summation of region scores. While head size did not significantly contribute toward fretting and corrosion, the remainder of the investigated properties observed to affect the damage.

The topography of circumferential machining lines on the surfaces of tapers was the focus of a novel study to investigate its role in fretting and corrosion [48]. Unlike the previous studies [49, 50] that had used average roughness as the descriptor of surface topography, this study used machining mark height and spacing to quantify the topography. A modified Goldberg method was used to score damage at 140 Co-Cr/Co-Cr and 129 Co-Cr/Ti head-stem couples. Two investigators scored fretting and corrosion combined. Between the two material couples, very similar average damage scores of stem and head tapers were observed. Also, Co-Cr and Ti tapers appeared to have almost identical surface topography which maintained to be aligned with the findings of some previous similar studies. For Co-Cr/Co-Cr couples, the outcome of this work suggests that stem taper machining mark height was associated with higher stem and head taper damage.

This section reviewed several large-scale retrieval studies that semi-quantitatively measured corrosion at the head and stem junction of hip replacement devices. It was observed that Goldberg scoring method is still widely in use and some modified version of that have been introduced since 2002. Also, the analysis of associations between patient/implant properties and corrosion has led to conflicting results in some studies which can be due to variations in their statistical power and the difference within the set of the investigated patient and implant properties.

With approximately 42,000 hip replacement surgeries per year, Australia is one of the main users of these products. 4,500 revision surgeries are undertaken each year to remove or reinsert malfunctioning implants [51]. Despite the clinical significance and high-incurred costs of revision surgeries together with the availability of large pools of failed implants, retrieval investigations on the fretting corrosion failure of THRs are still new in Australia and need development. Review of the literature shows only a limited number of retrieval works on the fretting corrosion damage to the head-neck taper junction of explants in Australia. The size of investigation pools ranges from 7 to 57 retrievals [23, 25, 43, 52, 53]. In the three larger studies, the variety of implants were either quite limited or belonged to some groups which are not popular in Australia anymore. Annual reports of Australian National Joint Replacement Registry indicate that Exeter V40 (Stryker) and CPT (Zimmer), have been among the ten most used femoral stems for THRs for more than a decade. None of these five studies included these popular designs in their investigations. These facts justify and signify the necessity of establishing retrieval libraries and programs that systematically investigate the failure of retrieved hip replacement devices in Australia.

2.3 Reliability of Visual Scoring

More recently, the reliability of visual assessments has been questioned by some researchers [6, 30, 31, 39, 54]. Some have tried to evaluate the reliability of damage scores by measuring their association with the measurements of quantitative methods (i.e. surface profilometry). These investigations endeavoured to verify inter-observer reproducibility and single-observer repeatability of measurement performed using visual scoring (Figure 2-1). Inter-observer reproducibility is the uncertainty in measurements when different observers measure the same thing under the same circumstances. Single-observer repeatability is the uncertainty in measurements when an observer measures the same thing a number of times under the same circumstances. It can be calculated as the standard deviation of a measurement that is replicated several times.

[Image removed due to copyright restriction]

Figure 2-1. Reliability of a method to perform an experiment is based on its repeatability and reproducibility *http://blog.f1000research.com*

The statistical analysis performed by Hothi [31] revealed that the inter-observer reproducibility of fretting scores could be vulnerable. It stems from the fact that fretting scars might be obscured by the material deposited on different parts of an explant surface due to corrosion. Also, some scars on the surface of an explant (such as taper junctions) can be due to damage caused by impaction during assembly or disassembly of tapers. Goldberg et al. [24] maintained that distinguishing between fretting damage attributable to impaction and actual fretting may be done differently amongst the observers. Hence, qualitative assessment has a predisposition to observer subjectivity. Also, Higgs et al. [6] believed different laboratories conducting such observations are subject to the lack of standardisation in visual scoring.

These concerns can give rise to an increased risk of false evaluation which has been the motivation for some studies [31, 54-57] to evaluate the reliability of visual scoring. Assuming VML as a valid reference point for damage, they measured it at a (limited) number of taper junctions or bearing interfaces, and the results were compared with the associated visual scores. Based on the calculated inter-observer reproducibility and single-observer repeatability, these studies concluded that the obtained visual scores could be moderately correlated with the VML measurements, yet visual scoring cannot substitute measuring VML.

This situation is not confined to scoring fretting or corrosion damage in retrievals. The majority of condition assessment units require periodic inspection and monitoring of tribounit systems that similar to THR devices operate under chemical-mechanical conditions and are subject to some combination of oxidation and erosive wear (fretting corrosion) [58]. The condition rating methods proposed by expert bodies such as ASTM (D610-08) or Japan Association of Steel Bridge Construction (Specification for Highway Bridges I & II) uses visual criteria to classify the degree of rusting on coated steels which require qualified staff and the use of subjective criteria [59-71]. In the field of Structural Health Monitoring, painting warranty clauses mandate both the owner and contractor inspect painting of steel bridges regularly for any defect (i.e. rust). Through this process, a defect is assigned a level based on its percentage, and accordingly, the painting has to be redone partially or entirely. Here, the subjectivity of the inspection may result in an argument between the owner and contractor [72].

Besides the visual inspection, there are some studies that questioned the suitability of measuring VML for evaluation of non-uniform corrosion damage. They raise two general concerns in relation to VML quantifying corrosion damage. First, VML does not reflect the distribution of corrosion damage. A corroded component with a mild level of damage which has spread widely across its surface may return a similar VML value to another component with severe but local corrosion. Second, the damage level 4 in Goldberg's scoring method has observed to have a relatively large variation in VML values [31].

These reliability concerns in regard to visual scoring and VML are among the reasons for researchers to search for alternative methods that can provide feasible yet reliable outcomes in various corrosion related applications.

2.4 An Alternative Method

Processing optical and electron microscopy images obtained from various surfaces have been successfully practised in the past to detect and classify the severity and typology of corrosion generated defects. Digital image acquisition, processing, and interpretation techniques offer several benefits such as finding information hidden from the human eye, identification and quantification of common numerical features in the same class images, and developing image databases that can be shared and used in future for testing more advanced processing algorithms [73].

Here, two distinct disciplines are involved, namely Digital Image Processing (DIP) and machine learning. These two disciplines are so closely related that it is inevitable to discuss DIP applications without elaborating on techniques in machine learning [74].

DIP may assume two roles. First, it can be used for image enhancement (first step in many vision tasks) purposes by benefiting from tools such as noise suppression filters, histogram equalisation, and contrast stretching. These tools can be used to highlight regions of interest in an image (e.g. wear or corrosion spots) and attenuate the present noise. Second, it can be used to extract characteristic features from images via quantifying surface morphology which is believed to be closely related to the degree of corrosion damage [61].

On the other hand, machine learning acts as a decision-making tool in situations where the decision-maker is a single trusted person, or a committee of experts, who despite their long experience cannot express their reasoning in any sensible way [75].

The corrosion-related studies have deployed different (un)supervised techniques in various contexts such as steel bridge, atmospheric corrosion, and steam boilers for the following defect (rust, crack, and wear) classification purposes.

- Analysis of shape and size of defects (e.g. pits): [58, 76-81]
- Identification or classification of corrosion severity: [59, 60, 62, 64-66, 68-71, 73, 82]
- Identification or classification of corrosion typology: [61, 83-85]

In relation to corrosion typology, using optical images may not be a suitable option and electron microscopy methods (e.g. Scanning Electron Microscopy, Transmission Electron Microscopy, and Atomic Force Microscopy) which reveal the microstructure is conventional. There are a lot of studies that analysed microstructure images and derived information from them regarding the corrosion typology which is out of the scope of this research.

2.5 DIP and Machine Learning in Classification Problems

Large-scale retrieval studies need a metric for corrosion severity. They rely on visual scoring and use predetermined visual criteria to assign corrosion scores subjectively. In other words, the implants are classified into several groups according to their corrosion severity. Therefore, we are facing an image-based rating problem in which images are to be classified according to a specific scoring method such as Goldberg.

This approach has not been implemented in the past to score the corrosion objectively in large-scale retrieval studies. Therefore, this section reviews the works that endeavoured to detect or rate corrosion severity via an intelligent rating of optical images. Reviewing the literature revealed that such studies belong to a larger group that uses morphological attributes to extract a vector of characterising features from an image that later on is used as the input for machine learning algorithms (Figure 2-2). Therefore, it is essential to be reasonably familiar with the DIP and machine learning algorithms that are being developed in such studies.

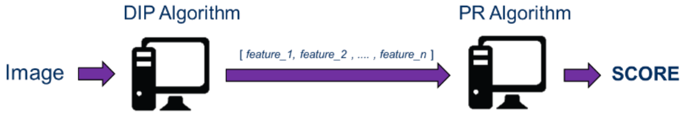


Figure 2-2. An overview of intelligent image-based rating methods

It should be noted that many of these algorithms have been around for many decades. However, they only became (or are still becoming) applicable after breakthroughs in technologies such as high-resolution cameras and powerful computer hardware [75].

2.5.1 DIP Algorithms

The morphology of defects can be used to formulate and extract three groups of features, namely colour, texture, and shape [61]. Corrosion defects usually appear as scattered random spots. Therefore, shape as a morphological attribute is not used for characterising corrosion defects.

2.5.1.1 *Colour Features*

This type of features is derived from the frequency histograms of colour components. Colour components result from segmentation of images according to the colours that they contain. Colour cameras reproduce colour by sensing the optical energy in three overlapping wavebands and thereby generate three separate signals (components) namely, Red-Green-Blue (RGB). Colour science suggest other representations for colour (i.e. colour models) for different vision tasks [75]. These alternate colour models express the RGB model in alternative coordinate systems [86].

[Image removed due to copyright restriction]

Figure 2-3. RGB colour space forms a 3D cubic coordinate system[75].

They present colour information in ways that make certain calculations more convenient (Figure 2-3). Another colour model in image analysis applications is the HSI (Hue, Saturation, and Intensity or Value) model. Hue and Intensity components are approximately proportional to the average and sum of RGB components, respectively, and Saturation quantifies the deficit of white colour [86]. In an HSI image, colour values are described by the Hue (e.g. red) component. As hue varies from 0 to 1.0, the corresponding colours vary from red through yellow, green, cyan, blue, magenta, and back to red, so that there are red values at 0 and 1.0. The corresponding colours (hues) vary from unsaturated (shades of grey) to fully saturated (no white component). As Intensity (luminance) varies from 0 to 1.0, the corresponding colours become increasingly brighter.

[Image removed due to copyright restriction]

Figure 2-4. HSI colour space[74]

HSI colour model allows for separation of colour from the lighting to a greater degree in comparison with the RGB model. This fact can benefit image analysis in two ways. First, image enhancement which is based on the manipulation of contrast, may not work effectively for RGB images. Transformation of the RGB values of the pixels to improve contrast may alter the chromatic (i.e. the colour) content of the image. However, HSI colour decouples the chromaticity (H and S) from the intensity which allows for fine-tuning the intensity without altering the colour information [74]. The second advantage of using an HSI colour model benefits characterising colour as an attribute of the morphology regardless of possible variations in luminance. This type of variation can increase the diversity among the images that belong to the same class by throwing away information and introducing noise into the system [87]. Luminance variation may stem from non-uniform scene illumination and image formation process in the image acquisition device.

Therefore, in many studies, before extracting colour features, the colour model of images is converted from RGB to HSI and the Intensity component is ignored [59, 64, 68, 71, 84, 88]. These studies calculated several numerical features from the brightness histograms of the Hue and Saturation components. The histogram for an image colour component demonstrates the frequency of occurrence for each pixel value and can be regarded as the probability density of an image pixel having a certain brightness [64, 86]. In an 8-bit image, a pixel brightness is a discrete number which may span from 0 (black) to 255 (white) [86]. The histogram is considered as one form of global information about an image. An image matrix may contain much data which slows down the processing tasks. Using global information in an image histogram is more concise and occupies less memory.

There are several features that have been extracted directly from the histograms of image colour components in various studies. These features are known as first-order image statistics. Some include moments, entropy dispersion, mean (an estimate of the average intensity level), variance (this second moment is a measure of the dispersion of the region intensity), mean square value or average energy, skewness (the third moment which gives an indication of the histograms symmetry), and kurtosis (cluster prominence or ‘‘peakedness’’) [75].

One limitation of first-order statistics is that they only reflect on individual pixel intensities and provide no information about the relative position of pixels (spatial information) to each other [75, 89].

2.5.1.2 *Texture Features*

Using intensity histograms is associated with some drawbacks, though. A histogram does not provide any information about the order of the measured intensity frequencies. In other words, a histogram does not reflect the spatial relationships between the pixels and only provides information about them individually.

The texture is considered as an important morphological attribute of corrosion defects because the corrosion process can deteriorate the surface of metals and produce rough surfaces. Consequently, the corroded surface will have a different texture to the rest of the image [61, 63, 64].

Texture is a term that refers to properties that represent the surface of an object. It is loosely used to describe the 'roughness' of something. Despite being widely used, and perhaps intuitively obvious, there is no precise objective definition for that due to its wide variability [74, 86]. It may be described subjectively using terms such as coarse, fine, smooth, granulated, rippled, regular, irregular, and linear. Therefore, some more precise properties must be defined to be able to use it in machine vision tasks.

One popular technique of texture analysis in corrosion-related studies is based on using grey-level co-occurrence matrix (GLCM) [60, 61, 64, 67, 68, 70, 82, 83]. GLCM is the matrix of relative frequencies $P_{\emptyset,d}$ (a,b) that describes how frequently two pixels with the grey levels *a* and *b* appear in a window separated by an offset distance *d* in direction Ø. The dimensions of a GLCM depends on the range of grey levels that it covers. In a digital 8-bit image, there exists up to 2^8 distinct grey levels which yield up to 256 × 256 GLCMs. There is no standard approach for finding the optimum distance and direction values in a particular problem [90]. Memory requirements may mandate scaling the grey level values to a smaller set, but it results in a reduced grey level accuracy [86].

The diagonal elements of GLCM correspond to the histogram. In an image with low contrast, the elements of the GLCM that are far from the diagonal are equal to zero or are very small. For high contrast images, the opposite is true. From GLCM, several numerical features can be extracted. Examples include Energy, Entropy, Contrast, Correlation, and Homogeneity [86, 91]. Due to the statistical nature of the GLCM method, the features that are extracted via this method are known as second-order image statistics.

2.5.1.3 *Wavelet Features*

It should be noted that texture description is highly scale-dependent. When texture elements are large enough, they might be defined in more than one scale. Therefore, image resolution (scale) must be a consistent part of the texture description [86]. maintained that those features extracted from second-order statistics could be inadequate since at a particular scale, some textures with the same numerical features can be easily discriminated by human visual systems. To decrease the problem of scale sensitivity, the texture needs to be described in multiple scales by using a coarse-to-fine multiresolution strategy [86]. Over the past two decades, multiresolution analysis techniques such as wavelet transform have received a great deal of attention in various applications of DIP [66, 67, 69, 71, 77, 78, 92-95].

It is well established that gradients of intensity in various directions over different scales reflect the texture of an image [96]. An effective method to capture image gradients at different magnitudes and directions is to use 2-D Fourier transform (FT). By using this transform, an image can be reconstructed to obtain a frequency-space representation. In the context of corrosion, defects appear as abrupt spatial changes in intensity which result in high frequencies in the corresponding FT reconstruction. By using this property, high-pass (HP) and low-pass (LP) filters can be created that pass specific spatial frequencies into the reconstruction while suppressing the others. These filters can be used to visually highlight texture features with particular magnitudes or directions [74].

A wavelet transform is computed by the convolution of the signal (image) and the scaled-shifted versions of a mother wavelet function (e.g. Daubechies, Haar, BiorSplines, and Gaussian). Figure 2-5 shows single decomposition wavelet transform that is implemented by a bank of 1-D HP and LP filters.

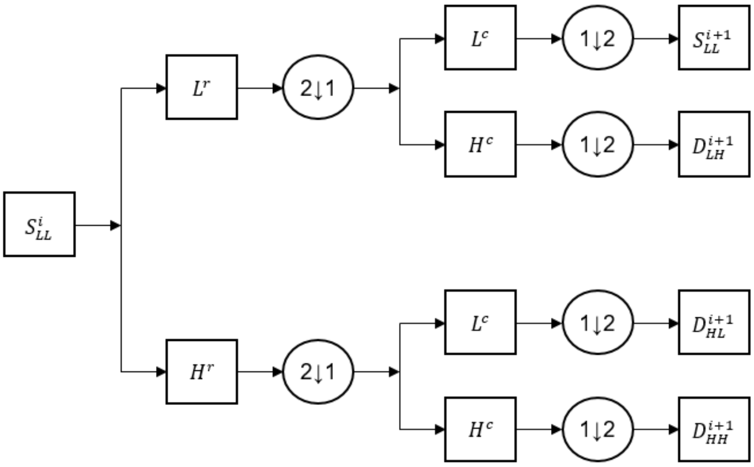


Figure 2-5. 2-D discrete wavelet decomposition

Here, S^{i}_{LL} is the input image at resolution level *i*. According to the Nyquist's rule, decomposition of an image with an LP or an HP filter yields almost twice as much data. In order to keep the amount of data almost the same size as the input, the data was down-sampled following each filtering (which is represented by the circle blocks in Figure 2-5). *H* and *L* represent 1-D HP and LP filtering and the *r* (rows) and c (columns) superscripts denote the direction of down-sampling. The circle blocks which perform down-sampling are 2↓1 (keeping one column out of two) or 1↓2 (keeping one row out of two) [66, 95].

The result is the decomposition of the input image into four sub-band images. D^{i+1}_{LH}, D^{i+1}_{HL}, and D^{i+1}_{HH} correspond to the low-high, high-low and high-high bands in the frequency domain, respectively. Also, these images highlight the horizontal, vertical, and diagonal details in the input image. S^{i+1}_{LL} (low-low component) is an LP filtered version of the input image that would be the input for further wavelet decompositions.

At each decomposition level, the wavelet transform coefficients of the three detail sub-band images produce several textural features. Among them, wavelet energy signatures which reflect the distribution of energy along the frequency domain are often employed as corrosion texture features. The energy signature arrays are denoted by E_{LH_i}, E_{HL_i}, and E_{HH_i}. The length of each arrays is equal to *i* (decomposition level) [66, 93, 95, 97].

2.5.1.4 *Local Features*

Although wavelet transforms address the scale sensitivity via multiresolution image processing, there exists other imaging deformations that wavelet transforms are not invariant to. These imaging conditions that have been considered by many studies [97-105] as serious issues against the application of global features are categorised as (1) geometric or affine (scale, translation, and rotation) distortions and (2) photometric (illumination, 3D camera viewpoint, background clutter, and occlusion) deformations.

Owing to this fact, the features extracted from an image can be grouped into global and local levels [105]. The aforementioned features extracted from intensity histogram, GLCM, and wavelet transforms are considered as global features because they characterise the entire content in an image [98, 105]. The sensitivity of these features to imaging conditions has led to the introduction of local features. These features are well localised in both the spatial and frequency domains which reduce the probability of disruption by imaging conditions [104, 106].

Unlike global features, their local counterparts refer to a pattern or distinct structure found in an image, such as a point or edge. They are usually associated with an image patch that differs from its immediate surroundings by texture, colour, or intensity. Here, what the feature actually represents does not matter, just that it is distinct from its surroundings along various orientations and scales [105].

Local features are selected as a number of ‘interest points’ at distinctive locations (more generally, regions) in an image by a detector that relies on gradient-based and intensity variation approaches. Alternatively, the interest points can be obtained by using a regular grid where the gridline intercepts define locations for interest points. Good local features are superior in three aspects. First, when given two images of the same scene, most features that a detector finds in both images are (ideally) the same (repeatable detection). In other words, the features are robust to changes in viewing conditions and noise. Second, the neighbourhood around a feature centre varies enough to allow for reliable comparison between the features. Third, a feature has a unique location assigned to it. Hence, changes in viewing conditions do not affect its location [99, 101, 104].

From the intensity pattern within a region, a vector of region descriptor is calculated. Here, a local pixel neighbourhood is transformed into a compact vector representation which is robust to local shape distortions and change in illumination within the Images [106].

After extracting vectors of region descriptors, it is required to encode them into a single vector of local features. There are several approaches to fulfil this task. A popular method is Bag Of Visual Words (BOVW) which was initiated originally in the context of texture classification. BOVW is analogues to the frequency of the words used in a text [107]. Words are the constituent elements of a text, and their frequency can be used to represent a text. This property can be used to match two texts and establish word clouds which highlight the most frequently used word/s in a text. By looking at the word cloud of a text, texts can be classified according to their topic, etc. [108].

Unlike a text in which the words come from a known vocabulary, there exist no vocabulary for images to represent them. Therefore, it is required to come up with a vocabulary for them. A BOVW model reduces a large number of vectors of region descriptors by quantising them using cluster analysis. Each cluster centre represents a feature or visual word. From the frequency histogram of the visual words in an image, the vector of local features is extracted. The number of clusters (words) required in a (classification) problem depends on the situation. Therefore, local features can derive a large number of features relative to their global counterparts.

Since the 1980s, a wide variety of local feature detectors and descriptors have been proposed in many image-based domain-specific applications such as content-based image retrieval [109, 110], face recognition, medical image annotation [111-113], scene classification [87, 98, 105, 107, 114-116], and object recognition [106, 116]. They vary mostly by the amount of invariance they theoretically ensure, the image property they exploit to achieve invariance and the type of image structures they are designed to detect [98].

Although extending local features to be invariant to full affine transformations sounds desirable, at first glance, it may result in an increased level of sensitivity to noise. Lowe [106] maintained affine frames are more sensitive to noise than those of the scale-invariant local features which in turn deteriorates repeatability of affine local features. evaluated the performance of several different local feature detectors and descriptors for texture and object classification on four texture and five object databases. The outcome of their research confirms that local features with the highest possible level of invariance do not yield the best performance. Therefore, a combination of multiple detectors and descriptors usually may achieve better results than even the most discriminative individual detector/descriptor channel. This fact can be extended to not ruling out global features and using them in combination with their local counterparts which have observed in some studies [112, 117].

Reviewing the literature in other texture analysis applications revealed that Scale Invariant Feature Transform (SIFT) has been utilised as the preferred method for extracting local features [97-99, 101, 104, 106-108, 110-112, 114, 116, 118-123]. Among the reasons for SIFT to be popular in texture classification problems are robustness to reasonable levels of 2-D affine transformations, scale, and viewpoint changes [103, 119]. More recently, though, due to the wider utilisation of local features in new image-based applications, different versions of SIFT or joint utilisation of different techniques based on the problem at hand are being proposed [97, 109, 112, 116, 117].

In choosing a feature detector and descriptor, the decision must be weighed up against the specific application (e.g. image classification, recognition, retrieval, and so on) and the nature of the images (resolution and presence of clutter or occlusion). There are studies which recommend a specific method over the others in a particular application. For instance, in image-based dietary assessment, Anthimopoulos et al. [103] maintained that SIFT detectors are not generally suitable for food classification problems. In large scale image retrieval, pointed out two limitations of SIFT local features.

SIFT was introduced in 2004 by Lowe [106], and it can be used as a detector or a descriptor. A SIFT detector searches over all scales and image locations by using a difference-of-Gaussian function to locate potential interest points that are invariant to orientation and scale. The nominated points are local maxima in scale space of Laplacian of Gaussian (LOG) filtered image. The scale space is generated by using different smoothed versions of an image filtered by this LP spatial filter.

When it comes to using a spatial LP filter, it is usually unknown at what scale the details may appear in an image. Therefore, using a whole spectrum of scales, the zerocrossings versus scales are plotted in a scale-space. Zerocrossings are the locations of edges in a LOG filtered image. They are associated with the locations at which the second derivative along a row/column becomes zero which indicates a maximum/minimum for the first derivative (intensity gradient) along that direction. Figure 2-6 displays the intensity variation along a particular row (red line) in the original and filtered image. In the intensity graphs, the red plot represents the gradients in the original image while the green plot belongs to the smoothed image which approximates the original gradient.

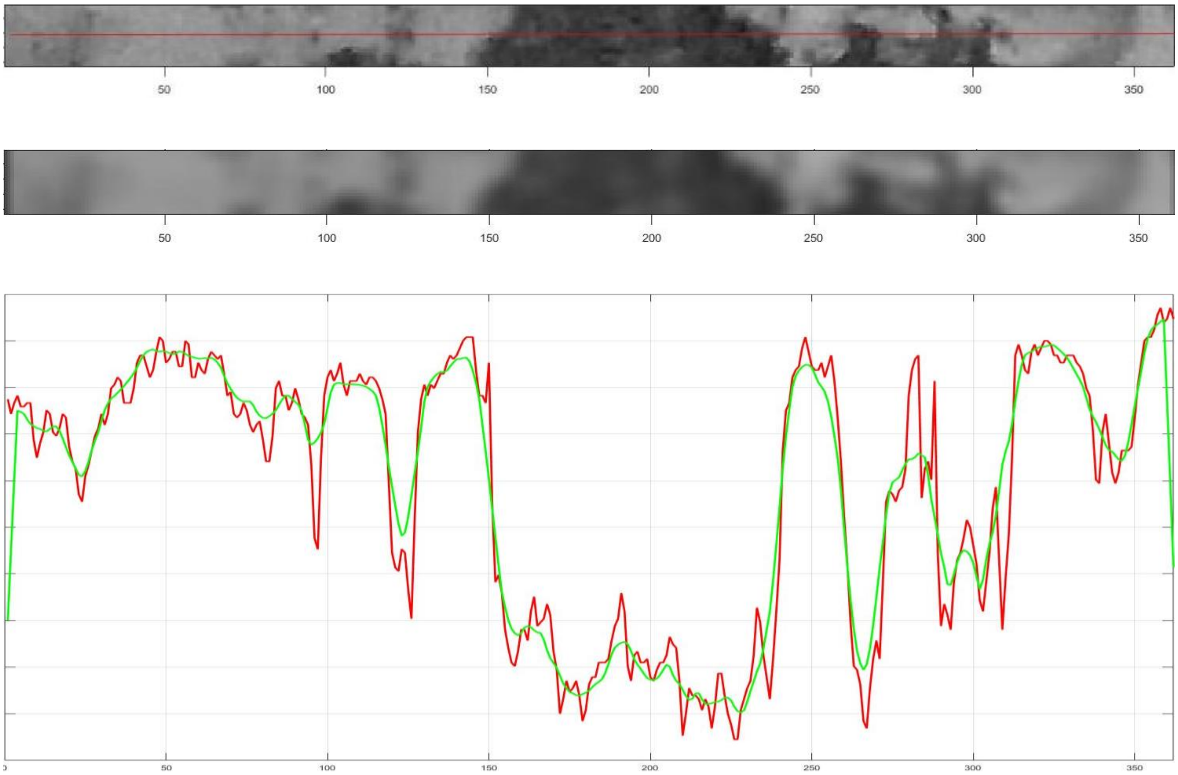


Figure 2-6. The influence of smoothing out operation on intensity gradients

This process can be iterated at higher scales to further smooth out the gradient graph (signal). Through this process, some zerocrossings persist while some are eliminated, and as zerocrossings can be associated with location of interest points, a number of potential points can be ultimately selected as the interest points. Here, the eliminated zerocrossings are associated with noise in the image.

Figure 2-7 displays 15,235 potential interest points that were originally identified from the zerocrossings. By resampling the original image and applying three different sigma values at each octave (sample level), 114 points are identified as the interest points. Each resampling is carried out by eliminating every other row and column which results in 50% reduction in the resolution. Each octave (row in the image) was filtered by three different sigma values which increases from left to right.

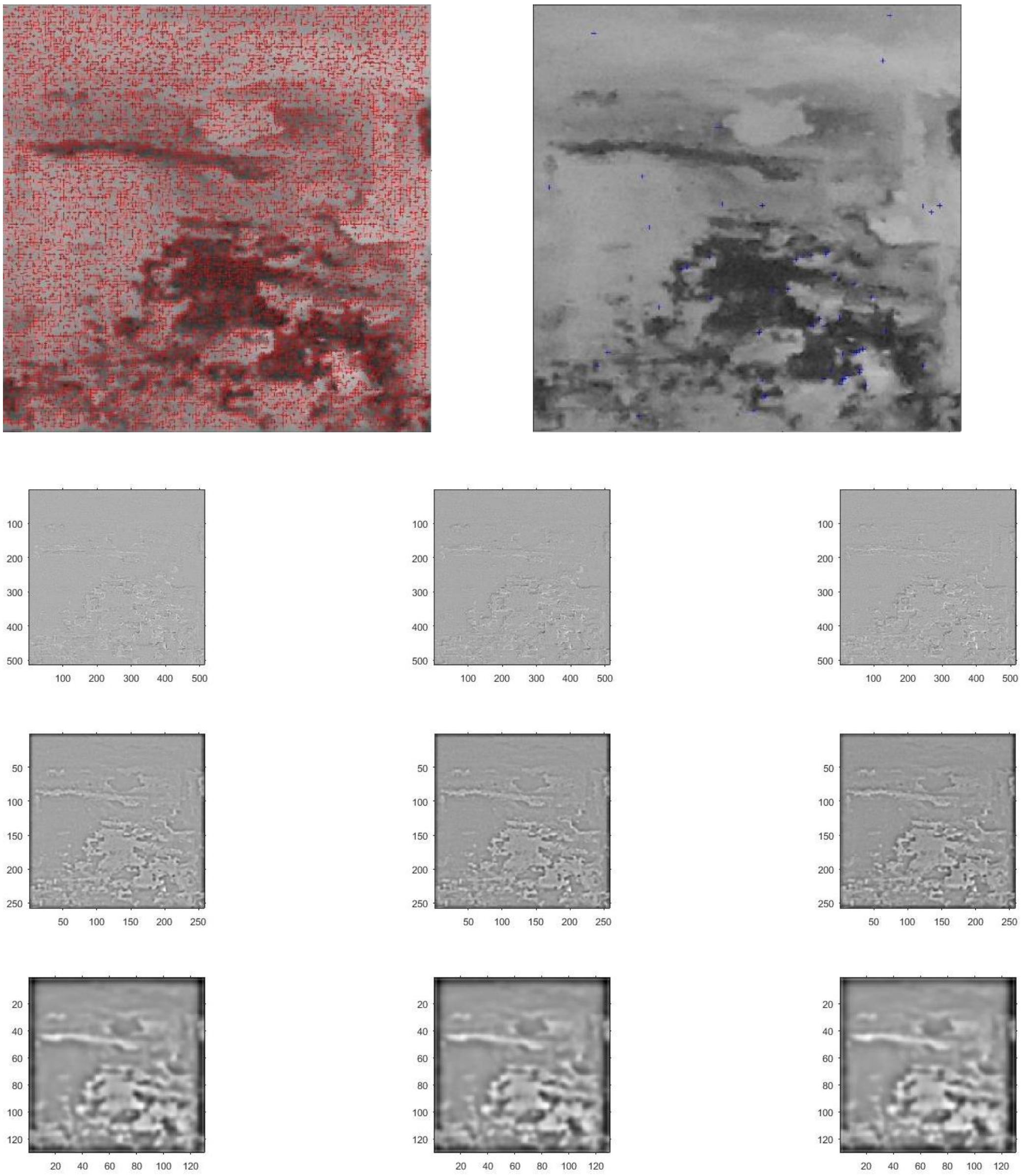


Figure 2-7. Calculation of SIFT interest points via down-sampling

A SIFT descriptor is used to derive a vector of descriptors at a local region (usually a 16×16 pixels neighbourhood) surrounding each interest point. The magnitude of the local gradients and the gradient orientations at each pixel within the surrounding patch is computed. Gradient looks at change, and it is more robust to noise than the actual intensity values. When there is little change in illumination, the intensity values may significantly change while the gradients may not be that much influenced. Next, the 16×16 neighbourhood is divided into a 4×4 sub-region and a histogram with a certain number of bins summarises the frequency of gradient orientations in each sub-region. For instance, a histogram with eight bins displays the frequency of orientations at 45° intervals. Figure 2-8 displays this procedure for the aforementioned neighbourhood and histogram. The gradient magnitude and orientation at each point are weighted by a Gaussian window (the overlaid circle) at the left image. They are then accumulated into orientation histograms at the middle image. The length of each arrow corresponds to the sum of the gradient magnitudes near that direction within the region.

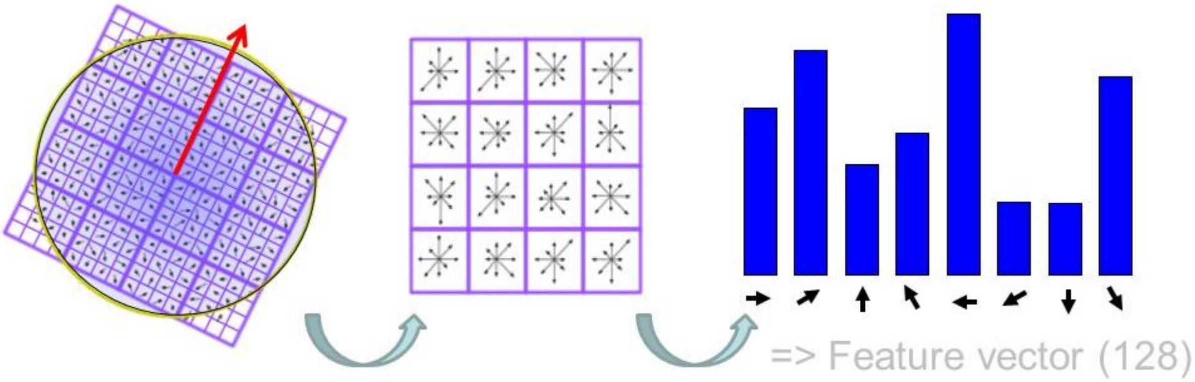


Figure 2-8. Calculation of frequency of gradient orientations

From a 16×16 neighbourhood, a 4×4 descriptor array is computed in which each element (histogram) is an eight-dimensional vector. Therefore, a SIFT descriptor is a 128 dimensional vector [97, 106, 111, 124, 125]. Lowe [106] justified the selected size for the neighbourhoods, sub-regions, and histogram bins by comparing the percent of interest points giving a correct match to a database of 40,000 interest points using different sizes.

In 2008, a version of SIFT knowns as Speeded-Up Robust Features (SURF) was introduced to address the concerns associated with the large dimension of region descriptors and consequently computation costs associated with SIFT descriptors [99]. SURF features are designed for classification problems. Figure 2-9displays the strongest 100 detected SURF interest points along with their corresponding neighbourhoods in two different images of the same object (Eiffel tower) that had been captured under widely different photometric conditions. In the left and right figures, 746 and 572 interest points were identified respectively. As can be seen, the radii of the neighbourhoods are different. Their values are calculated based on how large a region around an interest point needs to be extracted to form a feature vector.

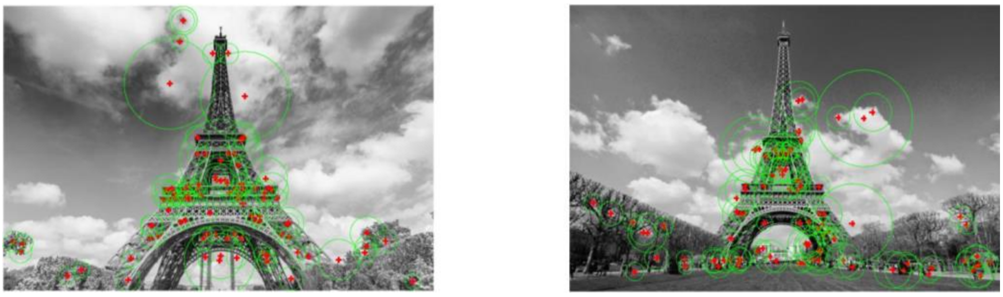


Figure 2-9. SURF interest points and their patch neighbourhoods

These are very distinct regions in the image which can be used to match with other images of the same class or matching an image with another image of the same scene or object. Figure 2-10 displays the eight matching interest points on these two images that are connected by the yellow lines.

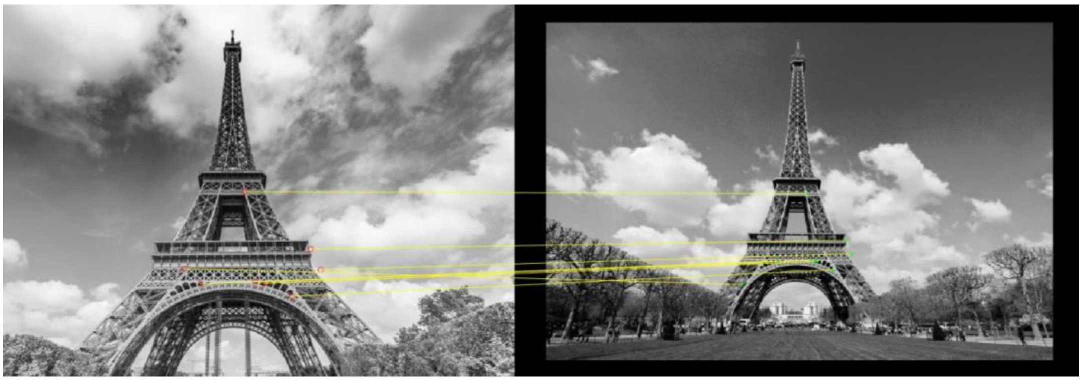


Figure 2-10. SURF can be used for the purpose of object detection

Local features have application in many domains. For instance, Content-Based Image Retrieval (CBIR) takes advantage of them to search within an image database to identify a suspect whose image was taken by a CCTV camera in a crime scene. In remote sensing, local features can be used to match interest points which belong to urban areas in several images which have been taken over a time span to detect measure any change. In text recognition, local features can be used to read alphanumerical characters in images of a scanned book or read the license plates of a car driving over the speed limit.

Besides SIFT and SURF, the Computer Vision Toolbox in MATLAB offers several types of local feature detectors such as the Features from Accelerated Segment Test (FAST), Harris, Shi & Tomasi, and Maximally Stable Extremal Regions (MSER) methods. It also provides local feature descriptors such as Fast Retina Keypoint (FREAK), Binary Robust Invariant Scalable Keypoints (BRISK), and Histogram Of Gradient (HOG). One can mix and match the detectors and the descriptors depending on the requirements of the problem to find the best configuration that yields the highest classification accuracy down the track. There are some general rules of thumb that may recommend utilisation of a specific method over the others based on the type of features in images, the context in which the features are used (i.e. classification and registration), and performance requirements (e.g. real-time performance and accuracy versus speed). Needless to say, the type of distortions present in images may limit the choice of local features as each is robust to one or some particular distortions.

In the context of corrosion, the application of local features has been limited to only a few studies that used them for matching purposes. used SIFT features to identify bolts and the cracks in their vicinity as a part of an automated vision-based crack detection for automated inspection of large scale bridge structures. This algorithm can be applied to aerial images without advanced knowledge of the crack locations or special control of camera position and orientation. An automated method for monitoring the evolution of corrosion at industrial structures was introduced by . They used SIFT features to match UAV images of the same area to unify their alignments before comparing their corrosion severity.

To learn more the application of local features to classification problems, other spheres of texture analysis literature was reviewed. introduced an automatic image modality classification method for three medical image databases with 18 and 31 distinct classes. The texture was quantified using local binary patterns, colour and edge directivity descriptors, fuzzy colour and texture histogram, and SIFT. They observed SIFT detectors provide the best performing descriptors for modality classification. SIFT method was used by another study for a five-level liver fibrosis scoring. In the visual codebook, the influence of grid spacing, bin size, and the codebook size was investigated across 168 possible combinations of them. Detecting lesion in retinal images that are induced by diabetic Retinopathy was the subject of another study in which SURF features were deployed. Three large retinograph datasets with diverse resolutions were used to test the performance of their algorithm [113]. A SIFT detector was also used in a food recognition system to classify 11 classes of food in 4868 images that had been collected from the web [103].

To summarise, local features are invariant to many image deformations and also can provide large quantities of numerical features in comparison to global features. Having more characterising features can potentially enhance the performance of subsequent machine learning algorithms.

Hence, introducing these methods in the analysis of corrosion damage at taper junctions of retrieved hip replacement implants may address the reliability concerns associated with the subjectivity nature of visual scoring methods.

2.5.2 Pre-processing the Features

Describing morphological attributes of corrosion damage is a challenging task. There exists no integrated theory to offer a set of specific features that provide a desirable level of discriminatory power to classify the intensity or typology of corrosion damage. Each study, according to the acquired images, the situation at hand (e.g. type of the alloy and defect), and the available computational technology has used a specific set of features for their problem. Section 2.5.1 categorised these features into four major groups (Figure 2-11). Features that characterise the shape of objects of interest were excluded from this study as explained in that section.

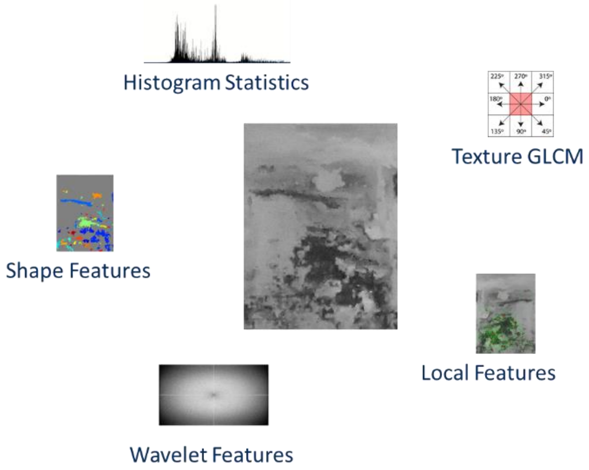


Figure 2-11. The raw information from an image can be captured and characterised by features

Generally, a larger feature set demands for more computer memory and leads to higher computation time, while fewer features can produce a poor classifier. Therefore, it is essential to preprocess the extracted features to ensure the most relevant information are passed to the subsequent machine learning algorithms. This section discusses some methods to assess the quality of the extracted features according to their discriminatory power in classification problems.

Some of these methods can determine the contribution of each feature and identify those with negligible levels of power that can be discarded. Reducing the dimensions of the feature space by dropping less descriptive features simplifies the classification process which may come at a computationally expensive cost. Alternatively, the feature can be combined to obtain a transformed version which produces a simpler description of the system in terms of capturing as much of the variability in the data. These two approaches (i.e. feature selection and feature transformation) are for the purpose of dimensionality reduction, and each can be implemented in several different ways. It is the context of the problem which determines which one is applicable.

Feature transformation methods such as Principal Component Analysis (PCA) aims to capture and reconstruct the variability of the original features in its most succinct and compact way. PCA produces a transformed version of the original features (principal components) which are a linear combination of them. PCA computes and sorts the PCs in the order of their contribution in explaining the variability of the original data points. This property of PCA helps with identifying PCs with no or negligible contribution and removing them to reduce the dimensionality of the feature space. It should be noted that PCA can act as a feature selection tool as well. It can find PCs that have high correlations with one smaller set of features and little or no correlation with another set of features. Therefore, those features with no higher correlation with any PC are discarded [61].

Capturing the variability of the features is of interest when it is desired to identify the number of clusters that are naturally formed by the data. This fact is not applicable in this study since the number of classes is already known, and there are sensible visual criteria to distinguish them. Also, feature transformation is applicable when the number of observations (e.g. images) is significantly greater than the number of descriptors (e.g. features). However, as mentioned before, DIP may produce hundreds of numerical features that exceed the quantity of the available images. Therefore, feature transformation was not deemed applicable in this study.

Feature selection aims to identify a subset of the original features that are most suitable for prediction of the known classes. Therefore, it is applicable in classification problems with pre-labelled data which is the case in this study. Unlike feature transformation techniques such as PCA and multidimensional scaling which were used by several studies [60, 61, 63, 64, 82, 83], feature selection has been deployed by fewer studies in the context of image-based corrosion assessment. Among the reasons for this fact is the low number (usually below 10) of features that are usually extracted in such studies. Gamarra Acosta et al. [68] used Fished scores that are a metric of interclass distance to reduce the dimensionality of their feature space from 6 to 1. In another study, used two filter model methods (Wilk's Lambda and data range analysis) to reduce the dimension of the feature space to three in a corrosion detection problem. Considering feature selection as a dimensionality reduction tool, it may not be suitable for feature spaces with small dimensions due to losing discriminative information required by the classification algorithms.

One method for feature selection that can handle large dimensional data sets as well as large feature spaces is Neighbourhood Component Analysis (NCA). NCA is a model that is embedded into a randomised classification construction and calculates a vector of feature weights which indicates the relevance between the corresponding features and the target labels. This feature selection method is less prone to overfitting concerns and computationally more efficient in comparison with the filter model (e.g. Fisher Score) and wrapper model feature selection algorithms [126]. NCA has a regularisation parameter, Lambda, which can be tuned to further reduce the generalisation error. It provides control over the sparsity as well as minimising the redundancy within the feature space.

2.5.3 Machine Learning Algorithms

After extracting and selecting numerical descriptors of corrosion, each image can be replaced by a representative feature vector. Using these vectors, an image is mapped into an abstract feature space as a point where its coordinates are derived from the elements of its corresponding feature vector. In the context of classification, machine learning algorithms search for decision boundaries or clusters that partition the images (points) according to their actual class (supervised) or a sensible reason (unsupervised) in the feature space. There exists a broad spectrum of (un)supervised techniques that serve this goal.

2.5.3.1 *Unsupervised Learning*

Unsupervised learning techniques aim to find how many clusters are naturally formed by the feature vectors in the feature space. The most natural clustering method is to visualise the feature space and subsequently, the cluster locations. Since the dimensions of the feature vectors are usually greater than three, for a human observer, direct visualisation of the feature space is impossible. Therefore, alternative methods have been developed that quantitatively measure the similarity of the images in the feature space.

A similarity measure describes how near or far the images are mapped in the feature space. The closer the distance between two distinct points on the space, the more their similarity. On the same token, the farther the distances are, the less alike they are in terms of their features. Based on the nature of the data, one can choose from several different pairwise distance metrics such as Euclidean, Cityblock, Minkowski, and Mahalanobis.

In accordance with the measured distances, clusters are formed on the feature space. Additionally, cluster evaluation determines the optimal number of clusters for the data via using different evaluation criteria. The task of unsupervised classification is to determine the ranges of the clusters corresponding to various images and to set the rules being the functions of their features, which are used to divide the images into classes [61].

Studies such as [58, 60-62, 64, 78, 83] used cluster analysis techniques such as multidimensional scaling and multivariate discriminant analysis in the classification of corrosion severity and typology. For instance, multidimensional scaling was used by two studies in a classification of corrosion typology to visualise pairwise distances of images in the feature space by producing a representation of them in a smaller number of dimensions. They maintained that Minkowski method is the most frequently used similarity metric in classification problems [61, 83]. Unsupervised learning techniques do not account for the actual class of the feature vectors and also the number of desired score levels. Therefore, they do not apply to the classification problem in this work. If unsupervised learning was not deemed as a suitable classification tool, then one would have to resort to supervised learning.

2.5.3.2 *Supervised Learning*

Supervised learning is carried out according to the true classes of a training set. These adaptive algorithms identify patterns in data by learning from the observations. Generally, classification can be considered as a supervised learning task when the classes have meaningful definitions. Classification by Machine Learning divides the images into test and training subsets. By using the training images, their corresponding feature vectors are extracted and mapped into a feature space. Then, the algorithm finds line/s (2-D feature vectors), plane/s (3-D feature vectors), or hyperplane/s (higher dimensional feature vectors) that best classify the images according to their corresponding visual scores. These borders are also called decision boundaries. Figure 2-12 displays this process for a 2-D feature space schematically. In this example, there are two distinct classes (denoted by square and triangle), and 11 images for each class. Therefore, there are 22 feature vectors where each vector maps a point into the feature space. Using the true class of each image that was already determined by an expert, the algorithm searches for a line (decision boundary) so that the sum of the vertical distances of the 22 points is minimised.

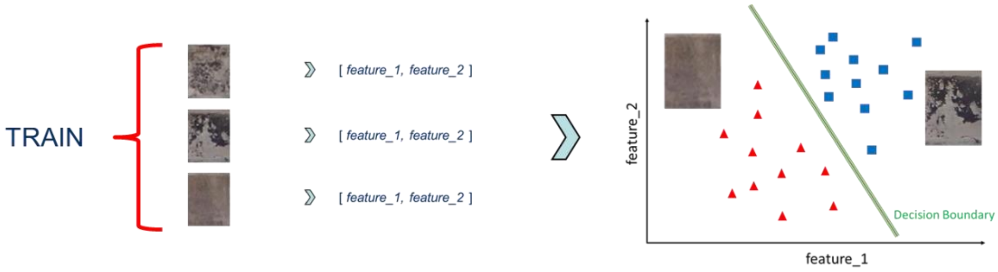


Figure 2-12 The training phase yields the decision boundaries.

To evaluate the performance of the predictive model, the images from the test set are used to see how the model can be applied to unseen images. Similar to the training stage, the feature vector for each image is calculated first. Next, this vector is mapped into the feature space and based on its location with respect to the decision boundaries, the class (score) is predicted (Figure 2-13).

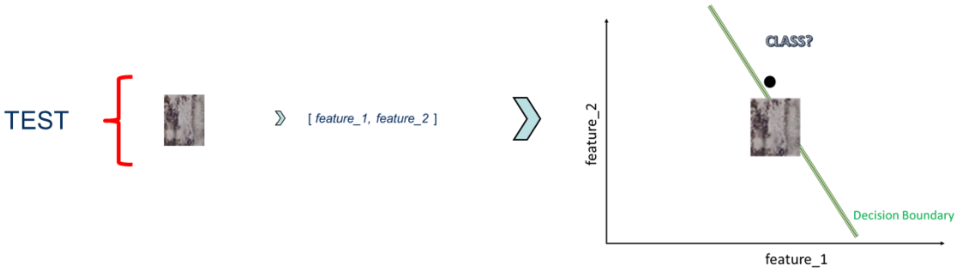


Figure 2-13 Decision boundaries determine the class of an unseen image.

The actual class of each image is used as the benchmark to determine what percentage of the test subset images are misclassified by the predictive model. Obviously, it is desired to utilise a machine learning algorithm which minimises the classification error.

Over time, several machine learning methods have been introduced. These methods differ in terms of their speed of training, memory usage, and predictive accuracy of new data. To find which method/s may be more suitable for a particular problem, the literature can be studied to find the recommended methods where similar types of data and objectives have been used.

In the context of corrosion-related applications, an emerging machine learning method is Support Vector Machine (SVM), a binary classification algorithm that has been mostly used for detection of corroded from intact surfaces [69]. It was used though for a multiclass problem by to distinguish pitting, uniform corrosion, and passivation of 304 stainless steel based on features that were extracted from electrochemical noise and achieved a 100% accuracy. In a binary classification problem, SVM was used to objectively decide on retiring or reusing the cross-arms of power poles. Interestingly, they did not deploy any DIP algorithm to extract numerical features and simply used the actual pixel values as the input for the machine learning algorithm [127]. 165 images of steel bridges were analysed in another study [59] to determine the rusted areas that needed blasting. They maintained that Radial Basis Function (RBF) is the most commonly used kernel function in SVM algorithms. Yan et al. [69] used SVM to classify the severity of corrosion in 530 weathering steel images. They also maintained that RBF kernel functions outperform their counterparts such as linear, polynomial, and sigmoid function in that they need fewer model parameters and can achieve easier convergence on numerical training. In their work, a maximum classification rate of 85% was achieved.

In addition to analysing corrosion, this method has shown promising results in other computer vision and text categorization applications, especially those involving large dimensional feature spaces [97, 98, 121, 128, 129]. For further information about the other types of machine learning algorithms, one can refer to two survey studies that list and compare them in the context of corrosion-related applications [63, 130].

SVM classifies images by finding the best hyperplane that separates all data points of one class from those of the other class. Hence, SVM is only suitable for binary classification problems, and it cannot handle multi-class problems directly. The best hyperplane for an SVM is the one with the largest margin between any two classes. Margin is the maximal width of the slab parallel to the hyperplane that has no interior data points. The support vectors are the data points that are closest to the separating hyperplane and constitute the boundary of the slab [95]. Figure 2-14 shows the support vectors and the margin for a 2-D feature space. As mentioned before, the decision boundary appears as a line in a 2-D feature space.

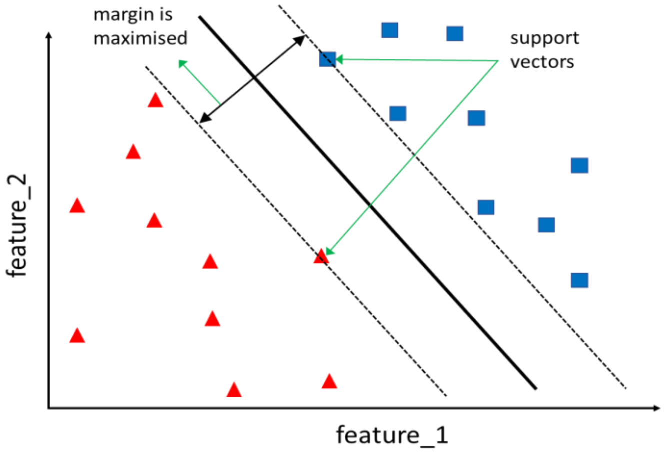


Figure 2-14. Classification by binary SVM learners.

In order to deploy SVM in a machine learning algorithm, there exist other considerations that will be discussed in the next chapter which elaborates on implementing SVM.

2.6 Statistical Analysis Methods

Upon quantification of damage semi-quantitatively, each study deploys a particular set of statistical analyses to investigate the effect of a particular set of factors (predictors) on the response (damage score). This step is quite critical since it extracts knowledge out of the undertaken cumbersome observations and measurements, and constitutes the outcome of these investigations. Although the nature of many retrieval studies is similar in terms of the existing predictors and response variables, a diverse range of statistical methods is used by the literature to study their associations.

Generally, these analyses belong to three groups of methods, (1) correlations (e.g. Pearson product-moment and Spearman's rank-order), (2) comparing average (median) values between groups (e.g. student t-test, ANOVA, Mann-Whitney, and Kruskal-Wallis test), and (3) prediction of outcome variable (e.g. multiple linear, binomial, and ordinal regression). Reviewing the literature of large-scale retrieval studies reveals that while the first two categories have been the popular methods [37, 39, 131-133], the last group, more specifically Ordinal Logistic Regression (OLR), has been seldom used by retrieval studies [46, 56, 134].

2.6.1 Correlations

This group only works for univariate analyses. While Spearman's rank-order has been used frequently by many studies [6, 24, 31, 36, 135], Pearson product-moment has been used less often [136, 137].

Spearman's rank-order provides a measure of the strength and direction of the association between two continuous or ordinal variables. The ordinal nature of visual scores justifies the higher popularity of this method compared with its counterpart.

Pearson product-moment measures the strength and direction of a linear relationship between two continuous variables. For instance, multicollinearity of continuous patient/implant factors as well as associations between these factors and material loss have been quantified via this method.

2.6.2 Comparing Differences Between Groups

This group can be deployed to conduct univariate as well as multivariate analyses (e.g. student t-test, ANOVA, Mann-Whitney, and Kruskal-Wallis test). However, the number of included factors in multivariate analyses may not go higher than two or three which makes this method inefficient for conducting multivariate analyses of several patient/implant factors. Since this group of analyses compares the mean (median) between groups of categorical factors, only categorical factors can be analysed via this group, not continuous factors. The strict assumptions of this group of methods demand a reasonable quantity of observations in each cell of the design to check for outliers, the distribution or variance of damage scores.

2.6.3 Ordinal Logistic Regression

OLR can also handle both univariate and multivariate analyses (e.g. multiple linear, binomial, and ordinal regression). Since the assumptions of OLR do not mandate a particular distribution of data or similar variance in each cell of design, more factors can be included in multivariate regression models compared with the previous group of methods. Unlike the previous two groups, OLR models can take in both numerical as well as categorical factors. More importantly, besides the determination of the influence of several factors on an ordinal variable, OLR provides predictive models that can be used to predict damage scores based on the available patient/implant data. This attribute of OLR sets it different from the other two groups since the outcome would not be retrospective analytics any more, but predictive analytics.

OLR belongs to the larger group of machine learning techniques that achieve a model using the training data when equation and laws are not promising. Besides OLR that fits a linear model to predict visual scores, there exist several other machine learning techniques that address situations where more complex nonlinear models may provide higher classification (scoring) accuracy.

2.7 Summary of the Literature and Highlights of the Research Gap

This chapter elaborated on several large-scale retrieval studies of hip replacement implants that had investigated many potential associations between patient/implant properties and fretting or corrosion damage at various junctions of these prostheses. Reviewing these studies revealed several issues that may pose concerns over the reliability of the study outcomes.

In the next chapter, the reviewed methodologies are tailored to an intelligent algorithm that utilizes DIP and machine learning to objectively score corrosion damage in a large-scale retrieval study. Also, OLR, as well as other machine learning techniques, will be developed to generate both retrospective and prospective analytics based on the available patient/implant data.

2.7.1 Comparison of Corrosion at the Zones

There are several studies that scored stem tapers locally according to a set of predefined zones [31, 44, 53, 55, 131, 132, 138-141]. The literature usually assumes the overall value as the global score for each implant [31, 55, 131, 132, 138]. This has led to presuming that this global score is a continuous variable, and statistical analyses for continuous variables have been utilised. Analysing a continuous variable with an interval or ratio level of measurement is generally less complex in nature. However, an increased number of levels in the global score does not imply a known "distance" between the score levels. Therefore, it is required to search for more reliable alternatives. Therefore, this approach was treated with suspicion in this study and was not adopted.

An alternative approach can be for zone to go through only univariate analysis based on the local scores, and the other predictors to go through univariate or multivariate analyses by holistically scoring the damage. Based on the advantages of OLR over the other statistical methods, it can fulfil both these tasks.

The number of zones scored at stem tapers has seldom gone beyond four (anterior, medial, posterior, and lateral quadrants). One reason for that could be the complexity of conducting pairwise comparisons within the groups of zone factor. With four zones, six combinations (order disregarded) would be required. If it is desired to consider the distal and proximal regions of each quadrant as well, 28 (i.e. $\frac{8!}{(8-2)!\times 2!}$) pairwise comparisons would be required to investigate the damage thoroughly. The studies that scored the distal and proximal regions separately have observed different damage patterns within these regions [44, 140, 141]. Therefore, it is required to look at stem taper zones with a higher level of granularity to explore whether any significant difference exists between the distal and proximal regions of the quadrants.

2.7.2 Visual Scoring Reliability

Visual scoring of fretting has been associated with several issues by the literature. According to several studies, fretting might be masked by corrosion damage and, consequently, hard to visually identify [31, 44]. Also, Hothi et al. [31] maintained that severity of fretting in Goldberg's method could not be measured consistently because the pitch of the machined thread of trunnion varies among different stem designs. Besides, fretting scars can be mixed up with scratches caused by attaching or detaching the head intraoperatively [132].

It was discovered that the damage severity is measured by only semi-quantitative methods of visual scoring in large-scale retrieval studies, whereas having a numerical (rather than an ordinal) measure may significantly improve the damage resolution. Numerical metrics are superior in capturing the variability of damage which in turn results in more accurate statistical inference outcomes. To measure material loss, complex instruments which perform surface profilometry are required, wherein time and cost feasibility limitations do not allow employing that for large sample sizes.

Due to these drawbacks, quantification of corrosion damage in large sample sizes was investigated in other domains of corrosion-related applications. It was discovered that a new group of methods had been developed over the past decade that is based on analysis of images or signals that are obtained from corroded samples. They capitalise on the principles of image/signal processing and machine learning to provide categorical or numerical measures of damage that might have been induced naturally or in a lab environment. More specifically, it was pointed out that they have been used to analyse the shape and location of defects, rate corrosion severity, or identify corrosion typology. While this new group has shown promising results in various corrosion-related application, no retrieval study has considered deploying them in the analysis of wear or corrosion at the hip or other orthopaedic implants.

The studies that have used this group of methods deployed various texture and shape analysis techniques which have been in use by areas other than corrosion for similar purposes (e.g. classification or prediction). Therefore, having a reasonable level of insight into the texture analysis methods that have been developed in other texture related applications may facilitate developing simple yet effective algorithms.

Rating the corrosion severity at hip replacement implants can be more challenging in comparison with many similar studies that have addressed corrosion as rust [67, 68, 70]. The morphological attributes that can be used in characterising corrosion at taper junctions confine to just texture while both colour and texture play a prominent role in characterising rust.

Also, there are additional complexities in the nature of this classification problem compared to those algorithms that work based on a set of well-defined classes (e.g. object classification). In retrieval studies, visual scores need to be benchmarked to train and evaluate the performance of machine learning algorithms. These scores are prone to the existing reliability concerns which may impact the quality of classifiers. It should be noted though that machine learning aims to take over the role of an expert. Therefore, to address this matter, it is essential for the images to be scored by an experienced investigator in the training phase so the resulting classifier can provide the same expertise which then can be used by anyone and anywhere in future. Looking at classification problems in popular contexts such as medical image analysis shows a similar situation in which objects (regions) of interest have to be labelled in training images by an expert to develop intelligent classification tools.

The application of local features in corrosion images are limited to a few studies that only used them for image registration purposes. However, it was observed that areas other than corrosion had used them successfully as texture descriptors. Hence, it needs to be investigated whether popular local feature detectors and descriptors (e.g. SIFT or SURF) can be used in characterising texture in corrosion images.

2.7.3 Multivariate Analysis

Considering the large number of factors that have been identified so far as potential contributors toward damage, and the synergistic role of these factors in the outcome of hip replacement operations, it is essential to study the role of these factors via multivariate analyses.

Although the nature of many retrieval studies is similar in terms of the existing predictors and response variables, a diverse range of statistical methods is used by the literature to study their associations. Review of the literature revealed that the majority of these methods face limitations to perform multivariate analysis. These limitations mostly arise from assumptions that demand specific distribution of scores and equal variance in each cell of design. However, the assumptions of OLR does not require maintaining such conditions. Also, unlike the other methods which can handle only predictors with specific levels of measurement, OLR can handle continuous, dichotomous, and polytomous predictors simultaneously.

OLR is a method that has been seldom used by the literature. The studies that used this method have not analysed polytomous variables. Analysing this type of variables such as zone and stem design requires several pairwise comparisons to obtain an overall test of statistical significance. These comparisons do not directly sort the groups in terms of the severity of the damage. Therefore, it is required to devise a method to fulfil this task.

One drawback of using regression models for multivariate analysis is that all the factors may not be included in a model due to the possibility to violate the multicollinearity and proportional odds assumptions. Another issue which exists for all statistical analyses is missing information which is an inherent characteristic of healthcare data. Increasing the number of included factors will result in a drop in the number of eligible records following listwise deletion. These two reasons turn the variable selection into a cumbersome task. Regression models with different sets of factors yield different outcomes. For instance, a particular factor grouped with two distinct sets may turn from significant into insignificant or vice versa. The significance of this matter needs to be investigated.

The retrieval studies that deployed OLR have only used it to produce retrospective analytics to investigate the influence of the factors that hypothesised to be contributing to damage scores. Hence, the quality of the regression models in the prediction of damage has been overlooked and not reported so far. Considering OLR as a baseline machine learning technique for more complex nonlinear models, a comparative investigation of linear (i.e. OLR) versus its nonlinear counterparts is required to identify suitable predictive analytics techniques for the context of retrieval studies.

2.8 Research Aim and Objectives

This section specifies the aim and objectives of this research study, following the review of three spheres of literature that were comprised of large-scale retrieval studies of hip replacement implant, DIP and machine learning algorithms in analyzing corrosion, and DIP and machine learning in texture analysis within application other than corrosion. This thesis is to address the reliability concerns associated with the visual scoring of corrosion by developing an intelligent method.

2.8.1 Aim

- Incorporating Artificial Intelligence (AI) to objectively determine corrosion scores based on the medical records data as well as features extracted from images of stem trunnions

2.8.2 Objectives

- to search and identify a set of explants in the retrieved implant library of Royal Adelaide Hospital that meets the inclusion criteria for this study
- to develop a photography setup that enables capturing images of stem trunnions under desirable lighting-optics-viewing conditions

- to determine a set of morphological attributes within images of corroded regions that provide superior discriminatory power for this classification problem
- to train, optimize, and cross-validate a machine learning algorithm that learns statistical patterns within images belonging to the same Goldberg score levels with an acceptable accuracy rate and computation cost
- to compare the severity of corrosion damage at eight distinct zones of stem tapers
- to investigate the variability of regression model outcomes when different sets of factors are selected
- compare the accuracy of several machine learning methods in prediction of the corrosion scores based on implant records data

3 METHODOLOGY

[Image removed due to copyright restriction]

3.1 Introduction

The core of this chapter is about deploying AI to objectively predict the sustained corrosion damage and find possible associations between that and several patient/implant attributes. Prediction of corrosion scores can be achieved via two different alternatives. The first approach utilises images of stem tapers to rate the severity of damage according to a desirable scoring method. The second approach is based on using the patient and implant data to achieve a predictive model.

Designing an intelligent algorithm that can be trained according to a desired visual scoring method and automatically score images of stem tapers of hip replacement implants was inspired by the literature of texture analysis in corrosion and some other applications that offers guidelines in choosing and implementing accurate yet efficient methods. When it comes to choosing the programming environment, one can choose software packages that provide libraries and toolboxes of common relevant workflows. In this study, MATLAB was selected as the programming tool. The Image Processing toolbox along with Statistics and Machine Learning toolbox offer several functions and interactive applications that can be used to perform tasks such as feature extraction and multidimensional data analysis. Rather than programming in a lower-level language (e.g. Java, C, C++, or Fortran), where it is required to perform tasks such as declaring variables, specifying data types, and allocating memory, MATLAB enables one to focus on the actual problem and apply concepts in a wide range of engineering, science, and mathematics applications. It offers thousands of engineering and mathematical functions which eliminates the need to code and test them yourself. Once an algorithm was designed and tested, MATLAB coder workflow can generate a corresponding standalone C or C++ code to be used in desktop, mobile or web applications.

3.2 Acquiring the Implants

This study was approved by the Southern Adelaide Clinical Human Research Ethics Committee (Reference No. 485.13). The inclusion criteria associated with the suitable implants for this study were (1) revision operations between 1995 and 2015 (2) detached head and stem (3) available patient record. To identify the right retrievals for this study, 2131 records of the primary and revision operations at Royal Adelaide Hospital (RAH) were explored to find the revision procedures in which both the head and stem had been retrieved. These records were checked against the information in Our Patient Management and Outcomes Database (OPMOD) of RAH to resolve any missing or conflicting information. The outcome was 302 cases with 164 attached head-neck and 138 detached head-neck junctions. The attached cases were excluded from this study. Also, the record for one case with attached head-neck could not be found.

Table 3-1 summarises the demographics and the implant information of this pool. Despite the great efforts to resolve the missing data issue, it could not be entirely addressed.

Table 3-1 Demographics of the selected retrievals for this study

Predictor	Quantity (% frequency)	Median	Range
Head Material			
CoCr	60 (43.8)		
Stainless Steel (SS)	7 (5.1)		
Ceramic	8 (5.8)		
Stem Material			
CoCr	54 (39.4)		
Stainless Steel (SS)	41 (29.9)		
Titanium	31 (22.6)		
Stem Fixation			
Cemented	76 (55.5)		
Cementless	50 (36.5)		
Gender			
Female	57 (45.2)		
Male	69 (54.8)		
Stem Taper			
12/14	52(38.0)		
V40	19 (13.9)		
9/10	12 (8.8)		
6°	8 (5.8)		
C-TAPER	8 (5.8)		
TYPE 1	2 (1.5)		
11/13	3 (2.2)		
10/12	1 (0.7)		
Joint Side			
Right	69 (54.8)		
Left	57 (45.2)		
Head Diameter (mm)		28	22-55
Time to Revision (year)		6	0-35
Weight (kg)		77	51-178
Age at Primary (year)		63.5	22-85

According to the Implant Retrieval, Cleaning and Documentation Protocol developed at the Royal Adelaide Hospital, the retrieved implants were immersed in 70% ethanol for four days; and subsequently, washed with running water. The washed implants were then immersed in 4% Biogram solution (polyphenolic disinfectant and detergent with approximately 18% phenol) and left in a fume cupboard for 48–72 hours. During the decontamination process, biologic debris such as blood or proteinaceous films were carefully removed using a cotton bud without abrasion. For this study, the stem tapers were further cleaned by immersing them in acetone for three minutes followed by a gentle wipe with a soft nylon brush to ensure the surface was free of any biological film or dirt.

3.3 **Photography**

The performance of any DIP algorithm can be sensitive to affine distortions, scene illumination, 3D camera viewpoint, and background clutter. The metallic surface of stem tapers can produce high reflections of light which usually appear as strips of glare that cover the surface partially and deteriorate the quality of an image. Also, these surfaces reflect the colours of nearby objects which is undesirable. Hence, it was required to develop an overall plan for a vision system to minimise the presence of glare and reflections from the nearby objects such as camera and implant fixtures which deteriorates the quality of images. In doing so, the following matters have to be addressed via experimentation.

- Configuring an optical sub-system (i.e. placing the lighting, optical elements and camera)
- Calculating field-of-view, depth of focus, and image resolution
- Choosing an appropriate image sensor (camera)

To create a controlled set of conditions for the photography, imaging of implants was performed in a cubicle. An Olympus TG-4 camera with a resolution of 16 Megapixels was selected because that is the only compact camera that offers the focus stacking feature. This feature allows the camera to take multiple consecutive shots of a scene with several different degrees of depth of field and merges them so that the entire scene is in focus from the foreground to the background. This feature was initially believed to be vital as the taper surfaces are curved, and imaging from a close distance (below 100 mm) results in the focus to be on only a specific portion of the stem tapers. However, a shorter distance between the camera and an implant was found to give rise to the high reflection of the camera's image on the surfaces of the implant.

To cover the entire 360° of the circumferential areas of tapers, four shots (anterior, medial, posterior, and lateral) were taken by imaging the tapers every 90°. Distal ends of the stems were inserted inside a high-density foam in angles and levels for the tapers to be placed in the centre of images with a minimum level of inclination to the sides. Table 3-2 details out the camera settings.

Table 3-2 The camera settings

Feature	Setting
Optical Zoom	x4.0 - MAX
Digital Zoom	-
Mode	Microscope
Flash	Off
Picture Mode	Natural
Exposure Compensation	0
White Balance	Yes
ISO	Auto
Resolution	3456 x 3456 pixels
Aspect Ratio	1:01
Compression	Fine
Auto Focus Mode	Spot
Image Stabiliser	Off

These settings allowed for reduced levels of noise and consistent quality in images. To minimise the reflection of the surrounding objects in the tapers, an implant and its fixation foam were placed inside a cylindrical shade with a diameter of 420 mm. The camera was fixed via a flexible arm in place and placed outside and tangent to the shade. A hole in the wall of the shade with a diameter of 10 mm allowed the camera to capture images, and avoid the camera's reflection in the implants (Figure 3-1).

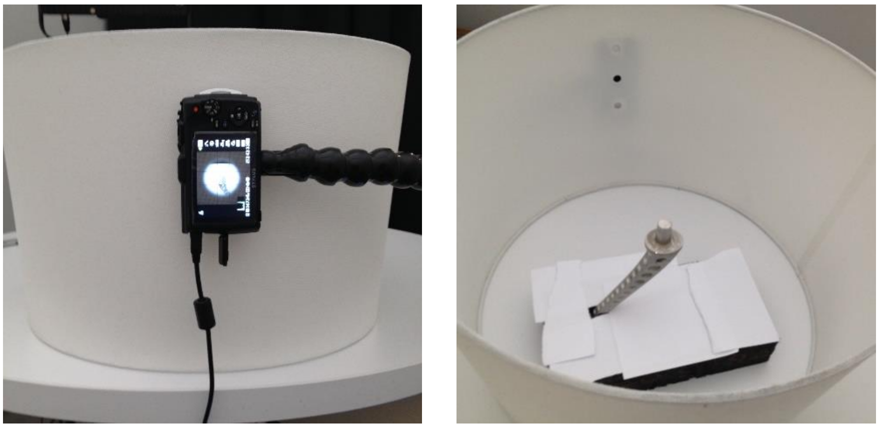


Figure 3-1 The position of the camera with respect to the shade (left image). The distance between the implant the camera was 250mm (right image)

To further minimise the reflection, the implants were placed 250 mm away from the camera which was the maximum distance that still allowed for the camera and the fixation foam to freely rotate 360° inside the enclosed area of the shade (Figure 3-1).

The lighting source was a DC Ikan Daylight LED spotlight that provides illumination with a reasonably low level of fluctuation. To suppress the glare, a fabric diffuser was used to cover the top section of the shade to provide a uniform illumination inside the shade.

Four shots from each of the 138 tapers were captured which resulted in 552 images each displaying a 90° region of the cylindrical geometry of the tapers. Otsu's method was used to segment the stems from the background. Using the oblique sides of the taper, each image was rotated to ensure the top side of the taper is horizontal. Next, the coordinates of the four corners of the taper were obtained and used to locate the middle section (i.e. the 90° region) which constitutes the region of interest (ROI) that represents a quadrant.

Figure 3-2, schematically, illustrates the front view of a stem taper in which P and D subscripts denote proximal and distal regions, respectively. The coordinates of the four top-left (B_P), top-right (E_P), bottom-left (B_D), and bottom-right (E_D) extrema points of the segmented tapers were obtained via blob analysis. Since each image covers the entire 180°, and it was desired to crop the middle 90° ROI, the coordinates of the four corners of the ROI were calculated.

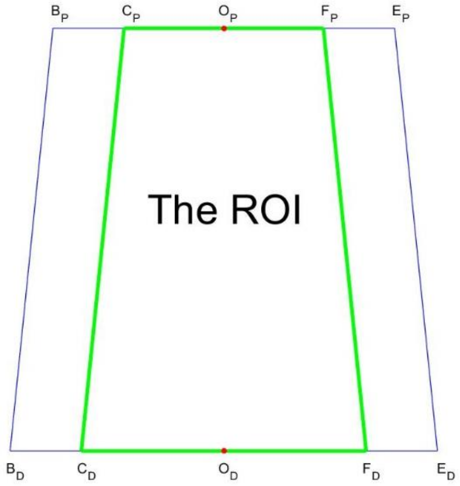


Figure 3-2. The front view of a typical stem taper along with the calculated ROI.

The stem tapers had a circular cross-sectional area, so from the coordinates of the four extrema points, the radii at the distal and proximal regions were calculated as half of the sides BE at these two regions. By using the radii and the extrema points’ coordinates, the midpoint O was located which represents the centre of circular sections.

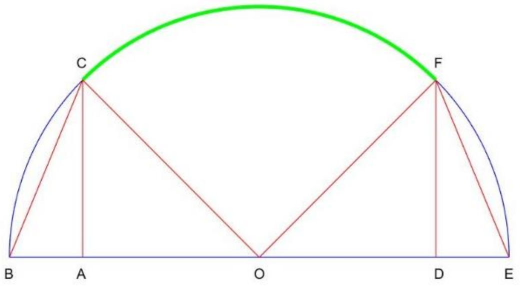


Figure 3-3. The top view of a typical stem taper along with the calculated ROI.

Figure 3-3, schematically shows half of the cross-sectional area of a taper which is the top view of Figure 3-2. Given the radius and the fact that the ROI is the middle 90° segment (arc CF), the sides AB and DE which are equivalent to CB and FE in Figure 3-2 were calculated as follows:

$$OC = OB = r$$

$$\widehat{COF} = 90°$$

$$\widehat{AOC} = \widehat{ACO} = B\widehat{OC} = 45°$$

$$AC = AO$$

$$AC = \frac{r}{\sqrt{2}}$$

$$\widehat{OBC} = \widehat{OCB} = \frac{180° - \widehat{BOC}}{2} = 67.5°$$

$$\tan \widehat{OBC} = \frac{AC}{AB} = \frac{{}^{r}/_{\sqrt{2}}}{AB}$$

$$AB = \frac{r}{\sqrt{2} \times \tan 67.5}$$

Considering the symmetry of this geometry, DE and AB are equal. By subtracting these two offsets from BE, the top and bottom sides of the ROI were determined. Figure 3-4 displays the result for a left joint stem taper where the red lines represent the boundaries of the ROI in each image.

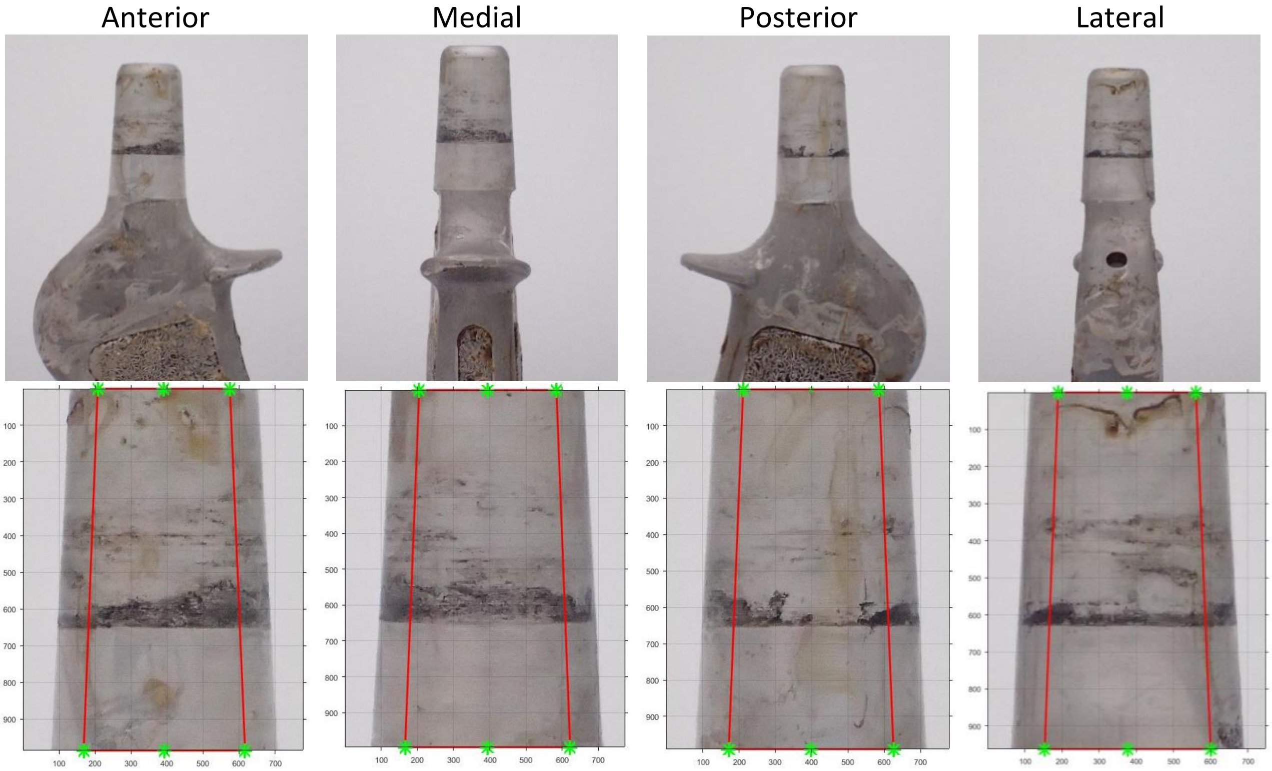


Figure 3-4. A sample of the four quadrants of a taper and the corresponding ROIs.

Subsequently, each image was split from the middle horizontally to obtain the distal and proximal zones at each quadrant. Therefore, 1104 images were obtained from the 138 stem tapers in this work. Figure 3-5 displays the final result of an example taper.

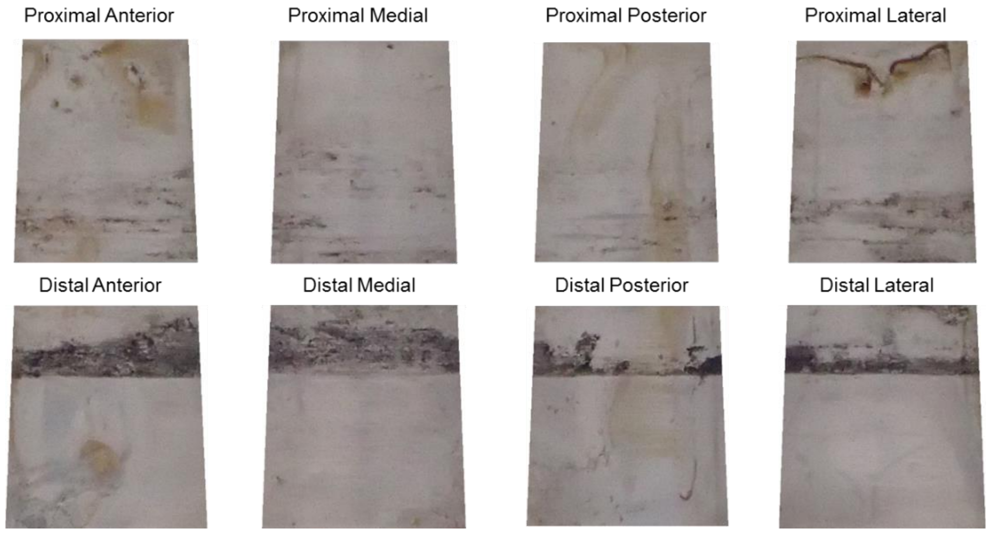


Figure 3-5. Eight images were obtained from each stem taper

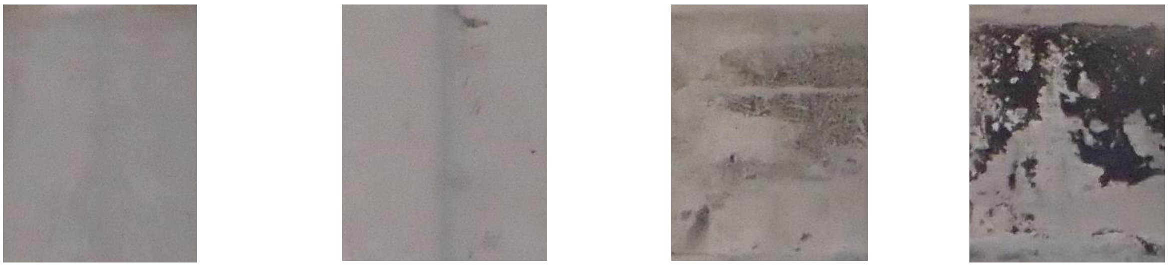
3.4 Visual Scoring

The accuracy of a classification algorithm was to be evaluated by comparing its predicted scores against the actual visual scores. Therefore, it was required to score the images visually to use them as the reference point. Since the aim was for the classification algorithm to utilise the expertise of the experienced investigator (RM) in this work, the investigator visually scored the images according to the Goldberg scoring method (Table) to train a classifier. From each stem taper, one score for each of the eight zones was obtained.

To score consistently, one trained investigator (RM) evaluated the damage. Eight images that correspond to posterior-distal, posterior-proximal, medial-distal, medial-proximal, anterior-distal, anterior-proximal, lateral-distal, and lateral-proximal zones of each stem taper were scored in randomly. Figure 3 displays a sample of each score level.

Table 3-3. The visual criteria for scoring corrosion damage [24].

Score	Corrosion Criteria
1 (None)	No Visible Corrosion
2 (Mild)	<30% Surface Discoloured / Dull
3 (Moderate)	>30% Surface Discoloured / Dull or <10% Containing Black Debris, Pits or Etch Marks
4 (Severe)	>10% of Surface Containing Black Debris, Pits, or Etch Marks



score: 1 score: 2 score: 3 score: 4

Figure 3-6. Scores of 1 through 4 for stem tapers.

3.5 The DIP Algorithm

After acquiring the images and visually scoring them, several features were extracted. These features are scalar properties of an image that provide a succinct and compact representation which can be very useful when dealing with large volumes of images with possibly high resolutions. There are potentially many features which can be selected and used, and this is a hot topic of research in the application of DIP in various contexts [70]. The features used in this study were selected according to the literature of corrosion-related DIP and focused on surface texture which is believed to be closely related to the degree of corrosion damage [61, 70, 77]. In this study, the Computer

Vision System Toolbox of MATLAB was used to extract three groups of global features, namely first and second-order image statistics, and wavelet features; plus one group of local features known as Speeded-Up Robust Features (SURF). The following sections present how these four groups are calculated and discuss their contribution toward the quality of this classification problem.

3.5.1 First Order Image Statistics

As a global descriptor of images, this group of features is derived from histograms of intensity distribution to characterise texture [86]. Similar to many conventional cameras, the colour model of the original images captured in this study was RGB (Red-Green-Blue). To calculate this group of features, first, the colour model was transformed to HSV (Hue, Saturation, and Value) by '*rgb2hsv*' command because it decouples chromaticity (H and S) from brightness (Value). In a set of images, due to luminance fluctuations, the Value colour component may not be reliable and is usually excluded from feature extraction [68, 71].

The first-order image statistics were obtained from the histograms of the Hue and Saturation colour components only. Equation 1 gives the first-order histogram *P(i)* in which *i* represents the grey levels of an image (i.e. 1 ≤ *i* ≤ 256).

$$P(i) = \frac{number\ of\ pixels\ with\ grey\ level\ i}{total\ number\ of\ pixels} \tag{1}$$

From each histogram, six features (mean, standard deviation, smoothness, third moment, uniformity, and entropy) were calculated according to which resulted in 12 (6 × 2) values. Equations 2-7 provide the corresponding mathematical expressions.

$$mean = \sum_{i=0}^{256-1} \left(\frac{i}{256-1}\right) \times P(i) \tag{2}$$

$$standard\ deviation = \sqrt{\sum_{i=0}^{256-1} \left(\frac{i}{256-1} - mean\right)^2 \times P(i)} \tag{3}$$

$$smoothness = 1 - \frac{1}{(1 + standard\ deviation^2)} \tag{4}$$

$$third\ moment = \sum_{i=0}^{256-1} \left(\frac{i}{256-1} - mean\right)^3 \times P(i) \tag{5}$$

$$uniformity = \sum_{i=0}^{256-1} \left(P(i)\right)^2 \tag{6}$$

$$entropy = -\sum_{i=0}^{256-1} P(i) \times \log_2 P(i) \tag{7}$$

Table 3-4 summarises these 12 (6 × 2) features.

Table 3-4. The first-order image features from Hue and saturation components.

1st Order Image Statistics	
Hue_Mean	Sat_Mean
Hue_Standard Deviation	Sat_Standard Deviation
Hue_Smoothness	Sat_Smoothness
Hue_Third Moment	Sat_Third Moment
Hue_Uniformity	Sat_Uniformity
Hue_Entropy	Sat_Entropy

3.5.2 Second-Order Image Statistics

Intensity histograms have limitations including that they do not provide information about the order of the measured intensity frequencies. In other words, they do not reflect the spatial relationships between the pixels and only provides information about their intensities. To overcome this issue, Grey Level Co-occurrence Matrix (GLCM) was used to extract the second group of features. GLCM is a popular method for statistical texture analysis of corrosion defects or other types of surfaces [61, 66, 77].

GLCM is the matrix of relative frequencies $P_{\emptyset,d}(i,j)$ that describes how frequently two pixels with the grey levels *i* and *j* appear in a window separated by an offset distance *d* in direction Ø. The dimensions of a GLCM depends on the range of grey levels that it covers. In a digital 8-bit image, there exists up to 2^8 distinct grey levels which yield up to 256 × 256 GLCMs. Memory requirements may mandate scaling the grey level values to a smaller set, but it results in a reduced grey level accuracy [86].

In this study, from a GLCM, at a particular distance and direction, five numerical features, namely energy, entropy, contrast, correlation, and homogeneity were calculated, according to Equations 8-16 [86, 91].

$$contrast = \sum_{i=1}^{256}\sum_{j=0}^{256}(i-j)^2 P_{\phi,d}(i,j) \tag{8}$$

$$correlation = \frac{\sum_{i=1}^{256}\sum_{j=1}^{256}\left[(ij)P_{\phi,d}(i,j)\right] - \mu_a\mu_b}{\sigma_a\sigma_b} \tag{9}$$

$$\mu_a = \sum_{i=1}^{256} i \sum_{j=1}^{256} P_{\phi,d}(i,j) \tag{10}$$

$$\mu_b = \sum_{j=1}^{256} j \sum_{i=1}^{256} P_{\phi,d}(i,j) \tag{11}$$

$$\sigma_a = \sum_{i=1}^{256}(a-\mu_a)^2 \sum_{j=1}^{256} P_{\phi,d}(i,j) \tag{12}$$

$$\sigma_b = \sum_{j=1}^{256}(j-\mu_b)^2 \sum_{i=1}^{256} P_{\phi,d}(i,j) \tag{13}$$

$$energy = \sum_{i=1}^{256}\sum_{j=1}^{256}\left(P_{\phi,d}(i,j)\right)^2 \tag{14}$$

$$homogeneity = \sum_{i=1}^{256}\sum_{j=1}^{256}\frac{P_{\phi,d}(i,j)}{1+|i-j|} \tag{15}$$

$$entropy = -\sum_{i=1}^{256}\sum_{j=1}^{256} P_{\phi,d}(i,j) \times \log_2 P_{\phi,d}(i,j) \tag{16}$$

The offset distance was considered as 1 pixel, and over four directions, the GLCMs were obtained which resulted in obtaining four GLCMs for each image. Table 3-5 summarises the matrix of distance and direction offsets which correspond to the top, top left, left, and bottom left directions. The positive horizontal and vertical coordinates correspond to the right and upward directions, respectively.

Table 3-5. One offset distance and four directions for GLCM.

Orientation	Horizontal Distance (pixels)	Vertical Distance (pixels)
top	0	1
top-left	-1	1
left	-1	0
bottom-left	-1	-1

A GLCM is symmetrical with respect to the main diagonal because the order of selecting the two pixels is not important. Therefore, the other four directions were not accounted for here. Having four GLCMs for an image from which five features had been calculated resulted in 20 second-order image statistics (Table 3-6). In this study, the '*graycomatrix*' and '*graycoprops*' commands were used to calculate GLCMs and the five features, respectively.

Table 3-6. The second-order image features based on the four directions.

2nd Order Image Statistics			
Direction_1	Direction_2	Direction_3	Direction_4
Energy_1	Energy_2	Energy_3	Energy_4
Entropy_1	Entropy_2	Entropy_3	Entropy_4
Contrast_1	Contrast_2	Contrast_3	Contrast_4
Correlation_1	Correlation_2	Correlation_3	Correlation_4
Homogeneity_1	Homogeneity_2	Homogeneity_3	Homogeneity_4

It should be noted that texture description is highly scale-dependent. When texture elements are large enough, they might be defined in more than one scale. Therefore, image resolution (scale) must be a consistent part of the texture description [86]. Huang et al. [94] maintained that those features extracted from second-order statistics could be inadequate since at a particular scale, some textures with the same numerical features can be easily discriminated by human visual systems. To decrease the problem of scale sensitivity, the texture needs to be described in multiple scales by using a coarse-to-fine multi-resolution strategy [86]. Over the past two decades, multi-resolution analysis techniques such as wavelet transform have received a great deal of attention in various applications of DIP, particularly the analysis of corrosion images [63]. Therefore, wavelet transform was used to extract the third group of global features in this study.

3.5.3 Wavelet Features

It is well established that gradients of intensity in various directions over different scales reflect the texture of an image [96]. An effective method to capture image gradients at different magnitudes and directions is to use 2-D Fourier Transform (FT). Using this transform, an image can be reconstructed to obtain a frequency-space representation. In the context of corrosion, defects appear as abrupt spatial changes in intensity which result in high frequencies in the corresponding FT reconstruction. Using this property, high-pass (HP) and low-pass (LP) filters can be created that pass specific spatial frequencies into the reconstruction while suppressing the others. These filters can be used to visually highlight texture features with particular magnitudes or directions [74].

Wavelet transform is computed by the convolution of the signal (image) and the scaled-shifted versions of a mother wavelet function (e.g. Daubechies, Haar, BiorSplines, and Gaussian). Figure 3-7 shows a single decomposition wavelet transform that is implemented by a bank of 1-D HP and LP filters.

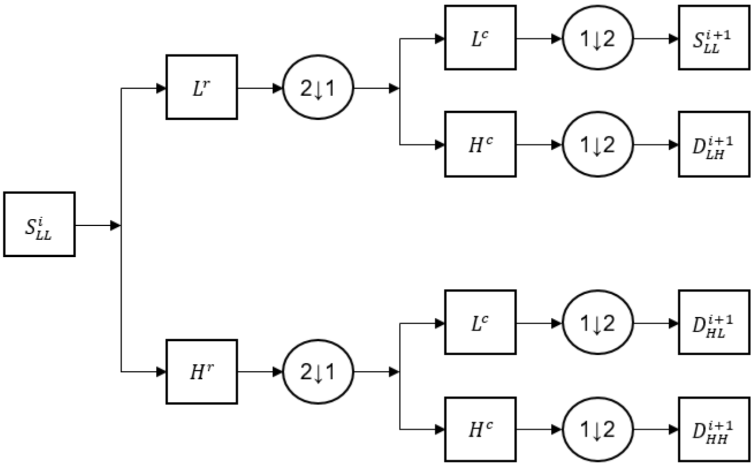


Figure 3-7. 1-D discrete wavelet filters applied first horizontally and then vertically.

Here, S^{i}_{LL} is the input image at resolution level *i*. According to the Nyquist's rule, decomposition of an image with an LP or HP filter yields almost twice as much data. To keep the amount of data almost the same size as the input, the data was down-sampled following each filtering (which is represented by the circle blocks in Figure 3-7). *H* and *L* represent 1-D HP and LP filtering and the *r* (rows) and c (columns) superscripts denote the direction of down-sampling. The circle blocks which perform down-sampling are 2↓1 (keeping one column out of two) or 1↓2 (keeping one row out of two) [66, 95].

The result is the decomposition of the input image into four sub-band images. D^{i+1}_{LH}, D^{i+1}_{HL}, and D^{i+1}_{HH} correspond to the low-high, high-low and high-high bands in the frequency domain, respectively. Also, these images highlight the horizontal, vertical, and diagonal details in the input image. S^{i+1}_{LL} (low-low component) is an LP filtered version of the input image that would be the input for further wavelet decompositions.

At each decomposition level, the wavelet transform coefficients of the three detailed sub-band images produce a number of textural features. Among them, wavelet energy signatures which reflect the distribution of energy along the frequency domain are often employed as corrosion texture features. The energy signature arrays are denoted by $[E_{LH_i}E_{HL_i}E_{HH_i}]$. The length of each array is equal to *i* (decomposition level) [66, 95, 97]. Yan et al. [69] computed a number of features from the energy signatures which were adopted in this study. From the summation of the three energy features, the global distribution of energy at a particular decomposition level (G_i) was calculated (equation 17). Also, the local distribution of energy was calculated (equation 18).

$$G_i = E_{LH_i} + E_{HL_i} + E_{HH_i} \tag{17}$$

$$L_i = \frac{E_{HH_i}}{E_{LH_i} + E_{HL_i}} \tag{18}$$

In general, from a *k* level decomposition of an image, 2*k* global and local features can be obtained. The images were decomposed down to three levels based on the assumption that the decomposition levels should be selected to ensure that the size of the smallest sub-image is greater than 10 × 10 pixels. Therefore, six energy features from the wavelet transform of the images were obtained as summarised in Table 3-7.

Table 3-7. The 6 extracted features from the wavelet transform of an image.

Wavelet Energy		
Decomposition Level 1	Global_1	Local_1
Decomposition Level 2	Global_2	Local_2
Decomposition Level 3	Global_3	Local_3

3.5.4 Speeded-Up Robust Features (SURF)

Unlike global features which are directly extracted from an entire image, two components are required for local feature extraction, namely detector and descriptor. A detector picks the location of a number of *interest points* within an image, and then a descriptor extracts a vector of region descriptors from a neighbourhood of each interest point. A literature review in the other texture analysis applications (e.g. land-use remote sensing and medical image annotation) revealed that Scale Invariant Feature Transform (SIFT) that was introduced by Lowe [142] could provide superior results in classification problems. One reason for SIFT to be popular in texture classification problems is robustness to reasonable levels of 2-D affine transformations, scale, and viewpoint changes [103, 119].

More recently, because of the wider utilisation of local features in new image-based applications, different versions of SIFT have been proposed [110, 111]. In 2008, a refinement on the basic scheme of SIFT known as Speeded-Up Robust Features (SURF) was introduced to address the concerns associated with SIFT computation costs [102, 113]. Therefore, this study used SURF according to Bay et al. [99].

In this work, two methods to define the SURF interest points were compared. The first method is based on using the SURF detector, while the second method relies on a regular grid that is superimposed over an image. The quality of the obtained features from both methods was near identical. Hence, the second method was adopted here due to its lower computation cost.

According to this approach, the interest points are considered as the gridline intercepts of a regular 8 × 8 grid. For each interest point, four neighbourhood blocks (patches) with dimensions of 32, 64, 96, and 128 pixels were used to investigate texture at different scales. In the next step, a SURF descriptor calculated vectors of region descriptors for the patches. The distribution of these vectors in the four classes is summarised in Table 3-8.

Table 3-8. The quantity of the vectors of region descriptors obtained from each class.

Class	Images	SURF
1	364	3307164
2	515	4743644
3	174	1641092
4	51	401788

At each class, 80% of the strongest region descriptors were preserved, and the rest were discarded. Also, to balance the number of descriptors across the four categories, only 321430 of them at each class were used which is equal to 80% of the minimum detected descriptors (i.e. class 4 with 401788 descriptors).

To calculate local features from the region descriptors, a Bag of Visual Words (BOVW) model was used. BOVW identifies visual words (numerical features) that best characterise the images. Since it was desired to extract 1000 SURF features, the descriptor vectors were quantised into 1000 clusters by using k-means clustering. The cluster centroids which act as the visual words were quantified in each image to find a frequency histogram. This histogram constitutes the feature vector, and its length corresponds to the number of visual words (1000). Figure 3-8 illustrates an overview of this process.

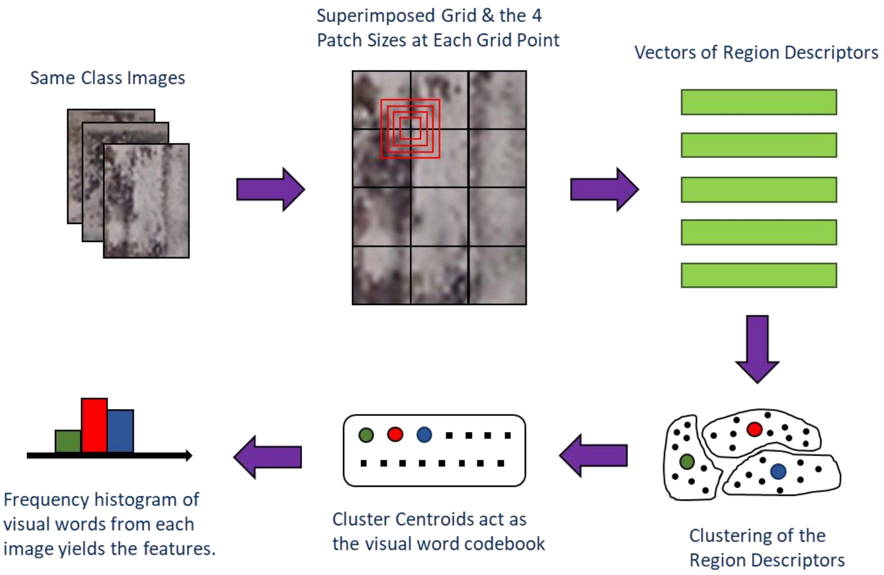

Same Class Images
Superimposed Grid & the 4 Patch Sizes at Each Grid Point
Vectors of Region Descriptors
Frequency histogram of visual words from each image yields the features.
Cluster Centroids act as the visual word codebook
Clustering of the Region Descriptors

Figure 3-8. An overview of the process of interest point selection, extraction of region descriptors, and BOVW.

3.6 Dimensionality Reduction

Describing morphological attributes of corrosion damage is a challenging task. There exists no integrated theory to offer a set of specific features that provide a desirable level of discriminatory power to classify the intensity or typology of corrosion damage. Each study, according to the acquired images, the classification objectives, and the available computational technology has used a specific set of features. To overcome this issue, a relatively wide variety of features compared with similar corrosion-related studies were extracted in this study.

Multivariate data analysis can be used to assess the discriminatory power of the obtained features. Through this process, the contribution of each feature can be determined and subsequently, those with negligible levels of power can be discarded. Dropping those features improves not only the quality of the feature vectors but also reduces their dimensions which in turn lowers the computation cost of the subsequent classification algorithm.

Alternatively, the feature can be combined to obtain a transformed version which produces a simpler description of the system in terms of capturing as much of the variability in data. These two approaches (i.e. feature selection and feature transformation) are for dimensionality reduction, and each can be implemented in several different ways.

It is the nature of the features which determines more suitable methods for a specific problem. Feature selection is preferable to feature transformation when the original units and meaning of features (e.g. categorical features) are important, and the modelling goal is to identify an influential subset. While this is not essential to for DIP features, it is required to maintain this condition for patient and implant factors since they directly influence the decisions of surgeons and manufacturers. Therefore, DIP features will be explored via both feature selection and transformation to identify which approach can yield superior features for machine learning algorithms. Since the number of patient and implant features is very low compared with DIP features, feature selection or transformation do not apply to this group of features.

Another criterion to select a feature selection or transformation technique is whether the data points come with actual classes. If it is desired to discover the patterns within the feature space, unsupervised feature selection techniques are required. However, if it is desired to obtain superior features in terms of predicting the actual classes of data points, supervised classification techniques should be used. The performance quality of one unsupervised and one supervised feature selection and transformation techniques were compared based on incorporating their outcomes in subsequent unsupervised and supervised machine learning methods.

3.6.1 Principal Component Analysis

A popular unsupervised method is Principal Component Analysis (PCA) which is also knowns as Hotelling transform [60, 63, 64, 74, 143, 144]. The common saying ‘Why use a hundred words when ten will do?’ embodies the central idea behind PCA, a very important and powerful statistical technique that aims to express information in its most succinct and compact form. In a feature space with many dimensions (variables), groups of variables often move together due to relatively high correlations between them. In other words, more than one variable might be measuring the same driving *principle* governing the behaviour of the system. This is the primary and essential requirement for PCA to be useful: the value of a given variable in the feature space, to some degree, is expected to be *predictive* of the values of the other variables.

PCA is a feature transformation tool which can be extended to feature selection. As a feature transformation means, PCA synthesises new variables (components) which are a linear combination of the original ones. These components are uncorrelated, obey the same statistical distribution as the original feature vectors, and have a descriptive power that is more easily ordered than the original features. In this case, less descriptive components can be dropped from consideration when building models. Sometimes the context of the problem needs for conducting the classification using (a subset of) the original features. In such situations, via rotation of the matrix of PC coefficients and comparing the correlation of the features with each PC, feature selection is performed.

PCA linearly transforms the feature space to obtain a set of principal axes which are uncorrelated. In doing so, the correlation or covariance matrix is calculated. Here, the latter will be used as methods for statistical inference based on the sample Principal Components (PCs) from the covariance matrix are easier and are available in the literature. Also, if the original features are normalised (mean-subtracted features divided by standard deviation), these two matrices will be identical. The PC coefficients will be calculated by '*pca*' command in MATLAB.

The dimensionality of the PCs is identical to that of the original features. PCA returns the PCs in the order of their contribution in explaining the variability of the original data points. This property of PCA helps with identifying PCs with no or negligible contribution and removing them to reduce the dimensionality of the feature space. Scree plot and Pareto chart can be used for this purpose. A Scree plot sketches the variance of the PCs. Looking at a Scree plot, one has to determine after how many PCs the plot converges to a horizontal line and retain them. Conventionally, PCs with variance values greater than 1 are preserved while the others are discarded. Using this criterion requires for the feature vectors to be standardised to have mean 0 and scaled to have a standard deviation (variance) 1. Alternatively, a Pareto chart can be used which cumulatively illustrates the contribution of the PCs. The first set of PCs which provide a cumulative contribution of at least 95% are preserved, and the rest are discarded. The latter approach is not sensitive to standardising the feature vectors as it addresses the percentages rather than their actual variance.

Once the final number of PCs were determined, the next step is to interpret them. The covariance of the coefficient matrix of the selected PCs may reveal strong correlations of a variable with several

PCs or variables with no strong correlations with any of the PCs. To make the location of the PC axes fit the actual data points better, these axes can be rotated.

There are many different types of rotations. An important difference between them is that they can create PCs that are correlated or uncorrelated with each other. Rotations that allow for correlation are called oblique rotations; rotations that assume the factors are not correlated are called orthogonal rotations. Brown [145] provides a detailed discussion about the selection criteria for the rotation method.

The outcome of the coefficient rotation is called loadings which can be used to find which variables tend to clump together. In other words, a matrix of loadings which has the same dimensions as the coefficient matrix produces factors that have high correlations with one smaller set of variables and little or no correlation with another set of variables. PCA conducts feature selection by discarding those variables with no higher loading at any PC [61]. After removing some of the features, the entire process can be iterated by recalculating the PCA of the reduced feature space and removing the weaker features.

3.6.2 Neighbourhood Component Analysis

As a supervised technique, Neighbourhood Component Analysis (NCA) was implemented by the '*fscnca*' command in MATLAB. NCA selects features intending to maximise prediction accuracy of classification problems. In doing so, it calculates a weight for each feature for minimising an objective function which measures the average classification loss (error) of a 1-nearest-neighbour classifier over the training data. Therefore, besides the feature vectors, the corresponding scores were required to implement NCA.

NCA has a regularisation parameter (λ) which may need tuning to further reduce the generalisation error. Normally, lambda is expected to be a multiple of 1/(number of images). Tuning lambda is carried out by an *n*-fold partitioning of the training data and splitting each fold into training and test subsets. A vector of lambda values which are multiples of 1/(number of observations) can be used to train an NCA model for each fold. Next, the generalisation errors for the corresponding test set in each fold using the vector of Lambda values were calculated. From the average loss values obtained over the folds for each lambda value, the lambda which returned the lowest average loss was selected.

3.7 Machine Learning

So far, the corrosion damage morphology was characterised by several local and global features that replace an image with a representative feature vector. By using these vectors, the images were mapped into an abstract feature space as a point. The coordinates of each point were derived from the elements of its feature vector. In the context of classification, machine learning searches for decision boundaries or clusters that partition the images (points) according to their class in the feature space. To serve this goal, there exists a broad spectrum of machine learning techniques that can be categorised into unsupervised and supervised learning methods.

In the context of visual scoring, the literature has used scores with a different number of levels. Two-level scoring is a binary classification of images which can be used to detect whether there is any corrosion or not. Scoring methods with more than two levels are not just about detection; they rate the severity of the damage.

Unsupervised learning (also known as cluster analysis) explores data and find how many clusters are naturally formed by the feature vectors in the feature space. Cluster analysis served two aims in this study. The first aim was to determine how many patterns exist naturally within the extracted DIP features. The second aim was to see whether the Goldberg scoring technique can be used in less or more than four levels. In other words, whether splitting a score level or combining two or three score levels can potentially improve the clustering results.

On the hand, supervised learning is a straightforward task since the number of desired classes is already known, and it is not necessary to explore the data to determine the number of natural patterns. These algorithms find hyperplanes (decision rules) that classify the images according to their corresponding visual scores.

3.7.1 Unsupervised Learning

One popular unsupervised learning technique is k-mean clustering. It partitions data into *k* mutually exclusive clusters and returns the index of the cluster to which it has assigned each observation. This technique which is considered more suitable for large amounts of data treats each observation in the data as an object having a location in space. It finds a partition in which objects within each cluster are as close to each other as possible, and as far from objects in other clusters as possible. In MATLAB, '*kmeans*' command offers five different distance measures which can be selected depending on the kind of data being clustered. Each cluster is defined by its member objects and its centroid. The centroid for each cluster is the point to which the sum of distances from all objects in that cluster is minimised. Cluster centroids are computed differently for each distance measure to minimize the sum with respect to that distance measure.

If the number of desired clusters is specified as *k*, this algorithm randomly picks *k* data points and assigns them as the centroids of the clusters. Based on the selected distance metric, the remainder of the data points are assigned to the closest cluster centre. Using the mean of the coordinates of the data points in each cluster, a new centroid is calculated. This process is iterated by finding the distances of the data points to the new centroids and forming new clusters. Therefore, every time the centroid and consequently the clusters are updated. This process is iterated until the centroids are not updated any more. The kmeans solution depends to a great extent to the initial cluster centres that were randomly selected. As the final cluster centroids are to minimise the sum of distances from all objects in their clusters, it is important to avoid local minimums. Therefore, MATLAB provides the option to replicate this process by choosing different initial centroid in each replicate. The replicate which yields the lowest total sum of distances (obtained by adding up the total distances of the data points in a cluster to the centroid) has got the best initial centroids.

In MATLAB, *kmeans* function can take in any values as the number of the desired clusters. Therefore, it is essential to evaluate the quality of the obtained clusters so that the optimum value for *k* can be identified. While visualisation of the cluster yields a preliminary idea about the quality of the clusters, the silhouette plot of the cluster indices for each clustering can be used for a more objective evaluation. In this type of plot, each horizontal bar represents a data point, and the bars are vertically grouped according to their assigned classes. That means the total number of bars is equal to the number of data points. The length of each bar is a value ranging from -1, through 0, to +1. This value which is represented by the horizontal axis explains how close a data point is to the points in the neighbouring clusters. It is desired for the bars to extend as much as possible toward +1, indicating points of a cluster that are very distant from neighbouring clusters. A value of 0 indicates a point that is not distinctly in one cluster or another, and -1 is a point that is probably assigned to the wrong cluster.

This process can turn from a graphical into an analytical one by calculating the average silhouette value for each clustering solution. As mentioned before, it is desired for these values to be as high as possible (closer to 1), so the same rule is applied to their average. In MATLAB, *silhouette* function can take in the feature vectors (data points coordinates) and their cluster indices and return the silhouette value for each data point as well as the silhouette graph.

Besides identifying the optimum number of clusters, it is required to evaluate the images within each cluster based on their actual scores similarity to ensure each cluster includes a reasonable number of same score images. It should be noted that while the order matters for the levels of visual scores, clusters do not follow any order. Therefore, several cluster-score configurations based on the desired number of clusters need to be explored to identify which configuration provides the best accuracy rate. Within these configurations, each score level must be assigned to a unique cluster, yet more than one distinct score levels can be assigned to a unique cluster. Also, when more than one score level is assigned to a cluster, the difference between the score levels must not exceed unity.

3.7.2 Supervised Learning

In the previous chapter, SVM was determined as the machine learning method that has outperformed other types of supervised learning algorithms in characterising corrosion and many other texture analysis applications. On this ground, it was adopted in this study as the classification method.

In this work, there were four classes (i.e. Goldberg score levels) which mandated using a multi-class classifier. To utilise SVM in a multi-class problem, '*fitecoc*' command in MATLAB was used which offers an ensemble method known as Error-Correcting Output Codes (ECOC). It includes a combination of binary SVM classifiers according to a particular coding design. The ECOC coding design used in this study was 'one-versus-one'. In this design, for each binary learner, one class is positive, another is negative, and the rest are ignored. This coding design exhausts all combinations of class pair assignments. Table 3-9 displays the coding design for this four-class problem.

Table 3-9. Six binary learners were trained for a four-level classification problem.

	Learner 1	Learner 2	Learner 3	Learner 4	Learner 5	Learner 6
Class 1	1	1	1	0	0	0
Class 2	-1	0	0	1	1	0
Class 3	0	-1	0	-1	0	1
Class 4	0	0	-1	0	-1	-1

Having *k* classes yields *k*(*k*-1)/2 binary learners according to 'one-versus-one' coding design. Hence, in this four-class problem, six binary learners were created. Learner 1 was trained on observations having Class 1 and Class 2 and treated Class 1 as the positive class and Class 2 as the negative class. The other learners were trained similarly. Each binary learner used 80% of the images in each of the two classes for training and the remainder 20% for testing. The classifier assigned an image to the class yielding the minimum average of the binary losses over the six binary learners. The binary loss was a function of the class and classification score that determined how well a binary learner classified an observation into the class assigned to it.

Two hyperparameters, namely Box-Constraint and Kernel-Scale, were optimised to enhance the performance of this SVM classifier. The incorporated Bayesian optimisation attempts to minimise the cross-validation error by varying the hyperparameters.

To evaluate the performance of the classification model, three standard measures were utilised, namely, no validation accuracy, cross-validation accuracy, and confusion matrix. The training sample accuracy was calculated using the entire images for training the classifier and finding what percentage of the images could be successfully classified by that model. This accuracy rate may not be reliable though due to overfitting concerns which occur when a model learns the detail and noise in the data to the extent that it negatively impacts the performance of the model on new data. This means that the noise or random fluctuations in the training data may be picked up and learned as concepts by the model. It usually leads to obtaining a model that classifies the data too well. To eliminate overfitting, fivefold cross-validation was used which partitioned the entire data into five disjoint sets (folds) randomly so that each fold has roughly equal size and roughly the same class proportions as in the vector of visual scores. The first classification model is trained using the observations in the last four folds and reserves the observations in the first fold for validation. The second classification model is trained using the observations in the first fold and the last three folds and reserves the observations in the second group for validation. The cross-validation proceeds in a similar fashion for the remainder classification models. The cross-validation accuracy was calculated as the mean of the five validation results. The third measure was confusion matrix which tabulates the predicted scores against the true scores in a classification model and reveals the quantity of misclassified images for each score level.

3.8 Statistical Analyses

In this study, two sets of scoring were conducted. First, each stem received eight visual corrosion scores that are associated with eight distinct zones on the surface. These scores were used to investigate the severity and spatial distribution of corrosion across the eight zones. Second, each stem taper was scored holistically so that a single score represents the corrosion damage. These scores were used to investigate the role that patient and implant factors play in corrosion. In this study, a univariate and a multivariate OLR were used for the first and the second groups of the statistical analysis. Selection of OLR over the other causal-explanatory statistical modelling is based on the advantages of regression over its counterparts, as discussed in Chapter 2.

3.8.1 Binomial Versus Ordinal Regression

The OLR model in this study uses cumulative logits. Selection of cumulative logits against other models (e.g. adjacent or continuation categories) was due to the interest of this study to use the entire response scale regardless of the score level. Specifically, cumulative odds ordinal logistic regression with proportional odds will be used.

The odds of an event occurring is the probability of it occurring versus the probability of it not occurring. The natural log of the odds of an event occurring is called logit. A logit is important as the log odds of an event occurring (a success) can be modelled as a linear expression of a set of Independent Variables (IVs), which occurs in binomial logistic regression. Assume the Dependent Variable (DV) is whether there is any corrosion on the surface or not with a dichotomous response of either ‘’Yes” (the success category) or “No” (the failure category). Equation 19 shows the relationship between the logit of success for this dichotomous DV and two IVs (age and weight).

$$logit(success) = \ln\left(\frac{Prob(success)}{Prob(failure)}\right) = \alpha + \beta_1 \times age + \beta_2 \times weight \tag{19}$$

Using this configuration, it is possible to generate intercept (α) and slope (β) coefficient terms through an iterative process that maximizes the log-likelihood of the outcome. The intercepts provide a useful and meaningful interpretation of the model, including the calculation of odds ratios and probabilities. This can allow the effect of the IV to be considered.

However, corrosion score is not dichotomous due to possessing more than two levels, and to maintain this characteristic for this DV, ordinal logistic regression has to be considered as a series of binomial logistic regressions that run simultaneously on cumulative logits. Cumulative logits split the DV in two with lower values (categories) considering the event/success and all higher values considering the non-event/failure.

By extending the previous example so that instead of simply asking whether there was any corrosion on the surface, we asked how much damage was present on the surface. The answer to this question could be “none” (score = 1), “mild” (score = 2), “moderate” (score = 3), or “severe” (score = 4). Given this setup, the first cumulative logit would be the natural log of the odds of the first category ("none") versus all higher categories (i.e., "mild", "moderate" and "severe"), which as expressed by equation 20.

$$logit(success) = \ln\left(\frac{Prob(score \leq 1)}{Prob(score > 1)}\right) = \alpha_1 + \beta_{1.age} \times age + \beta_{1.weight} \times weight \tag{20}$$

The subscript 1 used for the intercept and slope coefficient terms reflects the fact that this is the first cumulative logit.

This binomial logistic regression has now captured some of the ordinal nature of the ordinal DV. It has done this by allowing us to express the effect that the IVs have on whether a lower or higher score on the DV is probable, but only for this particular cumulative logit (i.e., "none" or lower versus higher than "none"). To extend this analysis, the next cumulative logit, which uses the first two categories ("none" and "mild") versus all higher categories (i.e., "moderate" and "severe" combined) could be expressed in a binomial logistic regression (equation 21).

$$\begin{aligned} logit(success) &= \ln\left(\frac{Prob(score \leq 2)}{Prob(score > 2)}\right) \\ &= \alpha_2 + \beta_{2.age} \times age + \beta_{2.weight} \times weight \end{aligned} \tag{21}$$

Similarly, the final cumulative logit (equation 22) uses the first three categories ("none", "mild" and "moderate") compared to higher categories (i.e., the "severe" category), and is expressed in a similar format.

$$\begin{aligned} logit(success) &= \ln\left(\frac{Prob(score \leq 3)}{Prob(score > 3)}\right) \\ &= \alpha_3 + \beta_{3.age} \times age + \beta_{3.weight} \times weight \end{aligned} \tag{22}$$

In summary, splitting the categories of the ordinal DV to run cumulative logits is shown in Table 3-10.

Table 3-10. An ordinal DV with four levels giving three cumulative probabilities and consequently logits.

Binomial Regression	Event Category	Non-Event Categories
1	Probability (score ≤ 1) “none”	Probability (score > 1) “mild”, “moderate”, and “severe”
2	Probability (score ≤ 2) “none” and “mild”	Probability (score > 2) “moderate” and “severe”
3	Probability (score ≤ 3) “none”, “mild”, and “moderate”	Probability (score > 3) “severe”

Each binomial logistic regression now predicts the probability of being classified into the 'lower' categories as opposed to the 'higher' categories for each dichotomization of the ordinal DV based on cumulative probabilities.

3.8.2 Assumptions of Ordinal Logistic Regression

In order to run an ordinal logistic regression, there are four assumptions that need to be considered. The first two assumptions relate to the nature of the study, while the other two assumptions relate to the characteristics of the data.

3.8.2.1 *Assumptions #1 & 2*

The first two assumptions of an ordinal logistic regression relate to the study design. The first assumption mandates the dependent variable is having an ordinal level of measurement, a categorical variable in which order is important. The nature of this study confirms this fact as corrosion scores have an ordinal nature with four distinct categories.

Under the second assumption, there should be at least one IV that is continuous, ordinal or categorical (including dichotomous variables). However, ordinal IVs must be treated as being either continuous or categorical. They cannot be treated as ordinal variables when running an ordinal logistic regression.

Table 3-1 has characterised the IVs in terms of their level of measurement and their categories. As can be seen, there exists no ordinal IV in our study. The other two assumptions are related to how the data fits the ordinal regression model.

3.8.2.2 *Assumption #3*

Under this assumption, there should be no multi-collinearity between the IVs. Multi-collinearity occurs when there are two or more IVs that are highly correlated with each other. This leads to problems with understanding which variable contributes to the explanation of the DV and technical issues in calculating an ordinal logistic regression.

Since multi-collinearity is associated with only the IVs, it can be determined by using the same method used for multiple linear regression, despite the dependent variable being ordinal and not continuous. However, before implementing this, one should consider that regression procedures do not accept categorical variables directly into the model. Based on the second assumption, categorical IVs can only be of a nominal type. To enter a categorical IV directly into the regression equation would be incorrect because the regression equation would assume that the variable is continuous and treat it accordingly.

When a categorical IV with *k* categories is present, it is required to recode it into *k-1* dummy variables. Each dummy variable indicates whether a case (e.g., participant) is a member of the category represented by that dummy variable. With dummy coding, a case is coded as "1" if it is a 'member' of that category and "0" if it isn't. Creating a dummy variable for the k^{th} category would be redundant since the last category will be coded "0" on the entire *k-1* dummy variables. The k^{th} category (the category with missing dummy variable) is called the reference category, and it plays an important role in the interpretation of the parameters of the logistic regression model. It should be noted that any category in a nominal IV can act as the reference category.

Equation 23 shows an example of a regression model with numerical *age* and categorical *stem_material* (CoCr, Stainless Steel, and Titanium) as the IVs.

$$\ln\left(\frac{Prob(score \leq j)}{Prob(score > j)}\right) = \alpha_j - \beta_{age} \times age + \beta_{stemCoCr} \times stemCoCr + \beta_{stemTitanium} \times stemTitanium \tag{23}$$

Here, the category of Stainless Steel in *stem_material* has been considered as the reference category, and therefore, *stem_material* has been recoded into two dummy variables *stemCoCr* and *stemTitanium*.

There are two methods to assess the multi-collinearity. The first method uses a correlation matrix (e.g. Pearson's product-moment or Spearman's rank-order) which is a measure of the strength and direction of the association/relationship between continuous or ordinal variables (Spearman) or just continuous variables (Pearson). Correlation coefficients above and below +0.8 and -0.8 respectively are considered too high conventionally. The second method can be implemented by incorporating collinearity diagnostic under linear regression which returns the Variance of Inflation Factor (VIF). VIF indicates to what extent a particular IV is contributing to multi-collinearity issues within the dataset. The higher the number, the bigger the problem caused by this IV. A general rule of thumb is that numbers in excess of 10 indicate a very big multi-collinearity problem. Unlike the first alternative, VIF does not indicate which IVs are showing strong correlation together. Also, since the reference category of nominal IVs is not inserted into regression models, changing the reference category alters the obtained VIF values which limits the obtained information from this method.

3.8.2.3 *Assumption #4*

As pointed out previously, cumulative odds ordinal logistic regression fits the data into more than one regression model due to the splitting of the categories of the ordinal DV to run cumulative logits. The problem with this approach of separately running multiple binomial logistic regressions is that the effect of the independent variables can differ for each cumulative logit. This means that it would not be possible to make an overall statement about the effect of an IV on the ordinal DV, but must qualify the statement to each specific cumulative logit. The above problem is expressed by the subscript 1, 2 and 3 in equations 20-22 (i.e., intercept and slope coefficients).

To overcome this problem, the fourth assumption comes into play. Under this assumption, the effect of each IV is assumed to be identical at each cumulative logit and, as such, the slope coefficients are constrained to be the same. Consequently, the odds would be the same for each cumulative logit.

This assumption will lead to a model (expressed in logit terms) where the intercepts (called thresholds) will differ for each cumulative logit, but the slope coefficients will remain the same. The result is transforming the three cumulative logit equations into three new equations 24-26 in which the slope coefficients for the corresponding IVs would be the same, and only the thresholds vary.

$$logit(success) = \ln\left(\frac{Prob(score \leq 1)}{Prob(score > 1)}\right) = \alpha_1 - (\beta_{age} \times age + \beta_{weight} \times weight) \qquad (24)$$

$$logit(success) = \ln\left(\frac{Prob(score \leq 2)}{Prob(score > 2)}\right) = \alpha_2 - (\beta_{age} \times age + \beta_{weight} \times weight) \qquad (25)$$

$$logit(success) = \ln\left(\frac{Prob(score \leq 3)}{Prob(score > 3)}\right) = \alpha_3 - (\beta_{age} \times age + \beta_{weight} \times weight) \qquad (26)$$

Therefore, these three binomial regression models transform into three new ordinal logistic regression model as in equation 27.

$$\ln\left(\frac{Prob(score \leq j}{Prob(score > j}\right) = \alpha_j - \beta_{age} \times age + \beta_{weight} \times weight \qquad (27)$$

In this equation, *J* categories of the ordinal dependent variable with *J* – 1 cumulative logits exist. This set up means that it would not be necessary to qualify our statements to particular cumulative logits when declaring the effect of an IV on the DV. This is more similar to linear regression. However, this parsimonious model comes at a price: the assumption of proportional odds has to be ascertained whenever a cumulative odds ordinal logistic regression is deployed.

3.8.3 Ordinal Logistic Regression Outcomes

This section provides an overview of various statistics that are generated after running Ordinal Logistic Regression in SPSS. The outcomes and their interpretation which are of more interest will be elaborated on. There are two main objectives that can be achieved with the output from an ordinal logistic regression: (a) determining which IVs (if any) have a statistically significant effect on the DV, and (b) determining how well the ordinal logistic regression model predicts the DV.

Based on the literature of retrieval studies, only the first type of outcome has always been of interest. However, the second type of outcome which determines the accuracy of regression models can be used to compare the power of this linear approach with more advanced supervised learning approaches. This matter requires immediate attention due to the potential to create dynamic prediction models that can act as decision support systems. Being trained by the vast amount of existing operation records and outcomes, these models use the patient characteristics and implant properties to assist clinicians in choosing implants with lower risks of failure. DVs such as time to revision or failure risk within a specific time-frame can be predicted by these models. This may serve the main aim of retrieval studies which is about learning lessons from the past experience to make more reliable decisions in the future.

3.8.3.1 *Cell Patterns and Overall Fit*

There are three categories of tests to interpret the overall model fit.

- two overall goodness-of-fit-tests
- the likelihood-ratio test
- three pseudo-R^2 measures

However, to do this correctly, the observed frequencies in each cell pattern should be large. Here, a cell pattern represents a unique combination of values of the IVs and DV that exist in the dataset. When a continuous variable is included as an IV (e.g. age), it is unlikely that cell sizes will be adequate due to having a large number of cells with zero frequency. This is because a continuous variable can take on a large number of different values. Therefore, overall goodness-of-fit statistics described later should be treated with suspicion when a continuous IV is present and/or there are a large number of cells with zero frequency.

Overall goodness-of-fit-tests: SPSS Statistics generates two tests of the overall goodness-of-fit of the model. These are the *Pearson* and *Deviance* goodness-of-fit tests. They are designed to provide a measure of how poorly the model fits the data (or the variation in the model that cannot be explained).

The *Pearson* goodness-of-fit test provides this measure by calculating an overall summary measure of the Pearson residuals. Alternatively, the *Deviance* goodness-of-fit statistic is the difference in fit between the current model and a full model; a full model being a model that fits the data perfectly. Hence, one is looking for these tests to be not statistically significant to indicate a good model fit. Neither of these tests will provide reliable tests of goodness-of-fit if there are many cells with zero frequencies and/or small expected frequencies and are generally not recommended.

The likelihood-ratio test: A better method of assessing model fit is to look at the change in model fit when comparing the full model to the intercept-only model. Here, a full model represents a model with the intercept and all independent variables. In this model, the coefficients of the IVs have been estimated using an iterative process that maximizes the log-likelihood of the outcome.

The intercept-only model describes a model that does not control for any IVs and simply fits an intercept to predict the DV.

The smaller the -2 log-likelihood value, the better the fit. As such, a greater difference between the -2 log-likelihood of these two models implies the IVs are better at explaining the DV. Here, a significant difference indicates that the IVs add statistically significantly to the model or, put another way, at least one IV is statistically significant.

pseudo-R^2 measures: Logistic regression does not have an equivalent to the R^2 that is found in ordinary least-squares linear regression which quantifies the proportion of variance for the DV explained by the IVs. Therefore, they are referred to as "pseudo" R^2 measures, and it is recommended to interpret them with great caution.

3.8.3.2 *Overall Parameter Estimates*

The parameters in this model consist of the thresholds (intercepts) and the slope coefficients. As pointed out earlier, the type of ordinal regression model we are running produces an equation for each cumulative logit. As there are four categories of the DV, there are three cumulative logits and, therefore, three equations. Also, the assumption of proportional odds constrains the slope coefficients to be the same for all the three equations, so it is just the thresholds that differ between the three equations.

The slope coefficients represent the change in the log odds of being in a specific category with respect to the reference category (or a unit change in a numerical IV). However, changes in log-odds do not have much intuitive meaning. Much better is to report changes in terms of the odds; that is, the ratio of the odds between any two categories (or a unit change in a numerical IV), which is called the odds ratio (OR). The OR is the exponential of the log odds of the slope coefficient. To further clarify the effect of a particular IV (categorical/numerical) on the odds ratio, the 95% confidence intervals of the OR, and whether the effect is statistically significant have to be reported.

While analysing numerical and dichotomous IVs is fairly straightforward, polytomous IVs incur additional calculations to complete an overall test of statistical significance. SPSS Statistics requires as many orthogonal contrasts as there are degrees of freedom (i.e., one less than the number of categories in an IV) to provide an omnibus test of statistical significance. The reason for this is the fact that for categorical variables, always one category is taken as reference, and the rest of the categories are compared with that. Hence, to exhaust the entire pairwise comparison of the categories, more than one significance test would be required. In each significance test, the polytomous IV has to be recoded into a new variable with the desirable reference category coded as the last category (highest level).

If the study design is interested in investigating the OR of a specific set of categories, not all orthogonal contrasts would be required to explore. Also, it would be wise first to establish whether that polytomous IV is statistically significant overall before exploring any specific contrasts. SPSS statistics provide the Test of Model Effects table which reports an overall test of significance for each variable (not the dummy variables) entered into the logistic regression model.

3.8.3.3 *Predictions and Model Fit*

One way of assessing whether a model fits the data well is to see how well it can predict the DV. However, what one would like to know is how well the model predicted the correct response. In binomial logistic regression, only one estimated probability is required because this probability is of the event occurring or not.

However, in ordinal logistic regression, it is more complicated, with more than one level to predict. The ordinal regression estimates the probability that a case will be classified into each of the levels of the ordinal DV. As such, there will be as many predicted probabilities as there are categories of the DV (e.g., four predicted probabilities in this study) and they should sum to 1. The model selects the response (level) with the largest predicted probability.

SPSS statistics name these probabilities as EST1_1 through EST4_1 (four score levels). For each observation, the predicted level is provided by PRE_1. Also, for any observation, PCP_1 and ACP_1 carry the information about the highest probability and the probability of the actual level, respectively. In order to get a better idea of how the observed and predicted levels are related based on this version of assessing the model fit, a confusion table can be created.

As discussed previously, a table of confusion provides an overall idea about the performance of the model. Very high accuracy rates may indicate overfitting which implies the model will be performing poorly with new unseen data. One way to overcome this concern would be cross-validation.

3.8.3.4 *Alternative Predictive Models*

OLR is usually considered as a baseline for evaluating more complex machine learning methods. The Classification Learner app in MATLAB was used to compare the performance of a selection of different classification models.

Due to the limited size of the data, first, the models were fit to the entire dataset (no holdout or cross-validation). If a reasonable accuracy was achieved, five-fold cross-validation was also performed to eliminate overfitting concerns. The accuracy level usually drops after cross-validation. However, this approach ensures the model can be generalised to unseen new data without too much variation in the performance quality.

3.9 Summary

This chapter elaborated on the methodology that has been implemented in this work. The two important analytical phases, namely quantification of corrosion at stem tapers and exploring associations between corrosion and implant/patient attributes by ordinal logistic regression were detailed out.

Ordinal Logistic Regression models assume the classes (levels) are linearly separable. So, they may not be powerful in classifying nonlinear systems. Considering the diverse range of IVs, it might be required to compare the accuracy of these models with other peer-reviewed studies to learn where a particular model stands concerning those developed in other studies with similar IVs and DV. This aspect of regression models has not been explored in the literature of retrieval studies.

4 OBJECTIVE VISUAL SCORING[1]

[Image removed due to copyright restriction]

[1] This chapter is based on the work that has been published in Medical Engineering and Physics; Milimonfared, R., et al., *An intelligent system for image-based rating of corrosion severity at stem taper of retrieved hip replacement implants.* Medical engineering & physics, 2018.

4.1 Introduction

Following the proposed methodology, this chapter will provide the results of applying DIP and machine learning to objectively score corrosion at stem tapers. DIP was used to extract several characteristic features from the images. Two types of unsupervised as well as supervised feature selection reduced the dimensionality of the feature space to investigate whether the performance of machine learning algorithms can be improved. A range of unsupervised and supervised machine learning techniques was deployed, and their performance metrics were compared.

4.2 Feature Extraction

Table 4-1 summarises the three groups of global features obtained from the HSV and grey-scale images. There are 12, 20, and six features that belong to the first-order image statistics, second-order image statistics, and wavelet groups, respectively.

Table 4-1. An overview of the 38 extracted global features.

1st Order Image Statistics		2nd Order Image Statistics			
Hue_Mean	Sat_Mean	Direction_	Direction_	Direction_	Direction_
Hue Standard	Sat Standard	Energy 1	Energy 2	Energy 3	Energy 4
Hue_Smoothnes	Sat_Smoothnes	Entropy_1	Entropy_2	Entropy_3	Entropy_4
Hue_Third	Sat_Third	Contrast_1	Contrast_2	Contrast_3	Contrast_4
Hue_Uniformity	Sat_Uniformity	Correlation	Correlation	Correlation	Correlation
Hue_Entropy	Sat_Entropy	Homogene	Homogene	Homogene	Homogene
Wavelet Energy					
Global_1	Global_2	Global_3	Local_1	Local_2	Local_3

These 38 global features and 1000 local features (SURF) yielded 1038-D feature vectors. From the 1104 images, an 1104 × 1038 matrix was obtained in which the rows represent the images, and the columns are the features.

4.3 Dimensionality Reduction

Feature selection and transformation techniques were utilised to identify features suitable for supervised and unsupervised machine learning. Section 4.3.1 fulfils this task for unsupervised machine learning, while section 4.3.2 identify a subset of features for supervised machine learning.

4.3.1 Principal Component Analysis

PCA was applied to the 1104 × 1038 matrix of DIP features. The high number of features prevents visualising them by scatter plots. Hence, Pareto charts and Scree plots were generated to evaluate the quality of the PCs in capturing as much of the variability in the features. These methods use different mechanisms to reduce the dimensionality of the feature space.

Figure 4-1 illustrates the Pareto chart which cumulatively plots the contribution of the PCs. Pareto function in MATLAB preserves the first set of PCs which provide a cumulative contribution of at least 95% and discards the rest.

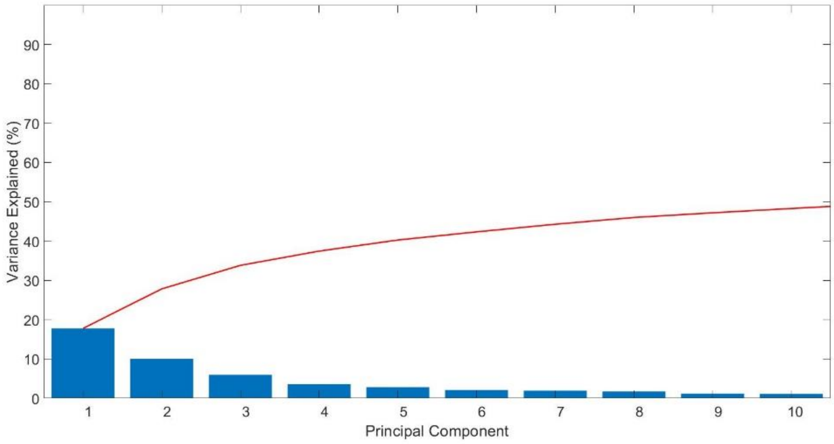


Figure 4-1. The Pareto chart for the first 10 PCs extracted from the 1038 features

As can be seen, the cumulative contribution did not reach 95%. Checking the array of cumulative variance percentage explained by the PCs revealed that the index of the first PC that exceeds 95% had been 347. Since it was not possible to sketch the Pareto chart of the first 347 PCs, only the first 10 PCs were illustrated here. According to the Pareto chart, it can be concluded that preserving the first 347 PCs produces a simpler description of the system while as much (95%) of the variability in the data can be captured.

As an alternative to the Pareto chart, a Scree plot can be employed to visualise the variance of the PCs. Conventionally, PCs with variance values greater than one are preserved while the others are discarded. Scree plot demands the feature vectors to be standardised first. Figure 4-2 displays the scree plot for the same 10 PCs.

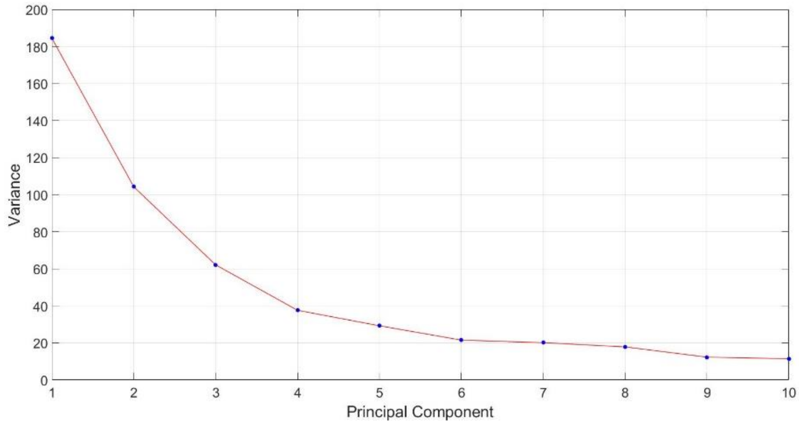


Figure 4-2. The Scree plot for the first 10 PCs extracted from the 1038 features

After looking at the array of PC variance values, PC 160 observed to be the first component that its variance falls below 1. Considering that PCs are in descending order of component variance, it is evident that the PCA of the feature vectors does not provide PCs that explain large portions of the data variance. This Pareto chart suggests using twice as large as the number of PCs that were identified via the Scree plot. Still, taking the first 347 PCs reduces the dimensionality by almost a third (1038/3).

From the first 347 PCs, a matrix of loadings (obtained from the rotation of PC axes) with 1038 rows (features) and 347 columns (rotated PCs) was established to identify variables with no higher loading at any of the 347 PCs to be consequently discarded. Conventionally, correlation values higher than 0.3 are considered, so a threshold of 0.3 for the correlations was set that resulted in the selection of 307 DIP (12 global and 295 local) features. Table 4-2 lists the selected 12 global features.

Table 4-2. The global features having higher correlation with the 347 PCs

text_stat_Hue_4	text_stat_Hue_5	text_stat_Sat_5	cooc_contrast_2
cooc_contrast_3	cooc_contrast_4	cooc_correlation_1	cooc_correlation_2
cooc_correlation_3	cooc_correlation_4	E_global_1	E_local_1

This process can be iterated to further reduce the dimensionality of the selected DIP features. Since the subsequent machine learning algorithms determine the suitability of the selected feature set for this classification problem, further feature selection was not carried out to see first how the dimensionality reduction will influence the performance of the classification algorithms in section 4.4.

Besides Pareto chart and Scree plot which uses some rules of thumb to decide on the number of the required PCs, the performance of machine learning methods can be compared when using the original feature space compared with subsets of the corresponding PCs. This matter will be investigated in section 4.4.1 to see whether feature selection or transformation offer the same quality of results.

4.3.2 Neighbourhood Component Analysis

Feature selection by NCA may not necessarily improve the quality of classification algorithms. Therefore, the first step was to evaluate the suitability of NCA. This was performed by dividing the images into training (80%) and test (20%) subsets and then computing the generalisation error using one as the weight for the entire features. In the second step, the error was calculated according to the feature weights that were determined by the NCA model. The result decreased the error from 0.3800 to 0.2900. Since the weights by NCA induced a lower generalisation error, NCA deemed suitable for improving the quality of the features.

Since the total number of images was 1104, 20 coefficients, ranging from 0 to 20, were randomly generated to be multiplied by 1/1104. Also, a 5-fold cross-validation partitioning of the training set with a division into training (80%) and test (20%) subset was performed. The NCA model used these 20 Lambda values to calculate 20 generalisation errors in each fold. Having five-folds, the result was a 20 × 5 matrix. To find the average error for each Lambda, the average of the elements in each row was calculated. The outcome was a vector with 20 generalisation errors that were averaged across the five folds. Figure 4-3 displays these twenty values.

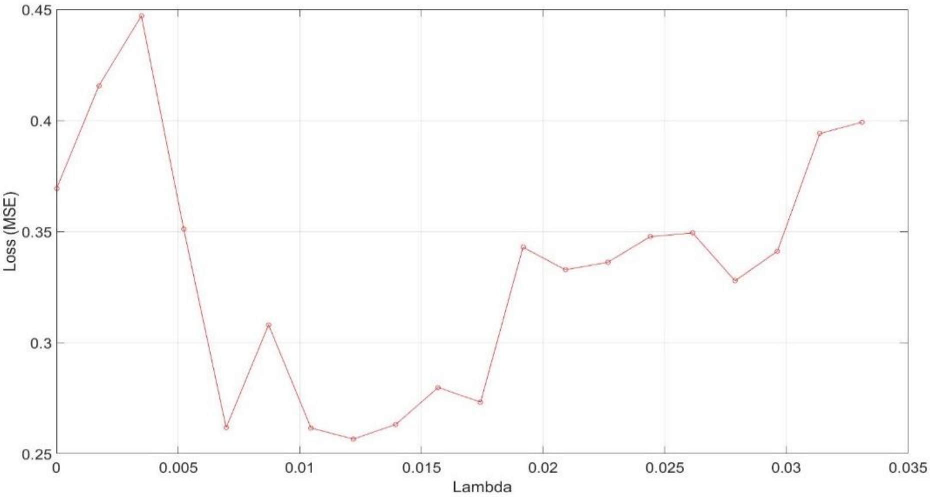


Figure 4-3. Optimization of Lambda in NCA minimises the loss.

The Lambda value which returned the lowest average generalisation error over its five-folds was 0.0122. According to this Lambda value, the generalisation error was further reduced to 0.2566. This

Lambda value was used by the NCA algorithm to perform feature selection. Optimising the Lambda parameter in the NCA algorithm reduced the dimensionality of the feature space from 1038 to 25. The feature weights obtained by NCA are shown in Figure 4-4.

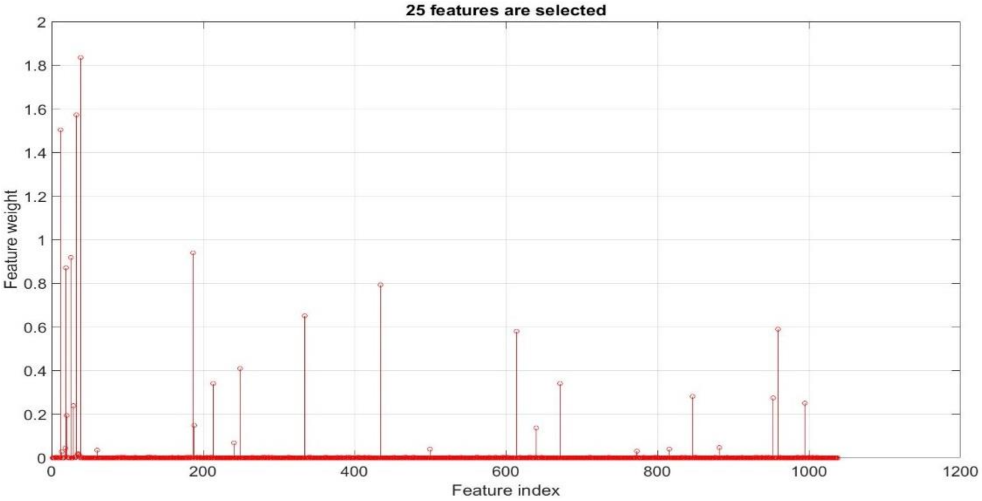


Figure 4-4. Feature loadings by NCA.

In this graph, the first 38 features are global, and the remainder is the 1000 SURF features. The maximum weight (1.836) was associated with the 38th feature which was the local energy at decomposition level 3. Many features were observed to have negligible weights, and only those with a weight greater than 2% of the maximum weight (0.036) were selected. The result was a selection of 25 features which were comprised of 8 global and 17 local features. The result was a selection of 25 features as summarised in descending order by weight in Table 4-3. The expressions of the features were stemmed from Table 4-3. The selected features sorted according to their NCA weight. The index of the SURF features reflects the index of the corresponding visual word.

Table 4-3. The selected features sorted according to their NCA weight.

Feature Name	Weight	Feature Name	Weight	Feature Name	Weight
Local_3	1.836	SURF_features_576	0.579	SURF_features_150	0.149
Global_1	1.573	SURF_features_211	0.408	SURF_features_602	0.135
Sat_Entropy	1.502	SURF_features_633	0.340	SURF_features_203	0.068
SURF_features_149	0.939	SURF_features_175	0.339	SURF_features_844	0.046
Homogeneity_1	0.918	SURF_features_808	0.280	Correlation_2	0.043
Correlation_3	0.869	SURF_features_915	0.273	SURF_features_462	0.039
SURF_features_396	0.792	SURF_features_957	0.248	SURF_features_778	0.038
SURF_features_296	0.650	Entropy_1	0.237		
SURF_features_921	0.588	Correlation_4	0.193		

Aside from the first-order image statistics that characterise colour in an image, the other three groups of global features had representatives in the subset of selected features. In this study, the colour was not expected to be a prominent feature within the captured images as it is well established that corrosion at taper junctions of hip replacement implants appears as regions with discolouration, dullness, or black debris [31].

4.4 Machine Learning

This section elaborates on the outcomes of the unsupervised and the supervised machine learning methods. Considering the diverse range of the available techniques, they were narrowed down to k-mean clustering in the unsupervised group based on its popularity and ease of implementation. For the supervised group, SVM was hypothesised as a potentially superior technique after the review of the literature. This needed to be verified for this study dataset before embarking on formulating a refined SVM algorithm.

4.4.1 Unsupervised Learning

Unsupervised learning analyses the data to find whether any pattern exists. Here, it is desired to learn how many clusters naturally exist within the DIP feature space. According to Goldberg scoring method, four clusters is desired in which the data points (images) of the same class should be clustered together. Cluster analysis can also be used to investigate whether increasing or decreasing the number of score levels produce better clustering solutions. study detected a relatively large variation of volumetric material loss at stem tapers with score level 4. It rose the question of whether that would be better to break this particular score level into two or even more levels. This matter was investigated here by setting the number of desired clusters to 5 and investigating the accuracy of clustering solutions in which score level 4 is broken into two distinct levels. The results of five clusters are compared with four clusters to determine whether this will improve the accuracy.

This approach was applied to three groups of input dataset to see whether feature selection and transformation by PCA improves the quality of the clusters. The *kmeans* function supports four pairwise distance measures for non-binary data that were used here to see which one provides superior clustering solutions.

Via looking at the silhouette graph, the best distance metric can be identified. This process can turn from graphical into analytical by calculating the average silhouette value for each clustering solution. Unlike the silhouette graph, this metric does not specify which cluster/s have been more problematic which is not a matter of concern. It is desired for average silhouette values to be as close as possible to unity, and the same is true for their average. Therefore, the distance metric with the largest average silhouette value was selected. To avoid falling in local optima, each clustering solution was obtained after ten replicates, and the replicate with the minimum total sums of point-to-centroid distances was selected for each clustering configuration.

The accuracy rates were obtained after comparing the actual scores and the assigned cluster indices for each image. Formulating the accuracy rate is a complex task, There exist several different ways that score levels and clusters can be grouped. While the score levels have order, the clusters do not follow any order which increases the number of possible solutions.

4.4.1.1 *The Original 1038 DIP Features*

The features obtained from section 4.2 were used to group the images into four and five clusters by using the four distance metrics (Table 4-4)

Table 4-4. The average silhouette values for each clustering solution using the 1038 features

k-value	sqeuclidean	cityblock	cosine	correlation
4	0.159	0.128	**0.239**	0.205
5	0.153	0.129	**0.224**	0.203

Cosine distance metric offered the best clustering for both clustering solutions. To evaluate the accuracy of the clusters, first, the corresponding confusion matrices were obtained. The rows represent the score levels, and the columns are associated with the clusters. Table 4-5 shows the confusion matrix for the four clustering solution with cosine distance metric.

Table 4-5. Confusion matrix for four clusterings of 1038 features using cosine distance metric

	C1	C2	C3	C4
S1	87	86	49	142
S2	108	104	62	241
S3	31	47	18	78
S4	11	12	5	23

Since the clusters do not follow any order, 24 cluster-score configurations wherein each score level is associated with just one cluster were established to quantify the accuracy rates. Table 4-6 summarises the accuracy rate for each configuration.

Table 4-6. The accuracies of the 24 cluster-score configurations for four clustering

	1				2				3				4			
	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4
S1	✓				✓				✓				✓			
S2		✓				✓					✓				✓	
S3			✓					✓		✓						✓
S4				✓			✓					✓		✓		
ACCURACY	21.0%				24.8%				19.8%				21.6%			

	5				6				7				8			
	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4
S1	✓				✓					✓				✓		
S2				✓				✓	✓				✓			
S3		✓					✓				✓					✓
S4			✓			✓						✓			✓	
ACCURACY	34.4%				32.4%				21.3%				25.1%			

	9				10				11				12			
	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4
S1		✓				✓				✓				✓		
S2			✓				✓					✓				✓
S3	✓							✓	✓						✓	
S4				✓	✓						✓		✓			
ACCURACY	18.3%				21.5%				32.9%				32.2%			

	13				14				15				16			
	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4
S1			✓				✓				✓				✓	
S2		✓				✓			✓				✓			
S3	✓							✓		✓						✓
S4				✓	✓							✓		✓		
ACCURACY	18.8%				21.9%				20.6%				22.4%			

	17				18				19				20			
	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4
S1			✓				✓					✓				✓
S2				✓				✓		✓				✓		
S3		✓			✓						✓		✓			
S4	✓					✓			✓						✓	
ACCURACY	31.5%				30.2%				24.9%				25.5%			

	21				22				23				24			
	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4
S1				✓				✓				✓				✓
S2			✓				✓		✓				✓			
S3		✓			✓					✓					✓	
S4	✓					✓					✓			✓		
ACCURACY	23.7%				22.4%				27.4%				25.4%			

The accuracy rates range between 18.3% (configuration 9) and 34.4% (configuration 5). These results will be used as the benchmark to investigate whether dividing score level 4 into two distinct score levels improves the accuracy rates. Table 4-7 shows the confusion matrix for five clusters.

Table 4-7. Confusion matrix for five clusterings of 1038 features using cosine distance metric

C1	C2	C3	C4	C5

S1	113	82	38	82	49
S2	184	100	42	128	61
S3	58	46	12	40	18
S4	19	12	5	10	5

Having five clusters and four score levels results in multitudes of distinct score-cluster configurations. Since the focus is on score level 4, only those configurations in which score level 4 is assigned to two clusters were investigated. The result was 24 score-cluster configurations (Table 4-8). Each accuracy rate corresponds to a specific configuration.

Table 4-8. The accuracies of the 24 cluster-score configurations for five clustering

	1					**2**					**3**					**4**				
	C1	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**
S1	✓					✓						✓					✓			
S2		✓						✓			✓							✓		
S3			✓				✓						✓			✓				
S4				✓	✓				✓	✓				✓	✓				✓	✓
ACCURACY	21.7%					19.6%					26.5%					17.8%				
	5					**6**					**7**					**8**				
	C1	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**
S1			✓					✓			✓					✓				
S2	✓						✓					✓								✓
S3		✓				✓									✓		✓			
S4				✓	✓				✓	✓			✓	✓				✓	✓	
ACCURACY	25.6%					19.1%					22.3%					21.3%				
	9					**10**					**11**					**12**				
	C1	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**
S1		✓					✓								✓					✓
S2	✓									✓	✓						✓			
S3					✓	✓						✓				✓				
S4			✓	✓				✓	✓				✓	✓			✓	✓		
ACCURACY	27.1%					19.6%					26.6%					20.1%				
	13					**14**					**15**					**16**				
	C1	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**
S1	✓					✓								✓					✓	
S2				✓						✓	✓									✓
S3					✓				✓						✓	✓				
S4		✓	✓				✓	✓				✓	✓				✓	✓		
ACCURACY	25.0%					20.9%					27.3%					19.7%				
	17					**18**					**19**					**20**				
	C1	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**
S1					✓					✓			✓					✓		
S2				✓		✓								✓						✓
S3	✓								✓						✓				✓	
S4		✓	✓				✓	✓			✓	✓				✓	✓			
ACCURACY	22.8%					26.3%					19.5%					15.4%				
	21					**22**					**23**					**24**				
	C1	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**
S1				✓					✓						✓					✓
S2			✓							✓				✓				✓		
S3					✓			✓					✓						✓	
S4	✓	✓				✓	✓				✓	✓				✓	✓			
ACCURACY	15.7%					16.8%					19.9%					14.7%				

The accuracy rates range between 14.7% (configuration 24) and 27.3% (configuration 15). The top accuracy rate drops from 34.4% to 27.3%. Hence, it is concluded that breaking score level 4 into two score levels will not improve the clustering solutions when the entire features are used.

4.4.1.2 *The First 347 PCs*

From the Pareto chart, the first 347 PCs observed explaining 95% of the data variance. As the input data, they went through the same process as in the previous section. Table 4-9 summarises the clustering results according to the four distance metrics.

Table 4-9. The average silhouette values for each clustering solution using the first 347 PCs

k-value	sqeuclidean	cityblock	cosine	correlation
4	0.174	-0.016	**0.264**	0.263
5	0.171	-0.021	**0.248**	0.248

Similar to the previous dataset, the cosine metric produced the maximum average silhouette values for both clustering solutions. To evaluate the accuracy of the clusters, first, the corresponding confusion matrix was obtained. Table 4-10 shows the confusion matrix for the four clustering solution with cosine distance metric.

Table 4-10. Confusion matrix for four clusterings of 347 PCS using cosine distance metric

	C1	C2	C3	C4
S1	72	114	45	133
S2	101	202	74	138
S3	49	64	20	41
S4	23	10	8	10

Table 4-11 summarises the accuracy rate for the 24 cluster-score configurations.

Table 4-11. The accuracies of the 24 cluster-score configurations for four clustering

	1				**2**				**3**				**4**			
	C1	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**
S1	✓				✓				✓				✓			
S2		✓				✓					✓				✓	
S3			✓					✓		✓						✓
S4				✓			✓					✓		✓		
ACCURACY	27.5%				29.3%				19.9%				17.8%			
	5				**6**				**7**				**8**			
	C1	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**
S1	✓				✓					✓				✓		
S2				✓				✓	✓				✓			
S3		✓					✓				✓					✓
S4			✓			✓						✓			✓	
ACCURACY	25.5%				21.7%				22.2%				23.9%			
	9				**10**				**11**				**12**			
	C1	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**
S1		✓				✓				✓				✓		
S2			✓				✓					✓				✓
S3	✓							✓	✓						✓	
S4				✓	✓						✓		✓			
ACCURACY	22.4%				22.8%				28.0%				26.7%			
	13				**14**				**15**				**16**			
	C1	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**
S1			✓				✓				✓				✓	
S2		✓				✓			✓				✓			
S3	✓							✓		✓						✓
S4				✓	✓							✓		✓		
ACCURACY	27.7%				28.2%				19.9%				17.8%			
	17				**18**				**19**				**20**			
	C1	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**
S1			✓				✓					✓				✓
S2				✓				✓		✓				✓		
S3		✓			✓						✓		✓			
S4	✓					✓			✓						✓	
ACCURACY	24.5%				21.9%				34.2%				35.5%			
	21				**22**				**23**				**24**			
	C1	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**
S1				✓				✓				✓				✓
S2			✓				✓		✓				✓			
S3		✓			✓					✓					✓	
S4	✓					✓					✓			✓		
ACCURACY	26.6%				24.1%				27.7%				23.9%			

The accuracy rates range between 17.8% (configurations 4 and 16) and 34.2% (configuration 19). These results will be used as the benchmark to investigate whether dividing score level 2 into two distinct score levels improves the accuracy rates. Table 4-12 shows the confusion matrix for five clusters.

Table 4-12. Confusion matrix for four clusterings of 347 PCS using cosine distance metric

	C1	C2	C3	C4	C5
S1	119	81	44	49	71
S2	123	132	70	94	96
S3	40	45	20	20	49
S4	9	6	8	5	23

The 24 score-cluster configurations wherein score level 4 is assigned to two distinct clusters were evaluated to identify the best configuration (Table 4-13).

Table 4-13. The accuracies of the 24 cluster-score configurations for five clustering

	1					**2**					**3**					**4**				
	C1	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**
S1	✓					✓						✓					✓			
S2		✓						✓			✓							✓		
S3			✓				✓						✓			✓				
S4				✓	✓				✓	✓				✓	✓				✓	✓
ACCURACY	27.1%					23.7%					22.8%					19.8%				
	5					**6**					**7**					**8**				
	C1	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**
S1			✓					✓			✓					✓				
S2	✓						✓					✓								✓
S3		✓				✓									✓		✓			
S4				✓	✓				✓	✓			✓	✓				✓	✓	
ACCURACY	21.7%					22.1%					**28.4%**					24.7%				
	9					**10**					**11**					**12**				
	C1	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**
S1		✓					✓								✓					✓
S2	✓									✓	✓						✓			
S3					✓	✓						✓				✓				
S4			✓	✓				✓	✓				✓	✓			✓	✓		
ACCURACY	24.1%					20.8%					22.8%					23.2%				
	13					**14**					**15**					**16**				
	C1	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**
S1	✓					✓								✓					✓	
S2				✓						✓	✓									✓
S3					✓				✓						✓	✓				
S4		✓	✓				✓	✓				✓	✓				✓	✓		
ACCURACY	25.0%					22.6%					21.3%					18.0%				
	17					**18**					**19**					**20**				
	C1	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**
S1					✓					✓			✓					✓		
S2				✓		✓								✓						✓
S3	✓								✓						✓				✓	
S4		✓	✓				✓	✓			✓	✓				✓	✓			
ACCURACY	19.8%					20.7%					18.3%					15.9%				
	21					**22**					**23**					**24**				
	C1	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**
S1				✓					✓						✓					✓
S2			✓							✓				✓				✓		
S3					✓			✓					✓						✓	
S4	✓	✓				✓	✓				✓	✓				✓	✓			
ACCURACY	16.6%					**16.3%**					18.1%					15.9%				

The accuracy rates range between 16.3% (configuration 22) and 28.4% (configuration 7). The top accuracy rate has dropped from 34.2% to 28.4%. Hence, it is concluded that breaking score level 4 into two score levels will not improve the clustering solutions when the first 347 PCs are used.

4.4.1.3 *The Selected 307 DIP Features*

Based on the 347 PCs from the previous section, 307 DIP features were selected due to showing relatively higher correlations with the rotated components. Table 4-14 summarises the clustering results of this set of data.

Table 4-14. The average silhouette values for each clustering solution using the selected 307 features

k-value	sqeuclidean	cityblock	cosine	correlation
4	0.087	0.082	**0.184**	0.164
5	0.088	0.084	**0.171**	0.162

Again, the cosine metric produced the maximum average silhouette values for both clustering solutions. To evaluate the accuracy of the clusters, first, the corresponding confusion matrix was obtained. Table 4-15 shows the confusion matrix for the four clustering solution with this distance metric.

Table 4-15. confusion matrix for four clusterings of 307 features using cosine distance metric

	C1	C2	C3	C4
S1	113	42	132	77
S2	206	70	137	102
S3	61	21	45	47
S4	10	7	10	24

Table 4-16 summarises the accuracy rate for the 24 cluster-score configurations.

Table 4-16. The accuracies of the 24 cluster-score configurations for four clustering

	1				**2**				**3**				**4**			
	C1	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**
S1	✓				✓				✓				✓			
S2		✓				✓					✓				✓	
S3			✓					✓		✓						✓
S4				✓			✓					✓		✓		
ACCURACY	22.8%				21.7%				26.7%				27.5%			
	5				**6**				**7**				**8**			
	C1	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**
S1	✓				✓					✓				✓		
S2				✓				✓	✓				✓			
S3		✓					✓				✓					✓
S4			✓			✓						✓			✓	

ACCURACY	22.3%				24.2%				28.7%				27.6%			
	9				10				11				12			
	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4
S1		✓				✓				✓				✓		
S2			✓				✓					✓				✓
S3	✓							✓	✓						✓	
S4				✓	✓						✓		✓			
ACCURACY	23.9%				21.4%				19.5%				18.0%			
	13				14				15				16			
	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4
S1			✓				✓				✓				✓	
S2		✓				✓			✓				✓			
S3	✓							✓		✓						✓
S4				✓	✓							✓		✓		
ACCURACY	26.0%				23.5%				34.7%				35.5%			
	17				18				19				20			
	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4
S1			✓				✓					✓				✓
S2				✓				✓		✓				✓		
S3		✓			✓						✓		✓			
S4	✓					✓			✓						✓	
ACCURACY	24.0%				27.4%				18.3%				19.7%			
	21				22				23				24			
	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4
S1				✓				✓				✓				✓
S2			✓				✓		✓				✓			
S3		✓			✓					✓					✓	
S4	✓					✓					✓			✓		
ACCURACY	22.2%				25.5%				28.4%				30.3%			

The accuracy rates range between 18.0% (configuration 12) and 35.5% (configuration 16). Table 4-17 shows the confusion matrix for five clusters.

Table 4-17. Confusion matrix for four clusterings of 307 features using cosine distance metric

	C1	C2	C3	C4	C5
S1	104	90	54	40	76
S2	106	144	100	67	98
S3	37	48	25	20	44
S4	8	6	8	7	22

The 24 score-cluster configurations wherein score level 4 is assigned to two distinct clusters were evaluated to identify the best configuration (Table 4-18).

Table 4-18. The accuracies of the 24 cluster-score configurations for five clustering

	1					2					3					4				
	C1	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**
S1	✓					✓						✓					✓			
S2		✓						✓			✓							✓		
S3			✓				✓						✓			✓				
S4				✓	✓				✓	✓				✓	✓				✓	✓
ACCURACY	27.1%					23.7%					22.8%					19.8%				
	5					**6**					**7**					**8**				
	C1	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**
S1			✓					✓			✓					✓				
S2	✓						✓					✓								✓
S3		✓				✓									✓		✓			
S4				✓	✓				✓	✓			✓	✓				✓	✓	
ACCURACY	21.7%					22.1%					**28.4%**					24.7%				
	9					**10**					**11**					**12**				
	C1	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**
S1		✓					✓								✓					✓
S2	✓									✓	✓						✓			
S3					✓	✓						✓				✓				
S4			✓	✓				✓	✓				✓	✓			✓	✓		
ACCURACY	24.1%					20.8%					22.8%					23.2%				
	13					**14**					**15**					**16**				
	C1	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**
S1	✓					✓								✓					✓	
S2				✓						✓	✓									✓
S3					✓				✓						✓	✓				
S4		✓	✓				✓	✓				✓	✓				✓	✓		
ACCURACY	25.0%					22.6%					21.3%					18.0%				
	17					**18**					**19**					**20**				
	C1	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**
S1					✓					✓			✓					✓		
S2				✓		✓								✓						✓
S3	✓								✓						✓				✓	
S4		✓	✓				✓	✓			✓	✓				✓	✓			
ACCURACY	19.8%					20.7%					18.3%					**15.9%**				
	21					**22**					**23**					**24**				
	C1	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**
S1				✓					✓						✓					✓
S2			✓							✓				✓				✓		
S3					✓			✓					✓						✓	
S4	✓	✓				✓	✓				✓	✓				✓	✓			
ACCURACY	16.6%					16.3%					18.1%					**15.9%**				

The accuracy rates range between 16.3% (configuration 20 and 24) and 28.4% (configuration 7). The top accuracy rate has dropped from 35.5% to 28.4%. Hence, it is concluded that breaking score level 4 into two score levels will not improve the clustering solutions when the selected 307 DIP features are used.

4.4.1.4 *The Second Round of Feature Selection*

Achieving relatively better solutions after conducting feature selection by reducing the dimensionality of the feature space from 1038 to 307 inspired one further replication of feature selection. The aim was to see whether a feature space with a dimension below that can achieve the same or even better clustering of the images.

From the PCA of these 307 features, the Pareto chart of the resultant 307 PCs was obtained. Figure 4-5 includes just the first 10 PCs. Similar to the Pareto chart of the PCs obtained from the original 1038 DIP features, this new set of PCs are not explaining high portions of data variability. It takes 186 PCs to explain more than 95% of this new dataset variance.

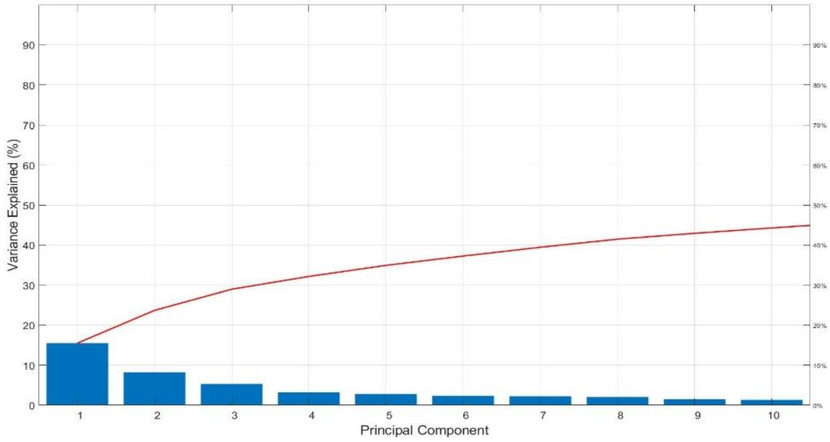


Figure 4-5. The Pareto chart for the first 10 PCs extracted from the selected 307 features

Using these PCs, a matrix of loadings with 307 rows (features) and 186 columns (rotated PCs) was established to discard variables with no higher loading at any of the 186 PCs. This second round of feature selection also used the conventional value of 0.3 as the threshold. The result was a selection of 177 out of 307 features wherein 19 and 158 features belong to global (Table 4-19) and local types.

Table 4-19. The global features having higher correlation with the 186 PCs

text_stat_Hue_1	cooc_contrast_2	cooc_contrast_3	cooc_contrast_4
cooc_correlation_1	cooc_correlation_2	cooc_energy_1	cooc_energy_4
cooc_homogeneity_1	cooc_homogeneity_2	cooc_homogeneity_4	cooc_entropy_2
cooc_entropy_3	cooc_entropy_4	E_global_1	E_global_3
E_local_1	E_local_2	E_local_3	

Table 4-20 summarises the clustering results of this subset of feature space.

Table 4-20. The average silhouette values for each clustering solution using the selected 177 features

k-value	sqeuclidean	cityblock	cosine	correlation
4	0.064	0.066	**0.153**	0.148
5	0.067	0.061	**0.148**	0.140

As in the other three input datasets, the cosine metric produced the maximum average silhouette values for both clustering solutions. To evaluate the accuracy of the clusters, first, the corresponding confusion matrix was obtained. Table 4-21 shows the confusion matrix for the four clustering solution with this distance metric and Table 4-22 summarises the accuracy rate for the 24 cluster-score configurations.

Table 4-21. Confusion matrix for four clusterings of 177 features using cosine distance metric

	C1	C2	C3	C4
S1	137	106	74	47
S2	142	197	104	72
S3	48	60	46	20
S4	11	9	24	7

Table 4-22. The accuracies of the 24 cluster-score configurations for four clustering

	1				**2**				**3**				**4**			
	C1	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**
S1	✓				✓				✓				✓			
S2		✓				✓					✓				✓	
S3			✓					✓		✓						✓
S4				✓			✓					✓		✓		
ACCURACY	35.1%				34.2%				27.9%				24.5%			
	5				**6**				**7**				**8**			
	C1	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**
S1	✓				✓					✓				✓		
S2				✓				✓	✓				✓			
S3		✓					✓				✓					✓
S4			✓			✓						✓			✓	
ACCURACY	26.5%				23.9%				27.3%				26.4%			
	9				**10**				**11**				**12**			
	C1	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**
S1		✓				✓				✓				✓		
S2			✓				✓					✓				✓
S3	✓							✓	✓						✓	
S4				✓	✓						✓		✓			
ACCURACY	24.0%				21.8%				22.6%				21.3%			
	13				**14**				**15**				**16**			
	C1	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**
S1			✓				✓				✓				✓	
S2		✓				✓			✓				✓			
S3	✓							✓		✓						✓
S4				✓	✓							✓		✓		
ACCURACY	29.5%				27.4%				25.6%				22.2%			
	17				**18**				**19**				**20**			
	C1	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**
S1			✓				✓					✓				✓
S2				✓				✓		✓				✓		
S3		✓			✓						✓		✓			
S4	✓					✓			✓						✓	
ACCURACY	19.7%				18.4%				27.3%				28.6%			
	21				**22**				**23**				**24**			
	C1	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**
S1				✓				✓				✓				✓
S2			✓				✓		✓				✓			
S3		✓			✓					✓					✓	
S4	✓					✓					✓			✓		
ACCURACY	20.1%				18.8%				24.7%				22.1%			

The accuracy of the configurations span from 18.4% (configuration 15) to 35.1% (configuration 3). Table 4-23 shows the confusion matrix for five clusters.

Table 4-23. Confusion matrix for four clusterings of 177 features using cosine distance metric

	C1	C2	C3	C4	C5
S1	72	100	45	48	99
S2	100	108	65	93	149
S3	44	40	20	24	46
S4	21	9	7	7	7

The 24 score-cluster configurations wherein score level 4 is assigned to two distinct clusters were evaluated to identify the best configuration (Table 4-24).

Table 4-24. The accuracies of the 24 cluster-score configurations for five clustering

	1					**2**					**3**					**4**				
	C1	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**
S1	✓					✓						✓					✓			
S2		✓						✓			✓							✓		
S3			✓				✓						✓			✓				
S4				✓	✓				✓	✓				✓	✓				✓	✓
ACCURACY	19.4%					17.3%					21.2%					20.2%				
	5					**6**					**7**					**8**				
	C1	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**
S1			✓					✓			✓					✓				
S2	✓						✓					✓								✓
S3		✓				✓									✓		✓			
S4				✓	✓				✓	✓			✓	✓				✓	✓	
ACCURACY	18.0%					19.1%					21.7%					**24.9%**				
	9					**10**					**11**					**12**				
	C1	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**
S1		✓					✓								✓					✓
S2	✓									✓	✓						✓			
S3					✓	✓						✓				✓				
S4			✓	✓				✓	✓				✓	✓			✓	✓		
ACCURACY	23.6%					27.8%					22.9%					24.0%				
	13					**14**					**15**					**16**				
	C1	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**
S1	✓					✓								✓					✓	
S2				✓						✓	✓									✓
S3					✓				✓						✓	✓				
S4		✓	✓				✓	✓				✓	✓				✓	✓		
ACCURACY	20.6%					23.6%					19.0%					23.3%				
	17					**18**					**19**					**20**				
	C1	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**
S1					✓					✓			✓					✓		
S2				✓		✓								✓						✓
S3	✓								✓						✓				✓	
S4		✓	✓				✓	✓			✓	✓				✓	✓			
ACCURACY	22.8%					21.6%					19.4%					22.5%				
	21					**22**					**23**					**24**				
	C1	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**
S1				✓					✓						✓					✓
S2			✓							✓				✓				✓		
S3					✓			✓					✓						✓	
S4	✓	✓				✓	✓				✓	✓				✓	✓			
ACCURACY	**17.1%**					22.4%					21.9%					19.7%				

The accuracy rates range between 17.1% (configuration 21) and 24.9% (configuration 8). The top accuracy rate has dropped from 35.1% to 24.9%. Hence, it is concluded that breaking score level 4 into two score levels will not improve the clustering solutions when the 177 selected DIP features are used. Table 4-25 summarises the top accuracy rates of 4 and 5 clustering solutions. By comparing the result of using 307 features with 177 features, it can be concluded that the second feature selection does not influence the quality of clustering. The lower dimensionality of the feature space which simplifies the cluster analysis justifies performing the second round of feature selection.

Table 4-25. The summary of the highest achieved accuracy rates via cluster analysis

k	1038 features	347 PCs	307 features	177 features
4	34.4%	34.2%	35.5%	35.1%
5	27.3%	28.4%	28.4%	24.9%

Generally, this method of clustering does not recommend breaking score level 4 into two distinct levels. Also, the accuracy rates across the four input datasets are not significantly different. Therefore, using the 177 features which reduces the dimensionality is recommended. The feature selection process can be further replicated and based on the observed quality of cluster-score configurations, it can be determined when to stop this process.

4.4.2 Supervised Learning

From NCA, 25 DIP features were identified as potentially superior for supervised classification of images. This section verifies this matter by using the entire features as well as this subset throughout the supervised learning.

4.4.2.1 *algorithm selection*

Via Classification Learner app in MATLAB, 22 classifications models were trained and fivefold cross-validated to compare their performance. They belong to five groups of classifiers that are comprised of Decision Trees, Discriminant Analysis, SVM, Nearest Neighbour Classifiers, and Ensemble Classifiers. Table 4-26 summarises the performance of the top three accurate algorithms with and without validation schemes using the 1038 features.

Table 4-26. Supervised learning outcomes of the original 1038 DIP features

No Validation (accuracy rate)	5-fold Cross-validation (accuracy rate)
Linear Discriminant (100%)	Support Vector Machine (80.8%)
k-Nearest Neighbour (100%)	Decision Tree (72.8%)
Support Vector Machine (99.3%)	k-Nearest Neighbour (65.5%)

The very high accuracy of no validation scheme raised overfitting concerns which were addressed via fivefold cross-validation to ensure the classification models are generalisable to unseen new images. This procedure was replicated for the 25 selected features to assess their discriminatory power. Table 4-27 summarises the corresponding results.

Table 4-27. Supervised learning outcomes of the selected 25 DIP features

No Validation (accuracy rate)	5-fold Cross-validation (accuracy rate)
k-Nearest Neighbour (100%)	Support Vector Machine (81.8%)
Support Vector Machine (99.4%)	k-Nearest Neighbour (76.0%)
Decision Tree (93.8%)	Decision Tree (75.4%)

The results confirm that SVM outperforms other supervised learning techniques as reported by the literature regardless of using the entire features or the selected features. Since dimensionality reduction simplified the classification problem and observed not to impact the accuracy rate, NCA approved to be a suitable feature selection technique for this dataset. Therefore, the selected features were used by an SVM classifier in this study.

4.4.2.2 *the SVM algorithm*

The script of the SVM model that had returned a relatively high cross-validation accuracy was extracted for amendments to further improve its performance. Accordingly, two hyperparameters, namely Box-Constraint and Kernel-Scale, were optimised. The process attempts to minimise the cross-validation error by varying the hyperparameters.

In MATLAB, the default values for Box-Constraint and Kernel-Scale hyperparameters are 1. However, the extracted script from Classification Learner App showed that Box-Constraint is set equal to 1, while MATLAB selects an appropriate scale factor using a heuristic procedure for Kernel-Scale. This heuristic procedure uses subsampling so that estimates can vary from one call to another. Therefore, the initial value for both of them was set equal to 1 before commencing hyperparameter optimisation.

Using these default values and no optimisation, '*fitcecoc*' was observed to return an unrealistically high (99.4%) accuracy with no validation scheme. To address the overfitting concerns, fivefold cross-validation accuracy was conducted, and the result was an accuracy rate of 60% approximately. Therefore, Bayesian optimisation was deployed iteratively (n = 30) by varying these two hyperparameters. The algorithm attempted to minimise the fivefold cross-validation accuracy by finding suitable values for the hyperparameters. The result improved the cross-validation accuracy from 60% to around 85%.

By updating the hyperparameters across all the binary learners, a no validation accuracy of 92% was achieved. Table 4-28 is the confusion matrix which quantifies the number as well as the percentage of the misclassified images.

Table 4-28. Confusion matrix for the entire 1104 images.

		Predicted Scores				Image set quantity
		1	2	3	4	
Actual Scores	1	338 (92.9%)	26 (7.1%)	0	0	364
	2	33 (6.4%)	469 (91.1%)	13 (2.5%)	0	515
	3	0	8 (4.6%)	162 (93.1%)	4 (2.3%)	174
	4	0	0	3 (5.9%)	48 (94.1%)	51

In this confusion matrix, the diagonal elements represent the quantity of the images that were correctly classified according to their actual Goldberg score while the rest of the elements are associated with the misclassified images. The difference between the actual and predicted scores never exceeds unity in this table.

4.5 Summary and Discussion of the Results

In this chapter, principles of computer vision, digital image processing, and machine learning were used to investigate the applicability of the well-established methods to rate the severity of corrosion at stem tapers of total hip replacement implants. Several groups of DIP features were extracted and went through feature selection and transformation to potentially improve the performance of the subsequent classification algorithms (Figure 4-6).

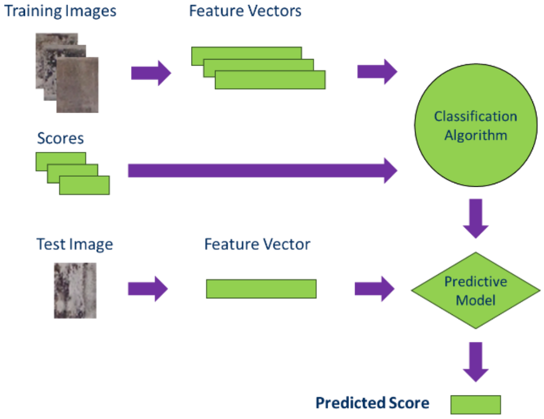


Figure 4-6. An overview of the objective corrosion rating pipeline

MATLAB was used as the environment to design and implement the algorithms. Two toolboxes of Image Processing & Computer Vision and Statistics & Machine Learning provides several functions and workflows to implement these tasks. The abundance of learning materials facilitated taking advantage of the available capabilities of MATLAB in a high capacity.

The performance of the classification phase is subject to the quality of the acquired images. Hence, throughout the photography phase, it was crucial to ensure that the images are captured with a reasonable level of quality. To fulfil this task, it is required to minimise affine distortions (i.e. translation, rotation, and scale transformations) and photometric deformations induced by scene illumination, 3D camera viewpoint, background clutter, and occlusion. The global features are not usually invariant to such imaging deformations. Out of the three groups of global features that were used in this study, only the wavelet transform provided robustness to scale transformations. While local features (e.g. SURF) are usually invariant to many of these deformations, they can be computationally expensive.

It should be noted that the images used in this study were captured in a laboratory environment under a strict lighting-optics-viewing condition to minimise these unfavourable conditions. Therefore, the main reason for considering local features in this study was to produce large numbers of numerical features which is another attribute of local features. This can provide the successive classification stage with more characterising information about the images. Local features have been introduced more recently in comparison to their global counterparts, and therefore, only a few studies have used them in texture analysis of corrosion defects to date [71].

In visual scoring of damage at stem tapers, it is routine to rely on microscopes in addition to the naked eye to magnify an area of interest and ensure the visual features are associated with corrosion and not anything else (e.g. biological debris). Despite the best efforts, it was not always possible to thoroughly remove dirt from stem tapers. The resolution of the camera was not in a top range, and the images were captured from 25 cm away outside the surrounding shade. Therefore, there is a possibility of mixing up corrosion with dirt. Also, using a professional camera with a long-reach macro lens can address this matter to some extent.

Rating corrosion severity via image-based analysis in large-scale retrieval studies resolves implant accessibility issues and the images can be shared and used by anyone and anywhere in future. The sensitivity of the SVM classifier to image sets obtained under different lighting-optics-viewing conditions need to be assessed to ensure reliable long-term performance.

This chapter elaborated on 1038 DIP features that were extracted from 1104 labelled images that belong to four classes (visual scores). One of the key decision to make is to extract what features from the images. Rating corrosion damage within this context was more challenging in comparison with many similar studies that have addressed corrosion as rust [67, 68, 70]. The morphological attributes that can be used in characterising corrosion at taper junctions confine to just texture while both colour and texture play a prominent role in characterising rust. This fact was verified through the features selection phase in this research. Out of the 12 extracted first-order image statistics that characterise colour, only one (text_stat_Sat_6) was picked as suitable to be used in the subsequent machine learning stage.

The global and local features that have been formulated for texture analysis are numerous. Since the novelty of this project was not about introducing new features, global and local features that have outperformed the others more frequently were used in this study. Nevertheless, there is always room for improvement by formulating superior texture features. This matter is looked upon as one disadvantage of many machine learning techniques against Convolutional Neural Networks (CNN) which has eliminated the need for image feature extraction. CNN is a group of machine learning algorithms with an architecture that can be divided into two parts. The first part receives the images as the input and takes advantage of a range of convolutional filters which enable them to adaptively learn and formulate a feature extractor. The second part is a fully connected neural network that classifies the feature vectors [146]. Although feature extraction and classification are integrated, CNN demands thousands of labelled images in each class to be properly trained. Since it is not practical in many problems, the other machine learning techniques have been more feasible to implement so far.

Unsupervised and supervised feature transformation and selection were observed to positively and sometimes adversely influence the performance of the subsequent machine learning algorithms. The matrix of features (1104 × 1038) was not too large in dimensions. Therefore, the computation time was not a matter of concern in this study. Yet, dimensionality reduction via supervised feature selection and transformation observed to yield approximately the same level of classification accuracy as when the entire features are used.

Two unsupervised and supervised dimensionality reduction techniques were deployed. Their efficacy levels were only determined after deploying their outcomes in the classification algorithms. As for the unsupervised dimensionality reduction, PCA was used for two rounds of feature selection. The second round observed achieving similar results to those of the first round. Considering the lower dimensionality of the feature space which in turn reduces the computation cost, it was recommended to continue this process until the accuracy of the obtained cluster-score configurations drops.

As for the supervised dimensionality reduction, NCA was observed to be effective in reducing the dimensionality of the feature vectors to 25 without adversely affecting the accuracy rate of the SVM algorithm. NCA determined which global features are better at the prediction of scores which can be used as a guideline in future studies to avoid spending time in the calculation of less relevant features.

The nature of this classification problem demands supervised learning. The number of classes is already known, and the images had been already labelled. The reason for incorporating unsupervised learning was to see whether the extracted feature vectors are better classified in five clusters if score level 4 breaks into two levels. The average silhouette value and the maximum accuracy rate dropped after increasing the number of clusters from 4 to 5 which indicates that by using the 1038 DIP features and kmeans clustering, still no improvement is achieved following the increase in the number of the score levels. Therefore, breaking score level 4 is not recommended. Generally, increasing the number of classes makes classification problems more complex which may adversely affect the accuracy rates. One way to improve the classification accuracy is by reducing the number of classes by combining them. It seems this problem abides the same rule.

The supervised algorithm introduced in this study endeavours to find patterns in data according to the scoring method that has been deployed to rate the corrosion severity. The algorithm does not surpass human expertise in evaluating damage severity. Therefore, the classification error entirely stems from the classification model, not the benchmark visual scores. This argument is based on the outcome of a previous study [31] that had identified substantial inter-observer reproducibility and single-observer repeatability of visual scoring techniques.

Several supervised classification algorithms were compared. As expected from the findings of the literature, SVM observed to be more robust to cross-validation. NCA features did not affect the SVM performance, so it was adopted as the feature selection tool to reduce the dimensionality of feature space. Optimising the two hyperparameters of the SVM algorithm improved the accuracy from about 60% to 85%. Taking into account the characteristics of this particular classification problem, the cross-validation accuracy achieved in this study sits well in the range of its counterpart reported by similar studies on other contexts of rating corrosion severity [59, 60, 73].

Generally, implementing supervised learning turned out to be a straightforward process that more accurately predicts the scores. Although SVM is highly more complicated than k-means clustering, MATLAB and similar software packages facilitate development, implementation, and tuning such complex techniques.

5 CORROSION ACROSS STEM TAPER ZONES[2]

[Image removed due to copyright restriction]

[2] *This chapter is based on the work that has been published in Metals; Milimonfared R, Oskouei RH, Taylor M, Solomon LB. The distribution and severity of corrosion damage at eight distinct zones of metallic femoral stem implants. Metals, 8(10) (2018) 840.*

5.1 Introduction

As mentioned in chapter three, the first set of scores measures corrosion semi-quantitatively at eight distinct zones that correspond to the Posterior-Distal (PD), Posterior-Proximal (PP), Medial-Distal (MD), Medial-Proximal (MP), Anterior-Distal (AD), Anterior-Proximal (AP), Lateral-Distal (LD), and Lateral-Proximal (LP). These scores were used to compare the frequency of score levels at each zone as well as the severity of damage across the eight zones.

5.2 Distribution of Scores

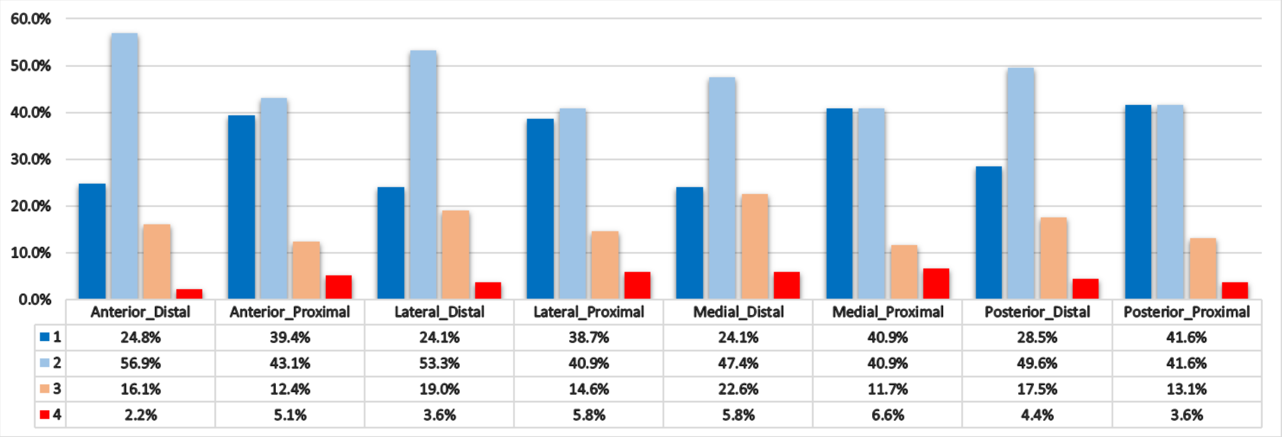
Weighted kappa (κ_W) with quadratic weights was run to determine the single-observer repeatability of the corrosion scores. A confusion matrix was established to quantify the disagreements. For quadratic weights, the further away from a disagreement from the perfect agreement, the more harshly that disagreement is considered. Weighted kappa (κ_W) with quadratic weights indicated a statistically significant agreement, $\kappa_W = 0.64$ (95% CI, 0.59 to 0.69), $p < 0.001$ between the two sets of scores. According to [147], the strength of the agreement was classified as good.

Visual scoring of the 137 stem trunnions across eight zones resulted in 1096 corrosion scores. Table 5-1 Table 1-1summarises the quantity and percentage of each score level. Score level 2 had the highest quantity (512) while the lowest quantity (51) belonged to score level 4.

Table 5-1. The quantity of the zones having each score level.

Score	Quantity (%)
1	359 (32.8)
2	512 (46.7)
3	174 (15.9)
4	51 (4.7)

Figure 1-1 compares the frequency of each score level over the entire eight zones. The values in this figure are portions of 137 (the scores assigned at each zone), and adding up the values in each column (not row) should give 100%. Score levels 1 through 4 stood in the first place at zones Posterior_Proximal, Anterior_Distal, Medial_Distal, and Medial_Proximal respectively.

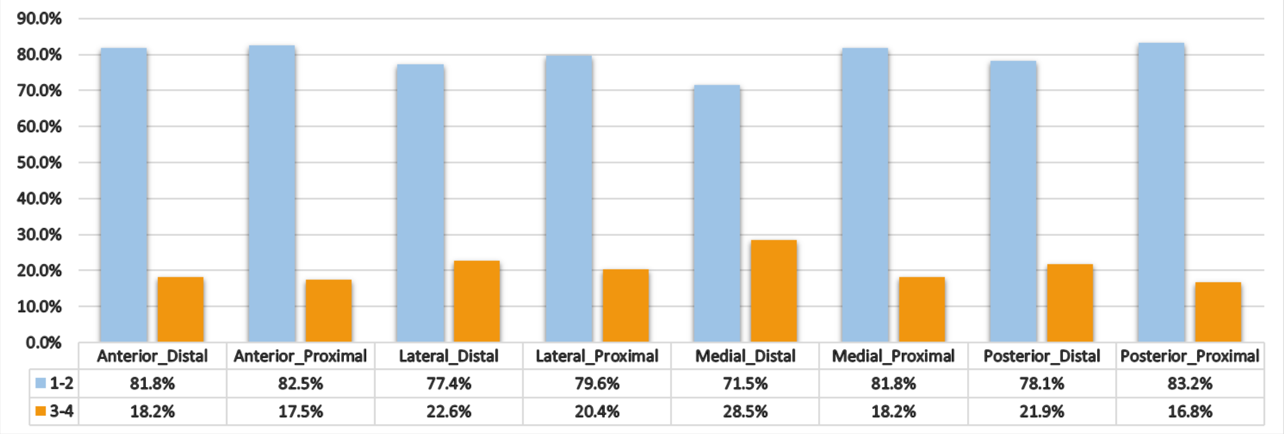


	Anterior_Distal	Anterior_Proximal	Lateral_Distal	Lateral_Proximal	Medial_Distal	Medial_Proximal	Posterior_Distal	Posterior_Proximal
1	24.8%	39.4%	24.1%	38.7%	24.1%	40.9%	28.5%	41.6%
2	56.9%	43.1%	53.3%	40.9%	47.4%	40.9%	49.6%	41.6%
3	16.1%	12.4%	19.0%	14.6%	22.6%	11.7%	17.5%	13.1%
4	2.2%	5.1%	3.6%	5.8%	5.8%	6.6%	4.4%	3.6%

Figure 5-1. Distribution of corrosion score levels across the entire eight stem taper zones of 137 retrieved implants.

Considering the unbalanced score levels, the first two score levels that are higher in quantity (i.e. 359 and 512) always show higher percentages compared with score levels 3 and 4 within each zone.

To better compare the severity of damage across the zones, two more configurations of scores (by combining the original score levels) were also explored. The first configuration groups the first and the last two score levels into low and high groups, respectively. Figure 5-2 visualizes this configuration and compares each score group across the eight zones.



	Anterior_Distal	Anterior_Proximal	Lateral_Distal	Lateral_Proximal	Medial_Distal	Medial_Proximal	Posterior_Distal	Posterior_Proximal
1-2	81.8%	82.5%	77.4%	79.6%	71.5%	81.8%	78.1%	83.2%
3-4	18.2%	17.5%	22.6%	20.4%	28.5%	18.2%	21.9%	16.8%

Figure 5-2. The quantity of the double score levels at each zone (scores 1 and 2 versus scores 3 and 4 combined).

As expected, the low score group which comprises 871 (359 + 512) scores has a higher frequency compared with the high score group (174 + 51). This configuration can better show which zones have more severe corrosion damage (for example, MD and LD zones). Also, at zones, MD and PP, the smallest and largest gaps between these two combined score levels were observed.

The third configuration preserves score level 1 and combines the other three score levels to form two new score groups of intact and corroded stem tapers. Figure 5-3 illustrates the frequencies of these two score groups.

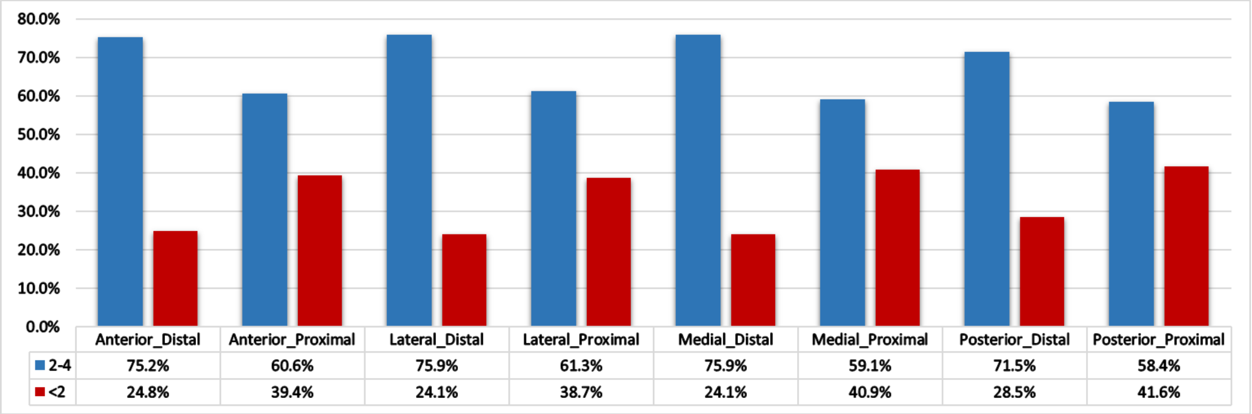


Figure 5-3. Distribution of corroded stem tapers against the intact group.

The medial distal zone had the largest difference between these two score groups which confirms that this particular zone is most damaged. Also, the posterior-proximal zone had the smallest difference between the two score groups (thus least damaged). As a key finding, the distal regions of the four quadrants showed more corrosion damage compared with the proximal regions.

These finding from the histogram can shed light on the likely outcome of the OLR model. Especially when the number of DV levels are higher, cumulative logits models may become infeasible. Histograms can determine which score levels are more important to be compared via using other types of OLR models such as adjacent categories.

5.3 Comparison of Corrosion in the Zones

Cumulative odds OLR with proportional odds was employed to conduct a pairwise comparison of the zones. First, it was established whether zone is statistically significant overall. By referring to the *Test of Model Effects* table at SPSS output report, zone was observed to be a significant (p = 0.002) predictor of corrosion scores in a univariate regression.

Since no specific zone was preferential to investigate, 28 pairwise comparisons had to be undertaken which incurred additional calculations to obtain the overall omnibus statistical test. Table 5-1 summarises the OR, p-values, and confidence intervals. Significant OR values are highlighted in grey. In this table, each zone has been used seven times either as the primary or reference (inside brackets) group to exhaust the combinations. OR values below 1 indicate that for the primary category, the odds of having a higher corrosion score is lower than that of the reference category.

For instance, the odds of receiving a higher score at Anterior_Proximal zone is 0.59 times that of Lateral_Distal zone. Since this number is below one, that implies the odds of observing more corrosion at Anterior_Proximal is 0.41 times lower than that of Lateral_Distal.

Table 5-2. The odds of observing a higher corrosion score at a primary zone compared with a reference

ZONE PAIR	OR	p-value	CI (*p* < 0.05)	
AD (AP)	1.493	0.077	0.957	2.329
AD (LD)	0.882	0.577	0.566	1.372
AD (LP)	1.365	0.169	0.876	2.128
AD (MD)	0.755	0.212	0.485	1.175
AD (MP)	1.524	0.063	0.977	2.378
AD (PD)	0.998	0.993	0.641	1.554
AD (PP)	1.634	0.031	1.047	2.551
AP (LD)	0.590	0.020	0.379	0.921
AP (LP)	0.914	0.693	0.586	1.427
AP (MD)	0.505	0.003	0.324	0.789
AP (MP)	1.021	0.928	0.654	1.594
AP (PD)	0.668	0.076	0.429	1.043
AP (PP)	1.094	0.692	0.701	1.709
LD (LP)	1.549	0.054	0.993	2.414
LD (MD)	0.856	0.490	0.550	1.331
LD (MP)	1.729	0.016	1.108	2.697
LD (PD)	1.132	0.583	0.727	1.762
LD (PP)	1.853	0.007	1.187	2.894
LP (MD)	0.553	0.009	0.355	0.862
LP (MP)	1.116	0.628	0.715	1.742
LP (PD)	0.731	0.167	0.469	1.140
LP (PP)	1.197	0.429	0.767	1.869
MD (MP)	2.019	0.002	1.294	3.152
MD (PD)	1.322	0.216	0.850	2.058
MD (PP)	2.165	0.001	1.386	3.382
MP (PD)	0.655	0.062	0.420	1.022
MP (PP)	1.072	0.760	0.686	1.675
PD (PP)	1.637	0.030	1.049	2.556

The reciprocal of odds ratios can be calculated to compare a reference group with a primary group. To compare the severity of corrosion across the entire eight zones, the odds ratios where sorted and plotted (Figure 5-4). The red and blue bars indicate significant and insignificant OR values, respectively. An OR equal to 1 indicates equal odds of observing a higher corrosion score at the primary and reference zone groups. By moving away from unity, the odds ratios are first insignificant which later on become significant. The speed by which this transition takes place is a function of the presumed statistical significance level.

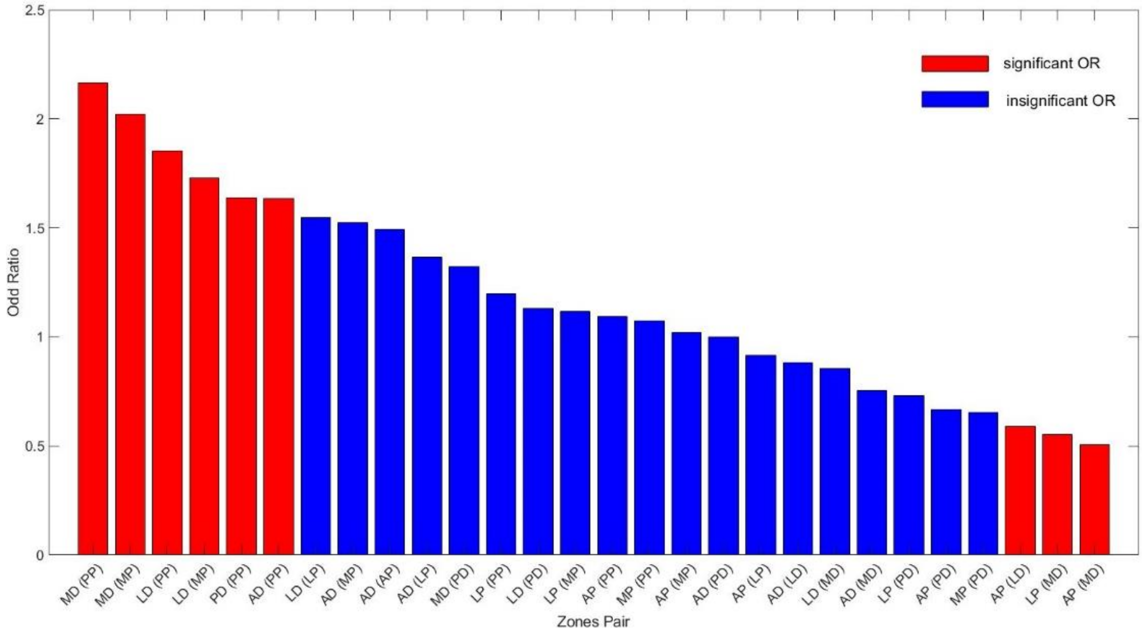


Figure 5-4. The 28 odds ratios sorted and colour-coded for 28 pairwise comparisons.

The severity of corrosion at each zone with respect to the other zones was assessed based on its corresponding OR values. For each zone, Table 5-2 has provided seven OR values wherein that particular zone appears as either primary or reference.

Table 5-3 sorts the eight zones from the least to the most severely damaged according to the value of C1 + C2. This value quantifies how many times each zone had a higher likelihood of damage compared with the other seven zones throughout the 28 pairwise comparisons. C1 indicates how many time a particular zone, as the primary, had an OR value above 1, while C2 indicates how many times that same zone, as the reference, had an OR value below 1. Therefore, both C1 and C2 reflects the frequency of each zone appearing as more severely damaged with respect to the other zones.

Table 5-3. The frequency of each zone showing statistically significant OR.

Zone	C1	C2	C1 + C2
Posterior_Proximal (PP)	0	0	0
Medial_Proximal (MP)	1	0	1
Anteriori_Proximal (AP)	2	0	2
Lateral_Proximal (LP)	2	1	3
Anterior_Distal (AD)	4	0	4
Posterior_Distal (PD)	1	4	5
Lateral_Distal (LD)	4	2	6
Medial_Distal (MD)	3	4	7

Zones PP and MD were identified as having the least and highest severity of corrosion. Interestingly, proximal and distal regions were found to be grouping together in this table with the distal region showing more damage compared with the proximal region across the four quadrants in the studied stem tapers.

5.1 Summary and Discussion of the Results

Eight distinct zones of the stem tapers including Anterior_Distal, Anterior_Proximal, Medial_Distal, Medial_Proximal, Posterior_Distal, Posterior_Proximal, Lateral_Distal, and Lateral_Proximal were scored and statistically compared to identify the zone(s) with the most severe corrosion damage in the retrieved implants studied in this work.

A univariate OLR was carried out to identify which zones sustained statistically higher levels of damage. The three histograms of corrosion scores highlighted that Medial _Distal and Posterior_Proximal zones have relatively higher and lower corrosion scores, respectively. To find more details, zone was used as the predictor in an OLR model with proportional odds. Out of 28 possible pairwise comparisons of zone groups, 11 turned out to be significant. To find the order by which the zone groups are damaged, regardless of the p-value, an index that measures the frequency of each zone group being more damaged against the other zones was introduced.

The outcome was aligned with the histograms of score levels across the eight zones. Medial_Distal and Posterior_Proximal zones observed having the highest and lowest severity of the damage. Also, the distal regions of the four quadrants were observed having more damage compared with the proximal regions.

There are several works in the literature that chose to score stem tapers holistically, not locally [42, 46, 56, 135, 148]. Within the studies [31, 44, 53, 55, 131, 132, 138-141] that have scored stem tapers locally, the pools of implants had limited diversity in terms of implant properties (e.g. head diameter, articulation type, and stem design). Therefore, it was deemed necessary to explore whether a similar distribution of corrosion damage can be seen in a more heterogeneous pool of implants.

To the best of our knowledge, there are only two studies [140, 141] in the literature that similar to this work have given eight local scores to the stems while the rest have given lower numbers of zones. Among them, one did not compare the scores between the zones [140]. The other compared the four quadrants and the two distal proximal regions separately in terms of corrosion severity and did not determine which zone(s) had the most severe damage [141].

In this study, since there was no particular hypothesis about the relative level of corrosion at the eight zones, 28 pairwise comparisons were carried out to exhaust the entire pairwise comparison of the zones. The distal region of the medial quadrant was found to have the highest odds ratio (meaning the highest corrosion scores) which is aligned with the findings of the literature that identified the distal region [141] and the medial quadrant [6, 24, 42, 53] having the highest corrosion scores. Also, this study shows that the medical region of all the four quadrants had more corrosion damage in comparison with the proximal region of those quadrants. It was therefore found that, regardless of the quadrant, corrosion damage is more present distally than proximally.

Generally, the higher severity of wear or corrosion at a specific zone has been attributed to several factors such as increased micro-motions at the interface, head or stem materials, head diameter, high friction moments, and poor lubrication of bearing articulation. While some act as root causes, the others play the role of causal factors. Also, damage at this junction usually appears as a combination of wear and corrosion mechanisms. Some of these factors may only contribute to a specific mode of damage, while others may contribute toward a set of damage mechanisms.

In a retrieval study of 231 implants [24], the stem tapers received four fretting and corrosion scores corresponding to the four quadrants. The medial and lateral scores were observed to be significantly higher than the scores at the other two quadrants (posterior and anterior). This was explained to be due to a higher likelihood of micro-motions between the head and neck about an axis in the sagittal plane. Similar to the present study, the pool of implants in [24] had a wide diversity, and higher corrosion scores at medial quadrant suggest that it could be a more general phenomenon than patient and implant factors only.

explains that at the double tapered cone design of Profemur Z, the proximal end of the neck experiences an almost pure compression and shear loading. High friction moments at taper junctions was related to poor lubrication of articulation interfaces by another study Bishop et al. [149]. Medial quadrant was identified having higher corrosion scores in a retrieval study of 52 S-ROM components Munir et al. [53]. It was hypothesised that greater micro-motions at this quadrant could result in a more frequent disruption of the passive oxide layer, and consequently more severe corrosion damage. Similar to the conclusion of the study, they maintained that this region is generally under a compression loading regime.

A computational modelling of the stem taper stresses paired with large diameter heads confirmed this hypothesis after witnessing maximum levels of principal stresses at the medial quadrant [150]. In this study, a 3D model of a 12/14 titanium taper paired with cobalt-chromium and alumina heads were used. Increasing the head diameter significantly increased this quadrant's stresses distal to the junction. It was highlighted that pairing of a small trunnion and a large head leads to a larger moment arm transmitting a higher force to a small surface area which facilitates tribo-corrosion.

These studies have used relatively homogenous pools of implants, yet they observed a higher level of corrosion at the medial quadrant or distal zones of stem tapers. Based on the findings of this study which show that the distal region of the medial quadrant sustains the most severe corrosion damage, it is understood that this particular zone is most severely damaged compared with all the other zones regardless of the properties and patient characteristics of the investigated pool of implants.

A relatively higher amount of load and stress at the medial quadrant causes elastic strains which appear as surface compression. This may lead to micro-motions of approximately 5 to 40 µm [137] which in turn may result in abrasion or fracture of the oxide layer. The subsequent changes in the metal surface potential and the continuous re-passivation of the oxide layer will change the chemistry of the crevice solution. Ultimately, the deaeration and pH decrease of the solution initiate crevice corrosive attacks [5, 151].

Besides micro-motion, galvanic corrosion at this interface due to using mixed metal components is a potential source of material loss. In this study, 18 (13.1%) implants had mixed head and stem materials, whereas 45 (32.8%) had similar materials. Therefore, galvanic corrosion cannot be nominated as the sole mechanism of corrosion.

Greater corrosion damage at the distal region of stem tapers has been reported by several studies [141, 152]. This has been attributed to using increasingly larger heads which induces a significant rise of stress at this region [139, 150]. Crevice corrosion tends to occur near the bore opening which may explain observing more severe corrosion at the distal region [153].

6 MULTIVARIATE ANALYSIS OF PATIENT/IMPLANT FACTORS

[Image removed due to copyright restriction]

6.1 Introduction

This chapter will use the information from the pool of 137 implants to discuss the issues regarding performing multivariate analysis of implant and patient factors. Unlike the previous chapter that looked at sustained corrosion damage at various zones of stem tapers and investigated the underlying reasons for that, this chapter aims to compare the different outcomes of OLR that were induced by variable selection.

The second set of scores will be used to investigate the influence of nine patient and implant factors on corrosion at stem tapers. OLR with cumulative odds will be used to identify the significant factors and compare the groups of polytomous factors. The outputs of three regression models will be compared to highlight the concerns raised in association with multivariate analysis.

Since OLR can also be used as a predictive model, the accuracy of the model in predicting the corrosion scores based on the available patient and information data will be explored. The performance of OLR is compared with more advanced machine learning methods to identify superior classification algorithms for prediction of scores.

6.2 Preliminary Analyses

The stem tapers were scored by two investigators (R.O and R.M). The differences in scores between the two investigators never exceeded one grade of the Goldberg scoring model, and this occurred in 20% of the examined stem tapers (28 of 137). Via a joint examination, a consensus was reached to resolve the discrepancies.

6.2.1 Visual Scores Statistics

Weighted kappa (κ_W) with quadratic weights was run to determine the inter-observer reproducibility of the corrosion scores. A confusion matrix was established to quantify the disagreements. For quadratic weights, the further away from a disagreement from the perfect agreement, the more harshly that disagreement is considered. Weighted kappa (κ_W) with quadratic weights indicated a statistically significant agreement, $\kappa_W = 0.79$ (95% CI, 0.71 to 0.87), $p < 0.001$ between the two sets of scores. According to Landis et al. [147], the strength of the agreement was classified as good.

Visual scoring of the 137 stem trunnions resulted in 137 corrosion scores. Table 6-1 summarises the quantity and percentage of each score level. Score level 3 had the highest quantity (64), while the lowest quantity (3) belonged to score level 1.

Table 6-1. The distribution of corrosion scores across 137 stem tapers

Score	Quantity (%)
1	3 (2.2)
2	54 (39.1)
3	64 (46.4)
4	17 (12.3)

6.2.2 Data Exploration

Since it was desired to determine a subset of the predictors for an adjusted multivariate analysis, first each factor went individually through univariate OLR to determine whether it is statistically significant. Table 6-2 sorts the significance levels from smallest to largest for the nine factors.

Table 6-2. Statistical significance level of each factor

Factor	p-value
Head material	0.193
Age	0.498
Head diameter	0.526
Hip fixation	0.544
Stem taper	0.605
Time to revision	0.638
Joint side	0.683
Stem material	0.889
Gender	0.925

Since none of them was observed to be statistically significant, a subset of seven confounding variables which have been investigated (individually or in a group) in the past and observed to be potential contributors toward damage by the literature were selected. This set is comprised of four categorical variables (Stem Material, Head Material, Stem Taper, and Hip Fixation) and three continuous variables (Time to Revision, Head Diameter, and Age at Insertion). The demography information is summarised by Table 6-3 which is representative of 52 (38%) implants with available information across the entire seven adjusted variables. For stem taper, 11/13 and Type_1 were eliminated due to having zero records in the set of 52 implants.

Table 6-3. The demography of the selected 52 operations

Predictor	Quantity (% of 52 implants)	Mean	Range
Head Material			
CoCr	40 (76.9)		
SS	6 (11.5)		
Ceramic	6 (11.5)		
Stem Material			
CoCr	27 (51.9)		
SS	15 (28.8)		
Titanium	10 (19.2)		
Stem Fixation			
Cemented	29 (55.8)		
Cementless	23 (44.2)		
Stem Taper			
12/14	30 (57.7)		
V40	5 (9.6)		
9/10	3 (5.8)		
6°	8 (15.4)		
C-TAPER	5 (9.6)		
TYPE 1	0 (0.0)		
11/13	0 (0.0)		
10/12	1 (1.9)		
Head Diameter (mm)		32.35	26 - 55
Time to Revision (year)		6.92	0 - 27
Age (year)		63.31	30 - 85

6.2.3 Assessing The Quality Of The Adjusted Variables

The validity of the OLR model obtained by adjusting for these variables from a statistical point of view was investigated. Accordingly, the quality of the model fit, the multi-collinearity, and proportional odds assumptions were considered.

6.2.3.1 *The Overall Model Fit*

Due to observing a high quantity of 156 (75.0%) cells with zero frequencies, overall goodness-of-fit statistics were not deemed to be reliable to test the overall goodness-of-fit of the model. Therefore, the quality of the model fit was investigated by comparing the full model to the intercept-only model. As such, a significant difference between the -2 log-likelihood of these two models was desired which implies the confounding predictors are better at explaining the scores. Table 6-4 presents model fitting information for these variables. The difference between these two models was observed to be significant as desired.

Table 6-4. Comparison between the full and intercept-only model

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	104.971			
Final	81.296	23.675	13	0.034

6.2.3.2 *Multi-Collinearity*

To test the multi-collinearity assumption, collinearity diagnostic which return Variance of Inflation Factor (VIF) was used. VIF indicates to what extent any particular predictor is contributing to multi-collinearity issues within the dataset. A general rule of thumb is that values over 10 indicate very strong multi-collinearity which was adopted here. The polytomous variables were dummy coded, and all the dummy variables except the reference group were included. Table 6-5 summarises the outcome. As can be seen, the VIF index never reached 10 which indicates that the adjusted variables are not associated with multi-collinearity concerns.

Table 6-5. VIF used as a metric to investigate multi-collinearity

(Dummy)	VIF
Head Diameter	2.583
Time to Revision	3.258
Age	1.891
Stem_CoCr	7.823
Stem_Ti	8.324
Head_CoCr	4.004
Head_SS	7.221
Stem Fixation	4.621
V40	3.892
9/10	4.177
6°	3.187
C_Taper	2.661
10/12	1.265

6.2.3.3 *Proportional Odds*

Proportional odds assumption was tested via the test of parallel lines that compare the fit of the proportional odds model to a model with varying slope coefficients. It was desired not to reject the null hypothesis that states the slope coefficients are the same across the three cumulative regression models. Table 6-6 displays the outcome of this test. The obtained p-value (0.613) did not reject the null hypothesis as desired.

Table 6-6. The outcomes of the test of parallel lines

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Null Hypothesis	81.296			
General	57.938	23.358	26	0.613

6.3 The Influence of the Factors

Two other OLR analyses were performed to investigate the influence of the predictors. The first one investigated the significance of the seven adjusted predictors individually using the 52 implants. The second group used these predictors in a multivariate regression model. The role of the first group of analyses was to ensure that any differences between the previous section analysis and the multivariate analysis of these seven predictors were due to the adjustments and not due to the different implant records.

Throughout these analyses, the polytomous predictors (i.e. stem *taper*, *stem material*, and *head material*) incurred additional calculations to obtain an overall test of statistical significance since always one group had to be taken as reference. For a polytomous predictor with *c* groups, *c*(*c*-1)/2 dichotomous dummy variables were generated. For instance, *taper* had six distinct groups; hence, 6(6-1)/2 or 15 tests of statistical significance did a pairwise comparison of the groups.

6.3.1 Univariate Analyses of the Confounding Variables

The first group of analyses performed a univariate regression of the seven predictors over the 52 implant records. Table 6-7 sorts the significance levels from smallest to largest.

Table 6-7. The significance levels of the factors from the univariate OLR

Factor	**p-value**
Head material	0.026
Age	0.245
Head diameter	0.357
Stem taper	0.528
Hip fixation	0.599
Stem material	0.739
Time to revision	0.973

Unlike the previous section that did a univariate analysis of the entire dataset and observed only stem taper having one out of 28 significant odds ratio values, only head material in this analysis was observed to be significant at SS (Ceramic) with an odds ratio of 0.031.

By comparing Table 6-7 with Table 6-2, it can be concluded that adjusting for the seven confounding variables does not dramatically influence the significance level of the predictors in univariate analysis.

6.3.2 Multivariate Analyses of the Confounding Variables

The second group of analyses performed a multivariate regression of the seven predictors. The overall significance of the predictors in a multivariate analysis can be investigated by referring to the *Test of Model Effects* table at SPSS output report. Table 6-8_displays the sorted values. Stem material and head material turned out to be significant predictors of corrosion scores in this group of analysis.

Table 6-8. The significance levels of the factors from the multivariate OLR

Factor	p-value
Stem material	0.010
Head material	0.012
Age	0.156
Stem taper	0.207
Hip fixation	0.484
Time to revision	0.654
Head diameter	0.980

Table 6-9 summarises the OR values, confidence intervals (CIs), and *p* values. The groups of polytomous variables were compared by each other by taking them as primary and reference (inside bracket). At each mutual comparison, the odds of receiving a higher score for the primary group compared with the reference group is provided. The majority of the significant odds ratios (grey rows) belong to stem material and head material comparisons which is in accordance with Table 6-8 findings. Stem taper had the remaining three significant odds ratios wherein 9/10 group shows a lower likelihood of receiving higher corrosion scores.

Table 6-9. The outcome of the multivariate OLR of the seven predictors

Analysis III		**OR**	**p-value**	**CI ($p < 0.05$)**	
head diameter (mm)		1.001	0.980	0.899	1.115
time to revision (year)		0.965	0.654	0.825	1.129
age (year)		0.952	0.156	0.890	1.019
stem material	CoCr (Ti)	2.581	0.468	0.199	33.381
	SS (Ti)	1165.610	0.005	8.706	156216.790
	CoCr (SS)	0.002	0.003	< 0.001	0.121
head material	CoCr (Ceramic)	0.066	0.061	0.004	1.137
	SS (Ceramic)	< 0.001	0.003	< 0.001	0.054
	CoCr (SS)	330.961	0.009	4.293	25539.971
stem fixation	cemented (cementless)	2.368	0.484	0.212	26.417
stem taper	12/14 (10/12)	0.366	0.694	0.002	55.054
	V40 (10/12)	0.006	0.103	< 0.001	2.780
	9/10 (10/12)	< 0.001	0.027	< 0.001	0.424
	6° (10/12)	0.013	0.160	< 0.001	5.508
	C-Taper (10/12)	0.969	0.992	0.002	518.506
	12/14 (C-Taper)	0.377	0.533	0.018	8.067
	V40 (C-Taper)	0.006	0.076	< 0.001	1.681
	9/10 (C-Taper)	< 0.001	0.028	< 0.001	0.445
	6° (C-Taper)	0.014	0.062	< 0.001	1.249
	12/14 (6°)	27.872	0.052	0.971	800.150
	V40 (6°)	0.475	0.707	0.010	23.195
	9/10 (6°)	0.033	0.174	< 0.001	4.521
	12/14 (9/10)	838.564	0.016	3.486	201746.806
	V40 (9/10)	14.281	0.155	0.364	559.935
	12/14 (V40)	58.720	0.066	0.769	4485.766

From Table 6-9, similar to the previous chapter, the order by which the groups of polytomous variables (stem material, head material, and stem taper) had higher severity of corrosion can be determined regardless of their significance levels. Table 6-10 summarises the results for these three polytomous predictors. C1 represents the number of times a group had an odds ratio above one as the primary group. C2 represents the number of times a group had an odds ratio below one as the reference group.

Table 6-10. Comparison of the severity of corrosion across the groups of the three polytomous predictors

Factor	Group	C1	C2	TOTAL
Stem material	Ti	0	0	0
	CoCr	1	0	1
	SS	1	1	2
Head material	SS	0	0	0
	CoCr	1	0	1
	Ceramic	0	2	2
Stem taper	9/10	0	0	0
	V40	1	0	1
	6°	0	2	2
	12/14	3	0	3
	C-Taper	0	4	4
	10/12	0	5	5

Stainless steel stems, ceramic heads, and 10/12 tapers were associated with higher corrosion scores compared with the other groups in each of their corresponding polytomous variable.

6.4 **Prediction of Scores**

Prediction of corrosion scores was carried out via different classification models. Unlike the causal-explanatory statistical modelling which demands to check several assumptions before adjusting for variables, predictive analytics methods do not impose such limitations. However, due to the missing information, including a higher number of predictors in the model results in a reduction in the number of implants with information available across the entire record fields.

Besides, there needs to be a balance between the number of included predictors and the number of available observations. Having too many predictors in a model may produce high multi-collinearity and introduce some noise into the classification model. On the other hand, incorporation very few variables may deprive the model of the useful information that could have enhanced the accuracy rate.

Considering that the inclusion of some particular variables might be preferential to the rest in the predictive models, a set of variables were selected as the predictors in the classification models. From a clinical decision-making perspective, selection of stem material and head material for a patient with a specific age can be preferential owing to the diversity and modularity of hip replacement implants. These factors were observed to be significant in the multivariate analysis. Also, time to revision is a key factor to predict the corrosion severity within any desired time frame after the operation. The polytomous predictors of stem material and head material were dummy coded, and their last categories were used as the reference. Therefore, the predictive models were comprised of age and time to revision as two continuous predictors and four binary predictors that were associated with two categories of stem material and head material.

The size of the available implant record for this set of predictors is 58 which is a reasonable value due to having around 15 observations per predictor. Table 6-11 summarises the quantity of the score levels for these 58 implant records.

Table 6-11. The distribution of the score levels across the 58 selected implants

Score	Quantity (%)
1	1 (1.7)
2	21 (36.2)
3	31 (53.5)
4	5 (8.6)

The classes are highly imbalanced with score levels 2 and 3 possessing the majority of implant records. Therefore, it is expected the same pattern to occur for the predicted scores. Also, this fact led to not using any validation scheme and just reporting the accuracy rate of models that use the entire 58 records for training.

6.4.1 Ordinal Logistic Regression

The prediction accuracies of two OLR models were determined here. Since the assumption of proportional odds is not required in prediction, the accuracy rates of OLR models with and without this assumption was determined to find how this condition may influence the prediction of the visual scores. After establishing the confusion matrix for these two predictive models, they were observed to be identical. Table 6-12 quantifies the actual scores against the predicted scores. The accuracy rate is 58.6% for both OLR models.

Table 6-12. Confusion tables for the OLR model

			Predicted Score				TOTAL
			1	2	3	4	
Actual Score	1	count	0	0	1	0	1
		%	0.0%	0.0%	100%	0.0%	100%
	2	count	0	5	16	0	21
		%	0.0%	23.8%	76.2%	0.0%	100%
	3	count	0	2	29	0	31
		%	0.0%	6.5%	93.5%	0.0%	100%
	4	count	0	0	5	0	5
		%	0.0%	0.0%	100%	0.0%	100%

The identical classification of implant records by these two OLR models shows that having proportional odds does not influence the prediction of scores.

6.4.2 Alternative Machine Learning Methods

Via Classification Learner app in MATLAB, three different classifications models were trained to compare the performance of OLR with these complex nonlinear algorithms. Table 6-13 summarises the performance of the top three accurate algorithms against the OLR models. As mentioned before, due to the limited size of the classes, no validation scheme was carried out.

Table 6-13. The summary of the accuracy rates in the prediction of scores

Technique	Accuracy rate (%)
k-Nearest Neighbour	98.3
Support Vector Machine	87.9
Decision Tree	74.1
OLR	58.6

6.5 Summary and Discussion of the Results

The statistical analysis in the literature of retrieval studies is dominated by univariate analysis of patient and implant factors. A large number of factors that have been identified so far as potential contributors toward damage and the synergistic role of these factors in the outcome of hip replacement operations demand looking at the influence of these factors, concurrently, via multivariate analysis. This study is among the few works in the literature that performed multivariate analyses of several patient and implant factors. OLR with proportional odds was used to carry out three groups of regression modelling.

The outcome of the first group showed that none of the ten predictors is individually significant. The second and third groups used a subset of 52 implant records with information about seven predictors. The univariate analysis of these factors revealed that head material has become significant. The multivariate analysis detected the statistical level of significance for head material as well as stem material. It can be concluded that adjusting for that seven predictors is the reason for stem material, yet such conclusion cannot be made with certainty for head material as its significance may be attributed to using a different dataset. For polytomous variables of stem taper, head material, and stem material, a method was introduced to sort their corresponding groups in order of increasing corrosion severity.

Also, the prediction accuracy of the corrosion scores by OLR was investigated. Due to having missing information, a subset of 58 implant record with information available about four predictors of head material, stem material, age, and time to revision was selected to be used in several prediction algorithms. The unbalanced size of the score levels did not allow for performing validation. The accuracy rates showed that nonlinear classification algorithms are potentially better at predicting the corrosion scores.

Regression is the sole method that can handle the multivariate analysis of continuous and categorical factors. The majority of the literature methods which are based on comparing the mean (median) of scores across different groups of nominal predictors [30, 37, 136, 138] or Pearson and Spearman correlations [30, 132] face limitations to perform multivariate analysis. These limitations mostly arise from assumptions that demand specific distribution of scores and equal variance in each cell of design. However, the assumptions of OLR does not require maintaining such conditions. Unlike the other methods which can handle only specific types of predictors in terms of the level of measurement, OLR can handle continuous, dichotomous, and polytomous predictors simultaneously.

The capability of OLR to handle multivariate analysis was investigated here through using the database of 137 retrieved implants with nine patient and implant fields (head material, head diameter, hip fixation, stem taper type, stem material, joint side, time to revision, age, gender). The assumptions of OLR did not allow for using the entire nine factors in the regression models which resulted in the need for variable selection. The high number of all possible subsets of predictors that could be considered in the regression model turns the variable selection into a complicated task. This matter was elaborated by performing three groups of regression analyses. The outcomes showed that the predictors might shift from being significant into insignificant and vice versa based on the available data (the first two groups of regressions) as well as the seven adjusted predictors (the third group versus groups one and two). For instance, head material was insignificant when all the available data (i.e. 75 out of 137 records) for this factor was used, yet it became significant after adjusting for the seven confounding variables and using this factor in the univariate analysis of the 52 implant records. Also, stem material which was not significant in both of the univariate analyses turned into significant in the multivariate analysis. The difference in the outcome is not confined to just the significance level of the factors, The odds ratio values vary as the main outcome of the OLR models as well which leads to different explanations about the associations between these factors and the corrosion scores.

These facts show that although OLR can be superior in conducting multivariate analyses compared with other types of methods, it may suffer when the number of predictors or the groups of polytomous predictors increases.

Several machine learning techniques were used to investigate how accurate they can predict the corrosion scores for the dataset that was available for this study. OLR with and without the assumption of proportional odds showed that maintaining this assumption does not change the accuracy (58.6%) of the model fitted to this particular dataset. Also, this method, as a linear model, did not perform as well as the other, more complex nonlinear methods. The very high non-validated accuracy of the nonlinear methods raised overfitting concerns which need to be addressed via a cross-validation scheme. However, the small sample size of score levels 1 and 4 did not permit a cross-validation in the dataset used as a sample for this purpose.

7 CONCLUSIONS AND SUGGESTIONS FOR FUTURE WORKS

[Image removed due to copyright restriction]

7.1 Introduction

This chapter will endeavour to use the obtained experience from this work to provide a future-oriented outlook for works that consider expanding the horizon of leveraging AI in orthopaedics and more specifically, large-scale retrieval studies of orthopaedic implants.

7.2 AI and Orthopaedics

The potential applications of AI tools (i.e. DIP and machine learning) in orthopaedics are elaborated in this section.

7.2.1 DIP

Image-based rating of corrosion severity at modular junctions as well as other surface areas of orthopaedic implants can resolve the concerns associated with the reliability of visual scoring. However, capturing images of hard to reach areas such as bore taper of femoral heads may pose limitations to the applicability and even feasibility of image-based rating. Despite that limitation, this approach offers other advantages such as creating large databases of implant images through time that can be shared and used in future. Also, it facilitates the visual scoring of various zones of a specific interface, as discussed in chapter 5.

Section 3.2 detailed out the process to clean the stem taper surfaces according to an existing protocol. Here, it should be noted that the total removal of dirt, the tissue remains, and bloodstains from these surfaces is not always possible. One reason for these limitations is the desire to keep the corroded and worn areas intact for the studies. In this study, it was endeavoured not to include implants with significant dirt to minimise its impacts on the results.

This work just relied on analysing RGB optical images of retrieved implants. However, image analysis is not just limited to post-revision. Through the life of an in-situ implant, medical images can be obtained and analysed continuously by DIP and machine learning algorithms. They can detect or measure variables representative of the current status of implant/host such the position and the orientation of implants components concerning each other or the nearby bones and tissues. Also, nearby tissues damage induced by particulate wear/corrosion debris can be detected and evaluated. Acquiring and analysing such images regularly can facilitate early detection of a problematic implant.

This information, along with the retrieved implants and the corresponding medical records, can support retrieval studies to gain a more detailed understanding of the causal factors and their interactions throughout the implantation time. The importance of leveraging AI in orthopaedic image analysis has been already highlighted by some clinicians [154].

7.2.2 Machine Learning

have reviewed the literature on musculoskeletal problems and related health conditions to learn the extent by which machine learning techniques have been employed since 2000. They listed several studies in which researchers developed and applied machine learning to support clinicians in their main tasks, whose essence is the need to make decisions in the absence of certitude and translate these decisions into choices of care for the betterment of patients.

The present work used machine learning for two distinct purposes. The first one served the development of a classification model for prediction of scores by using some textural features that characterise the severity of corrosion damage. As mentioned in Section 4.5, currently, the feature extraction process is getting automated by taking advantage of adaptive learning techniques. These techniques can learn a set of features that best fit in a predictive model of the response variable (e.g. corrosion scores).

Adaptive learning comes at a much higher computation cost compared with non-adaptive learning and usually demands a larger sample size within each class (e.g. score level). The advent of powerful processors, cloud computing tools, and data augmentation techniques have paved the path to overcoming this limitation. High volumes of data can be analysed on an ongoing basis to extract knowledge from experience to improve processes and decision outcomes.

On the same token, this study endeavoured to embed the knowledge and experience of a human expert to develop a classification model that visually scores modular junctions of implants with no human interventions (objectively).

Machine learning, also served the second purpose in this study. Rather than using DIP features from the implant images to predict the severity of corrosion damage, prediction of scores from the patient and implant records was investigated. That can be a very important milestone since the prediction of damage from images can be made only after failure and retrieval of implants which incur high costs for both the patients and the medical systems. On the contrary, using the empirical data from medical records from the past primary/revision operations allows for predicting the severity of corrosion through time (similar to survival analysis) or even earlier where decisions are made about the implant to be used for a prospective recipient.

The available dataset for this study was quite limited, and the size of the implants within each score level was small and unbalanced. This matter was highlighted (Sections 3.8, 3.8.3.3, and 3.8.3.4) and explored (Sections 6.4) using the available data.

Another reason that justifies a broader use of machine learning in large-scale retrieval studies is the limitations associated with the existing causal-explanatory statistical modelling techniques (elaborated in Section 6.5). Although OLR can be superior in conducting multivariate analyses compared with other types of methods, it may suffer when the number of predictors or the groups of polytomous predictors increases. To address this matter, alternative machine learning techniques can be employed to capture the underlying characteristics of the data. A classification model of damage scores is not subject to complexities associated with the variable selection. Recently, with the advent of powerful computers and the availability of several machine learning libraries and open-source toolkits, it has become feasible to incorporate these algorithms in various classification as well as regression contexts.

Overall, this study detailed out the process to apply supervised machine learning algorithms to predict corrosion scores based on the implant and patient information. It is a decision support tool to estimate the probability of severe corrosion damage at in-situ implants for surgeons any time after an operation. It can facilitate the decision-making process regarding when to remove a seemingly malfunctioning implant. The performance of these machine learning algorithms can be improved by refining (e.g. via hyperparameter optimisation) and incorporation of larger datasets which was out of the scope of this study, and it can be investigated through future works.

7.3 Electronic Medical Records and Clinical Decision Support Systems

The essence of retrieval studies is to produce evidence-based knowledge to achieve better clinical decision-making which in turn can improve the outcomes of primary and revision operations. This can be done over two stages of systematically collecting and analysing the data. In a larger scale, successful commercialisation and adoption of many Electronic Medical Record (EMR) systems have shifted the question from gathering the data to converting the massive amounts of available data into knowledge directly applicable to diagnosis, prognosis, or treatment.

The result has been the evolution of EMR-based Clinical Decision Support Systems (CDSS) in a variety of healthcare settings such as precision medicine which serve in capacities beyond generating alerts and reminders [155-157]. Considering the increasingly voluminous data associated with medical records as well as knowledge accumulated from the past retrieval studies, it is necessary to integrate similar systems that encode the knowledge in a way that makes it actionable by clinicians and manufacturers.

7.4 Concluding Remarks

1. Image-based rating facilitates visual scoring of stem taper zones.
2. The expertise in visual scoring can be captured and used objectively via image classification algorithms.
3. Global and local textural features used in rating corrosion severity within other corrosion-related domains can also be used for taper corrosion.
4. The significant difference observed in corrosion scores across eight stem taper zones mandates using local scores over scoring holistically.
5. Using overall visual scores as a continuous variable should be treated with suspicion.
6. Corrosion at stem tapers is more commonly seen distally and medially.
7. Despite the extensive advancements in characterising damage at modular orthopaedic implants, the data analysis techniques are just limited to causal-explanatory statistical modelling.
8. While the techniques to quantify the damage, analyse wear and corrosion mechanisms, and characterise corrosion by-products have advanced significantly, analysing the raw data produced by them is still limited to causal-explanatory statistical modelling and have not evolved properly.
9. No certain protocol exists for statistical analyses in this area. Different studies use different approaches to similar works. The impact of using different analyses on the results are not known.
10. The synergistic role of several patient/implant factors at the outcome of hip replacement operations demands for multivariate data analysis. Due to the limitation in sample size, the majority of studies are content with using just one or just a few factors.

11. Ordinal logistic regression is recommended over the other causal-explanatory statistical modelling.

12. Due to concerns associated with variable selection in regression models, causal-explanatory statistical modelling is recommended to be used alongside predictive modelling to capture the underlying characteristics of the data.

7.5 Suggestions for Future Work

1. The potential utilisation of machine learning and deep learning tools to be explored to generate predictive analytics from the vast amounts of primary and retrieval operation records.

2. This study endeavoured to predict corrosion scores from image descriptors. These descriptors can expand to several patient/implant/surgical factors for prediction of not just corrosion scores, but the likelihood of implant failure for a specific patient using a specific hip replacement

3. The successful experience of decision-making related areas with leveraging state-of-the-art big data storage and analysis tools need to be investigated to develop a clinical decision support system in the context of orthopaedic implants.

8 REFERENCES

1. Stephen Richard Knight, S.P.B., *Total Hip Arthroplasty - over 100 years of operative history.* Orthopaedic Reviews, 2011. **3**.
2. Harris, W.H., *A New Total Hip Implant.* Clinical Orthopaedics and Related Research, 1971. **81**: p. 105-113.
3. Collier, J.P., et al., *Mechanisms of Failure of Modular Prostheses.* Clinical Orthopaedics and Related Research, 1992. **285**: p. 129-139.
4. Gilbert, J.L., *Hip Implant Corrosion Mechanisms and Effects: Mechanically Assisted Corrosion, Crevices and Voltage Effects*. 2011, Syracus Biomaterial Institute.
5. Osman, K., et al., *Corrosion at the head-neck interface of current designs of modular femoral components ESSENTIAL QUESTIONS AND ANSWERS RELATING TO CORROSION IN MODULAR HEAD-NECK JUNCTIONS.* Bone & Joint Journal, 2016. **98B**(5): p. 579-584.
6. Higgs, G.B., et al., *Is Increased Modularity Associated With Increased Fretting and Corrosion Damage in Metal-On-Metal Total Hip Arthroplasty Devices?: A Retrieval Study.* The Journal of Arthroplasty, 2013. **28**(8, Supplement): p. 2-6.
7. Kocagöz, S.B., et al., *Does taper angle clearance influence fretting and corrosion damage at the head–stem interface? A matched cohort retrieval study.* Seminars in Arthroplasty, 2013. **24**(4): p. 246-254.
8. Pivec, R., et al., *Modular Taper Junction Corrosion and Failure: How to Approach a Recalled Total Hip Arthroplasty Implant.* The Journal of Arthroplasty, 2014. **29**(1): p. 1-6.
9. Hoffman, A.S., *Chapter I.2.1 - Introduction: The Diversity and Versatility of Biomaterials*, in *Biomaterials Science (Third Edition)*, B.D.R.S.H.J.S.E. Lemons, Editor. 2013, Academic Press. p. 63-64.
10. Brunski, J.B., *Chapter i.2.3 - Metals: Basic Principles*, in *Biomaterials Science (Third Edition)*, B.D.R.S.H.J.S.E. Lemons, Editor. 2013, Academic Press. p. 111-119.
11. Hallab, N.J. and J.J. Jacobs, *Chapter II.5.6 - Orthopedic Applications*, in *Biomaterials Science (Third Edition)*, B.D.R.S.H.J.S.E. Lemons, Editor. 2013, Academic Press. p. 841-882.
12. Gilbert, J.L., et al., *Direct in vivo inflammatory cell-induced corrosion of CoCrMo alloy orthopedic implant surfaces.* Journal of Biomedical Materials Research Part A, 2014. **103**(1): p. 211-223.
13. Hench, L.L. and S.M. Best, *Chapter I.2.4 - Ceramics, Glasses, and Glass-Ceramics: Basic Principles*, in *Biomaterials Science (Third Edition)*, B.D.R.S.H.J.S.E. Lemons, Editor. 2013, Academic Press. p. 128-151.
14. Chevalier, J., *What future for zirconia as a biomaterial?* Biomaterials, 2006. **27**(4): p. 535-543.
15. Gallo, J., M. Holinka, and C.S. Moucha, *Antibacterial surface treatment for orthopaedic implants.* International journal of molecular sciences, 2014. **15**(8): p. 13849-80.
16. Williams, D.F., *Definitions in biomaterials: proceedings of a consensus conference of the European Society for Biomaterials, Chester, England, March 3-5, 1986*. Vol. 4. 1987: Elsevier Science Ltd.
17. Ratner, B.D., *Chapter III.1.3 - Correlation, Materials Properties, Statistics and Biomaterials Science*, in *Biomaterials Science (Third Edition)*, B.D.R.S.H.J.S.E. Lemons, Editor. 2013, Academic Press. p. 1354-1361.
18. Horbett, T.A., *Chapter II.1.2 - Adsorbed Proteins on Biomaterials*, in *Biomaterials Science (Third Edition)*, B.D.R.S.H.J.S.E. Lemons, Editor. 2013, Academic Press. p. 394-408.

19. Walter, W.L., et al., *Retrieval analysis of squeaking alumina ceramic-on-ceramic bearings.* J Bone Joint Surg Br, 2011. **93**(12): p. 1597-601.

20. Williams, D.F. and R.L. Williams, *Chapter II.4.4 - Degradative Effects of the Biological Environment on Metals and Ceramics*, in *Biomaterials Science (Third Edition)*, B.D.R.S.H.J.S.E. Lemons, Editor. 2013, Academic Press. p. 728-738.

21. Billi, F. and P. Campbell, *Nanotoxicology of metal wear particles in total joint arthroplasty: a review of current concepts.* J Appl Biomater Biomech, 2010. **8**(1): p. 1-6.

22. Martin, J.R. and R.T. Trousdale, *Unique failure mechanism of a femoral component after revision total hip arthroplasty.* Orthopedics, 2013. **36**(10): p. e1327-9.

23. Molloy, D.O., et al., *Fretting and Corrosion in Modular-Neck Total Hip Arthroplasty Femoral Stems.* The Journal of Bone & Joint Surgery, 2014. **96**(6): p. 488-493.

24. Goldberg, J.R., et al., *A multicenter retrieval study of the taper interfaces of modular hip prostheses.* Clin Orthop Relat Res, 2002(401): p. 149-61.

25. Kop, A.M. and E. Swarts, *Corrosion of a Hip Stem With a Modular Neck Taper Junction: A Retrieval Study of 16 Cases.* The Journal of Arthroplasty, 2009. **24**(7): p. 1019-1023.

26. Schweitzer, P.A., *Corrosion Engineering Handbook, -3 Volume Set*. 1996: CRC Press.

27. Eliaz, N., *Degradation of implant materials*. 2012: Springer Science & Business Media.

28. Ryu, J.J., S. Letchuman, and P. Shrotriya, *Roughness evolution of metallic implant surfaces under contact loading and nanometer-scale chemical etching.* Journal of the Mechanical Behavior of Biomedical Materials, 2012. **14**(0): p. 55-66.

29. *fretting places*.

30. Kurtz, S., et al., *Do Ceramic Femoral Heads Reduce Taper Fretting Corrosion in Hip Arthroplasty? A Retrieval Study.* Clinical Orthopaedics and Related Research®, 2013. **471**(10): p. 3270-3282.

31. Hothi, H.S., et al., *The Reliability of a Scoring System for Corrosion and Fretting, and Its Relationship to Material Loss of Tapered, Modular Junctions of Retrieved Hip Implants.* The Journal of Arthroplasty, 2014. **29**(6): p. 1313-1317.

32. Affatato, S., et al., *Tribology and total hip joint replacement: Current concepts in mechanical simulation.* Medical Engineering & Physics, 2008. **30**(10): p. 1305-1317.

33. *testing set-up*.

34. Dyrkacz, R.M.R., et al., *Finite element analysis of the head–neck taper interface of modular hip prostheses.* Tribology International, 2015.

35. Hothi, H., et al., *The reliability of a semi-quantitative scoring method for taper corrosion and fretting, and its usefulness for predicting the volume of material loss*. 2013.

36. Dyrkacz, R.M.R., et al., *The Influence of Head Size on Corrosion and Fretting Behaviour at the Head-Neck Interface of Artificial Hip Joints.* The Journal of Arthroplasty, 2013. **28**(6): p. 1036-1040.

37. Goyal, N., et al., *Do You Have to Remove a Corroded Femoral Stem?* The Journal of Arthroplasty, 2014. **29**(9, Supplement): p. 139-142.

38. Hothi, H.S., et al., *Influence of stem type on material loss at the metal-on-metal pinnacle taper junction.* Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine, 2015. **229**(1): p. 91-97.

39. Nassif, N., et al., *Taper Design Affects Failure of Large-head Metal-on-metal Total Hip Replacements.* Clinical Orthopaedics and Related Research®, 2014. **472**(2): p. 564-571.

40. Collier, J.P., et al., *CORROSION BETWEEN THE COMPONENTS OF MODULAR FEMORAL HIP PROSTHESES.* Journal of Bone and Joint Surgery-British Volume, 1992. **74**(4): p. 511-517.

41. Gilbert, J.L., C.A. Buckley, and J.J. Jacobs, *In vivo corrosion of modular hip prosthesis components in mixed and similar metal combinations. The effect of crevice, stress, motion, and alloy coupling.* J Biomed Mater Res, 1993. **27**(12): p. 1533-44.
42. Higgs, G.B., et al., *Method of Characterizing Fretting and Corrosion at the Various Taper Connections of Retrieved Modular Components from Metal-on-Metal Total Hip Arthroplasty.* Metal-on-Metal Total Hip Replacement Devices, 2013. **1560**: p. 146-156.
43. Kop, A., C. Keogh, and E. Swarts, *Proximal component modularity in THA—at what cost?: an implant retrieval study.* Clinical Orthopaedics and Related Research®, 2012. **470**(7): p. 1885-1894.
44. De Martino, I., et al., *Corrosion and Fretting of a Modular Hip System: A Retrieval Analysis of 60 Rejuvenate Stems.* J Arthroplasty, 2015. **30**(8): p. 1470-5.
45. Lanting, B.A., et al., *Correlation of corrosion and biomechanics in the retrieval of a single modular neck total hip arthroplasty design: modular neck total hip arthroplasty system.* J Arthroplasty, 2015. **30**(1): p. 135-40.
46. Whittaker, R.K., et al., *The effect of using components from different manufacturers on the rate of wear and corrosion of the head-stem taper junction of metal-on-metal hip arthroplasties.* Bone & Joint Journal, 2016. **98B**(7): p. 917-924.
47. Triantafyllopoulos, G.K., et al., *Otto Aufranc Award: Large Heads Do Not Increase Damage at the Head-neck Taper of Metal-on-polyethylene Total Hip Arthroplasties.* Clinical Orthopaedics and Related Research, 2016. **474**(2): p. 330-338.
48. Pourzal, R., et al., *Does Surface Topography Play a Role in Taper Damage in Head-neck Modular Junctions?* Clinical Orthopaedics and Related Research, 2016. **474**(10): p. 2232-2242.
49. Arnholt, C.M., *Micro-grooved Surface Topography Does Not Influence Fretting Corrosion of Tapers in THA: Classification and Retrieval Analysis*. 2015, Drexel University.
50. dos Santos, C.T., et al., *Characterization of the fretting corrosion behavior, surface and debris from head-taper interface of two different modular hip prostheses.* Journal of the Mechanical Behavior of Biomedical Materials, 2016. **62**: p. 71-82.
51. *Annual Report Australia*. 2015, Australian Orthopaedic Association National Joint Replacement Registry: Australia.
52. Cross, M.B., et al., *Fretting and corrosion changes in modular total hip arthroplasty.* Bone & Joint Journal Orthopaedic Proceedings Supplement, 2013. **95**(SUPP 15): p. 127-127.
53. Munir, S., et al., *Corrosion in modular total hip replacements: An analysis of the head–neck and stem–sleeve taper connections.* Seminars in Arthroplasty, 2013. **24**(4): p. 240-245.
54. Kocagoz, S.B., et al., *Ceramic Heads Decrease Metal Release Caused by Head-taper Fretting and Corrosion.* Clinical Orthopaedics and Related Research®, 2016. **474**(4): p. 985-994.
55. Hothi, H.S., et al., *Detailed inspection of metal implants.* Hip international: the journal of clinical and experimental research on hip pathology and therapy, 2015: p. 0-0.
56. Matthies, A.K., et al., *Material Loss at the Taper Junction of Retrieved Large Head Metal-on-Metal Total Hip Replacements.* Journal of Orthopaedic Research, 2013. **31**(11): p. 1677-1685.
57. Langton, D., et al., *Taper junction failure in large-diameter metal-on-metal bearings.* Bone and Joint Research, 2012. **1**(4): p. 56-63.
58. Codaro, E.N., et al., *An image processing method for morphology characterization and pitting corrosion evaluation.* Materials Science and Engineering: A, 2002. **334**(1–2): p. 298-306.
59. Son, H., et al., *Rapid and automated determination of rusted surface areas of a steel bridge for robotic maintenance systems.* Automation in Construction, 2014. **42**: p. 13-24.

60. Trujillo, M. and M. Sadki. *Sensitivity analysis for texture models applied to rust steel classification*. in *Electronic Imaging 2004*. 2004. International Society for Optics and Photonics.

61. Choi, K.Y. and S.S. Kim, *Morphological analysis and classification of types of surface corrosion damage by digital image processing.* Corrosion Science, 2005. **47**(1): p. 1-15.

62. Lee, S., L.-M. Chang, and M. Skibniewski, *Automated recognition of surface defects using digital color image processing.* Automation in Construction, 2006. **15**(4): p. 540-549.

63. Jahanshahi, M.R., et al., *A survey and evaluation of promising approaches for automatic image-based defect detection of bridge structures.* Structure and Infrastructure Engineering, 2009. **5**(6): p. 455-486.

64. Medeiros, t.N.S., et al., *On the evaluation of texture and color features for nondestructive corrosion detection.* EURASIP J. Adv. Signal Process, 2010. **2010**: p. 1-7.

65. Ghosh, B., V. Pakrashi, and F. Schoefs, *High dynamic range image processing for non-destructive-testing.* European Journal of Environmental and Civil Engineering, 2011. **15**(7): p. 1085-1096.

66. Jahanshahi, M.R. and S.F. Masri, *Parametric Performance Evaluation of Wavelet-Based Corrosion Detection Algorithms for Condition Assessment of Civil Infrastructure Systems.* Journal of Computing in Civil Engineering, 2013. **27**(4): p. 345-357.

67. Shen, H.-K., P.-H. Chen, and L.-M. Chang, *Automated steel bridge coating rust defect recognition method based on color and texture feature.* Automation in Construction, 2013. **31**: p. 338-356.

68. Gamarra Acosta, M.R., J.C. Vélez Díaz, and N. Schettini Castro, *An innovative image-processing model for rust detection using Perlin Noise to simulate oxide textures.* Corrosion Science, 2014. **88**: p. 141-151.

69. Yan, B.F., et al., *Imaging-Based Rating for Corrosion States of Weathering Steel Using Wavelet Transform and PSO-SVM Techniques.* Journal of Computing in Civil Engineering, 2014. **28**(3): p. 13.

70. Feliciano, F.F., F.R. Leta, and F.B. Mainier, *Texture digital analysis for corrosion monitoring.* Corrosion Science, 2015. **93**: p. 138-147.

71. Yeum, C.M. and S.J. Dyke, *Vision-Based Automated Crack Detection for Bridge Inspection.* Computer-Aided Civil and Infrastructure Engineering, 2015. **30**(10): p. 759-770.

72. Chen, P.H. and L.M. Chang, *Intelligent steel bridge coating assessment using neuro-fuzzy recognition approach.* Computer-Aided Civil and Infrastructure Engineering, 2002. **17**(5): p. 307-319.

73. Kyvelidis, S.T., L. Lykouropoulos, and N. Kouloumbi, *Digital system for detecting, classifying, and fast retrieving corrosion generated defects.* Journal of Coatings Technology, 2001. **73**(915): p. 67-73.

74. Solomon, C. and T. Breckon, *Fundamentals of Digital Image Processing: A practical approach with examples in Matlab*. 2011: John Wiley & Sons.

75. Batchelor, B.G., *Machine Vision Handbook*. 2012: Springer.

76. Ledda, A., et al. *Quantitative image analysis with mathematical morphology*. in *ProRISC Ann. Workshop Circuits, Syst. Signal Process*. 2000.

77. Tao, L., et al., *Image analysis of periodic rain accelerated corrosion of aeronautical aluminium alloys.* Materials Science and Engineering a-Structural Materials Properties Microstructure and Processing, 2008. **476**(1-2): p. 210-216.

78. Pidaparti, R.M., et al., *Classification of corrosion defects in NiAl bronze through image analysis.* Corrosion Science, 2010. **52**(11): p. 3661-3666.

79. Lv, S.L., et al., *Influence of morphology of corrosion on fracture initiation in an aluminum alloy.* Materials & Design, 2013. **45**: p. 96-102.
80. Konovalenko, I.V., P.O. Marushchak, and R.T. Bishchak, *Automated Estimation of Damage to the Surface of Gas Main by Corrosion Pittings.* Materials Science, 2014. **49**(4): p. 493-500.
81. Wang, Y. and G. Cheng, *Quantitative evaluation of pit sizes for high strength steel: Electrochemical noise, 3-D measurement, and image-recognition-based statistical analysis.* Materials & Design, 2016. **94**: p. 176-185.
82. San-Martin, M.T. and M. Sadki, *Generalized and optimized classification framework for textural imagery*, in *Applications of Digital Image Processing Xxvii, Pts 1and 2*, A.G. Tescher, Editor. 2004. p. 128-136.
83. Choi, K.Y., A.Y. Grigoriev, and N.K. Myshkin, *Analysis of tribochemical surface damage by image processing (analysis of tribochemical damages).* Tribology Letters, 2002. **13**(2): p. 125-129.
84. Zimer, A.M., et al., *Investigation of AISI 1040 steel corrosion H2S solution containing chloride ions by digital image processing coupled with in electrochemical techniques.* Corrosion Science, 2011. **53**(10): p. 3193-3201.
85. Jian, L., et al., *Determination of corrosion types from electrochemical noise by artificial neural networks.* Int. J. Electrochem. Sci, 2013. **8**: p. 2365-2377.
86. Sonka, M., V. Hlavac, and R. Boyle, *Image processing, analysis, and machine vision*. 2014: Cengage Learning.
87. Hu, F., et al., *Unsupervised Feature Learning Via Spectral Clustering of Multidimensional Patches for Remotely Sensed Scene Classification.* Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2015. **8**(5): p. 2015-2030.
88. Peng, Y.P., et al., *Oxidation wear monitoring based on the color extraction of on-line wear debris.* Wear, 2015. **332**: p. 1151-1157.
89. Jhanwar, N., et al., *Content based image retrieval using motif cooccurrence matrix.* Image and Vision Computing, 2004. **22**(14): p. 1211-1220.
90. Xie, X., *A review of recent advances in surface defect detection using texture analysis techniques.* ELCVIA Electronic Letters on Computer Vision and Image Analysis, 2008. **7**(3).
91. Gonzalez, R.C. and R.E. Woods, *Digital Image processing*. 3 ed. 2007.
92. Wang, S.Y. and S.Z. Song, *Image analysis of atmospheric corrosion exposure of zinc.* Materials Science and Engineering a-Structural Materials Properties Microstructure and Processing, 2004. **385**(1-2): p. 377-381.
93. Arivazhagan, S. and L. Ganesan, *Texture classification using wavelet transform.* Pattern Recognition Letters, 2003. **24**(9-10): p. 1513-1521.
94. Huang, P.-W. and S. Dai, *Image retrieval by texture similarity.* Pattern recognition, 2003. **36**(3): p. 665-679.
95. Li, S.T., et al., *Texture classification using the support vector machines.* Pattern Recognition, 2003. **36**(12): p. 2883-2893.
96. Nixon, M.S. and A.S. Aguado, *Chapter 8 - Introduction to texture description, segmentation, and classification*, in *Feature Extraction & Image Processing for Computer Vision (Third edition)*. 2012, Academic Press: Oxford. p. 399-434.
97. Zhang, J., et al., *Local features and kernels for classification of texture and object categories: A comprehensive study.* International journal of computer vision, 2007. **73**(2): p. 213-238.
98. Quelhas, P., et al., *A thousand words in a scene.* Ieee Transactions on Pattern Analysis and Machine Intelligence, 2007. **29**(9): p. 1575-1589.
99. Bay, H., et al., *Speeded-up robust features (SURF).* Computer vision and image understanding, 2008. **110**(3): p. 346-359.

100. Leung, T. and J. Malik, *Representing and recognizing the visual appearance of materials using three-dimensional textons.* International journal of computer vision, 2001. **43**(1): p. 29-44.

101. Mikolajczyk, K., et al., *A Comparison of Affine Region Detectors.* International Journal of Computer Vision, 2005. **65**(1): p. 43-72.

102. Doshi, N.P., *Multi-dimensional local binary pattern texture descriptors and their application for medical image analysis*. 2014, © Niraj P. Doshi.

103. Anthimopoulos, M.M., et al., *A Food Recognition System for Diabetic Patients Based on an Optimized Bag-of-Features Model.* IEEE Journal of Biomedical and Health Informatics, 2014. **18**(4): p. 1261-1271.

104. Csurka, G., et al. *Visual categorization with bags of keypoints*. in *Workshop on statistical learning in computer vision, ECCV*. 2004. Prague.

105. Vetrivel, A., et al., *Identification of structurally damaged areas in airborne oblique images using a visual-Bag-of-Words approach.* Remote Sensing, 2016. **8**(3): p. 231.

106. Lowe, D.G., *Distinctive Image Features from Scale-Invariant Keypoints.* International Journal of Computer Vision, 2004. **60**(2): p. 91-110.

107. Zhao, L., P. Tang, and L. Huo, *A 2-D wavelet decomposition-based bag-of-visual-words model for land-use scene classification.* International Journal of Remote Sensing, 2014. **35**(6): p. 2296-2310.

108. Altintakan, U.L. and A. Yazici, *Towards Effective Image Classification Using Class-Specific Codebooks and Distinctive Local Features.* Ieee Transactions on Multimedia, 2015. **17**(3): p. 323-332.

109. Zheng, L., S.J. Wang, and Q. Tian, *Coupled Binary Embedding for Large-Scale Image Retrieval.* Ieee Transactions on Image Processing, 2014. **23**(8): p. 3368-3380.

110. van Gemert, J.C., et al., *Visual Word Ambiguity.* Ieee Transactions on Pattern Analysis and Machine Intelligence, 2010. **32**(7): p. 1271-1283.

111. Stanciu, S.G., et al., *Experimenting liver fibrosis diagnostic by two photon excitation microscopy and Bag-of-Features image classification.* Sci Rep, 2014. **4**: p. 4636.

112. Tommasi, T., F. Orabona, and B. Caputo, *Discriminative cue integration for medical image annotation.* Pattern Recognition Letters, 2008. **29**(15): p. 1996-2002.

113. Pires, R., et al., *Advancing bag-of-visual-words representations for lesion classification in retinal images.* PLoS One, 2014. **9**(6): p. e96814.

114. Van Gemert, J.C., et al. *Kernel codebooks for scene categorization*. in *European conference on computer vision*. 2008. Springer.

115. Zhao, L.J., P. Tang, and L.Z. Huo, *Land-Use Scene Classification Using a Concentric Circle-Structured Multiscale Bag-of-Visual-Words Model.* Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2014. **7**(12): p. 4620-4631.

116. Perronnin, F., *Universal and adapted vocabularies for Generic Visual Categorization.* Ieee Transactions on Pattern Analysis and Machine Intelligence, 2008. **30**(7): p. 1243-1256.

117. Dimitrovski, I., et al., *Improved medical image modality classification using a combination of visual and textual features.* Computerized Medical Imaging and Graphics, 2015. **39**: p. 14-26.

118. Bosch, A., A. Zisserman, and X. Muñoz. *Scene classification via pLSA*. in *European conference on computer vision*. 2006. Springer.

119. Angeli, A., et al., *Fast and Incremental Method for Loop-Closure Detection Using Bags of Visual Words.* IEEE Transactions on Robotics, 2008. **24**(5): p. 1027-1037.

120. Bosch, A., A. Zisserman, and X. Munoz, *Scene classification using a hybrid generative/discriminative approach.* Ieee Transactions on Pattern Analysis and Machine Intelligence, 2008. **30**(4): p. 712-727.

121. Tahir, M., et al. *Visual category recognition using spectral regression and kernel discriminant analysis*. in *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*. 2009. IEEE.

122. Perronnin, F., J. Sanchez, and T. Mensink, *Improving the Fisher Kernel for Large-Scale Image Classification.* Computer Vision-Eccv 2010, Pt Iv, 2010. **6314**: p. 143-156.

123. Sanchez, J., et al., *Image Classification with the Fisher Vector: Theory and Practice.* International Journal of Computer Vision, 2013. **105**(3): p. 222-245.

124. Krig, S., *Computer Vision Metrics, Survey, Taxonomy, and Analysis. Apress*. 2014: Oxford.

125. Nixon, M.S. and A.S. Aguado, *Chapter 4 - Low-level feature extraction (including edge detection)*, in *Feature Extraction & Image Processing for Computer Vision (Third edition)*. 2012, Academic Press: Oxford. p. 137-216.

126. Yang, W., K. Wang, and W. Zuo, *Neighborhood Component Feature Selection for High-Dimensional Data.* JCP, 2012. **7**(1): p. 161-168.

127. Yamana, M., et al. *Development of system for crossarm reuse judgment on the basis of classification of rust images using support vector machine*. in *Tools with Artificial Intelligence, 2005. ICTAI 05. 17th IEEE International Conference on*. 2005. IEEE.

128. Szeliski, R., *Computer vision: algorithms and applications*. 2010: Springer Science & Business Media.

129. Wang, R.G., et al., *A novel method for image classification based on bag of visual words.* Journal of Visual Communication and Image Representation, 2016. **40**: p. 24-33.

130. Zhang, Y., S. Wang, and G. Ji, *A Comprehensive Survey on Particle Swarm Optimization Algorithm and Its Applications.* Mathematical Problems in Engineering, 2015. **2015**: p. 1-38.

131. Tan, S.C., et al., *Tribocorrosion: Ceramic and Oxidized Zirconium vs Cobalt-Chromium Heads in Total Hip Arthroplasty.* Journal of Arthroplasty, 2016. **31**(9): p. 2064-2071.

132. Tan, S.C., et al., *Effect of Taper Design on Trunnionosis in Metal on Polyethylene Total Hip Arthroplasty.* Journal of Arthroplasty, 2015. **30**(7): p. 1269-1272.

133. Del Balso, C., et al., *Taperosis Does head length affect fretting and corrosion in total hip arthroplasty?* Bone & Joint Journal, 2015. **97**(7): p. 911-916.

134. Higgs, G.B., et al., *Does Taper Size Have an Effect on Taper Damage in Retrieved Metal-on-Polyethylene Total Hip Devices?* The Journal of Arthroplasty, 2016. **31**(9, Supplement): p. 277-281.

135. Carlson, J.C.H., et al., *Femoral stem fracture and in vivo corrosion of retrieved modular femoral hips.* The Journal of arthroplasty, 2012. **27**(7): p. 1389-1396. e1.

136. Fraitzl, C.R., et al., *Corrosion at the Stem-Sleeve Interface of a Modular Titanium Alloy Femoral Component as a Reason for Impaired Disengagement.* Journal of Arthroplasty, 2011. **26**(1): p. 113-119.

137. Porter, D., et al., *Modern Trunnions Are More Flexible: A Mechanical Analysis of THA Taper Designs.* Clinical Orthopaedics and Related Research®, 2014. **472**(12): p. 3963-3970.

138. Hothi, H.S., et al., *Clinical significance of corrosion of cemented femoral stems in metal-on-metal hips: a retrieval study.* International Orthopaedics, 2016. **40**(11): p. 2247-2254.

139. Stamer, C.M., *Assessment of Bore-Cone Taper Junctions in Explanted Modular Total Hip Replacements*. 2015.

140. Kao, Y.-Y.J., et al., *Flexural Rigidity, Taper Angle, and Contact Length Affect Fretting of the Femoral Stem Trunnion in Total Hip Arthroplasty.* Journal of Arthroplasty, 2016. **31**(9): p. S254-S258.

141. Triantafyllopoulos, G.K., et al., *Otto Aufranc Award: Large Heads Do Not Increase Damage at the Head-neck Taper of Metal-on-polyethylene Total Hip Arthroplasties.* Clinical Orthopaedics and Related Research®, 2015: p. 1-9.

142. Lowe, D.G. *Object recognition from local scale-invariant features*. in *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*. 1999. Ieee.

143. Duval, L., et al. *Image processing for materials characterization: Issues, challenges and opportunities*. in *Image Processing (ICIP), 2014 IEEE International Conference on*. 2014. IEEE.

144. Nixon, M. and A.S. Aguado, *Chapter 12 - Appendix 3: Principal components analysis*, in *Feature Extraction & Image Processing for Computer Vision (Third edition)*. 2012, Academic Press: Oxford. p. 525-540.

145. Brown, J.D., *Choosing the right type of rotation in PCA and EFA.* JALT testing & evaluation SIG newsletter, 2009. **13**(3): p. 20-25.

146. Kim, P., *MATLAB Deep Learning*, in *MATLAB Deep Learning*. 2017, Springer. p. 103-120.

147. Landis, J.R. and G.G. Koch, *The measurement of observer agreement for categorical data.* biometrics, 1977: p. 159-174.

148. Higgs, M.G., et al., *Is Corrosion a Threat to the Strength of the Taper Connection in Femoral Components of Total Hip Replacements?* CORROSION, 2017. **0**(0): p. null.

149. Bishop, N.E., A. Hothan, and M.M. Morlock, *High friction moments in large hard-on-hard hip replacement bearings in conditions of poor lubrication.* Journal of Orthopaedic Research, 2013. **31**(5): p. 807-813.

150. Lavernia, C.J., et al., *Trunnion-Head Stresses in THA: Are Big Heads Trouble?* Journal of Arthroplasty, 2015. **30**(6): p. 1085-1088.

151. Jacobs, J.J., J.L. Gilbert, and R.M. Urban, *Corrosion of metal orthopaedic implants.* Journal of Bone and Joint Surgery-American Volume, 1998. **80A**(2): p. 268-282.

152. Langton, D., et al., *A comparison study of stem taper material loss at similar and mixed metal head-neck taper junctions.* Bone Joint J, 2017. **99**(10): p. 1304-1312.

153. Cook, S.D., R.L. Barrack, and A.J.T. Clemow, *CORROSION AND WEAR AT THE MODULAR INTERFACE OF UNCEMENTED FEMORAL STEMS.* Journal of Bone and Joint Surgery-British Volume, 1994. **76B**(1): p. 68-72.

154. Berg, H.E., *Will intelligent machine learning revolutionize orthopedic imaging?* Acta orthopaedica, 2017. **88**(6): p. 577.

155. Castaneda, C., et al., *Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine.* Journal of Clinical Bioinformatics, 2015. **5**(1): p. 4.

156. Osheroff, J.A., et al., *A roadmap for national action on clinical decision support.* Journal of the American medical informatics association, 2007. **14**(2): p. 141-145.

157. Hamet, P. and J. Tremblay, *Artificial intelligence in medicine.* Metabolism, 2017. **69**: p. S36-S40.