UNIVERSITY OF OULU

FACULTY OF INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING

Daniel Uher

OPTIMIZATION OF 3D TEXTURE ANALYSIS OF MR CARTILAGE IMAGES FOR PREDICTION OF KNEE OSTEOARTHRITIS

Master's Thesis
Degree Programme in Biomedical Engineering: Signal and Image Processing
December 2020

Uher D. (2020) Optimization of 3D Texture Analysis of MR Cartilage Images for Prediction of Knee Osteoarthritis. University of Oulu, Degree Programme in Biomedical Engineering: Signal and Image Processing, 59 p.

ABSTRACT

This thesis attempted to optimize a novel GLCM-based 3D Texture Analysis software in terms of its input parameters in order to maximize the early prediction of knee osteoarthritis from 3D DESS MR images. 20 subjects (10 control subjects; 10 progressor subjects) containing image data from baseline and from a 36-month-follow-up were extracted from the Osteoarthritis Initiative database and used as the study dataset. Multiple sets of 3D Texture Analysis were conducted incorporating 22 static and dynamic grey level quantization schemes, 6 bin quantization schemes and 4 offset settings. Cliff's delta was calculated to measure the effect size between the patient cohorts. Multilayer perceptron, Naïve Bayes and Support Vector Machines were implemented to classify the patients into their respective cohorts and estimate the robustness of the 3D Texture Analysis outputs. The predictions were done using only the baseline data, where all patients showed no signs of osteoarthritis. Maximum achieved robustness was 87%. The 3D Texture Analysis was found to have a high potential for the early prediction of knee osteoarthritis based on the GLCM features and the results outlined the importance of the software's input parameters.

Keywords: GLCM, DESS, machine learning, Cliff's delta, quantization

TABLE OF CONTENTS

ABSTRACT
TABLE OF CONTENTS
FOREWORD
LIST OF ABBREVIATIONS AND SYMBOLS
1. INTRODUCTION 6
2. THEORETICAL BACKGROUND 8
  2.1. Anatomy of the Knee Joint 8
  2.2. Osteoarthritis 9
  2.3. Magnetic Resonance Imaging 10
  2.4. Digital Image Processing 13
  2.5. Texture Analysis 15
  2.6. Machine Learning 18
3. METHODS 22
  3.1. Subject Dataset 22
  3.2. Cartilage Histogram Analysis 24
  3.3. 3D Texture Analysis Tool 25
  3.4. Applied 3D Texture Analysis 27
  3.5. Statistical Analysis 29
  3.6. Machine Learning Analysis 30
4. RESULTS 34
  4.1. Cartilage Histogram Analysis 34
  4.2. Applied 3D Texture Analysis 36
  4.3. Statistical Analysis 36
    4.3.1. Grey Levels 38
    4.3.2. Bins 38
    4.3.3. Offset 40
  4.4. Machine Learning Analysis 43
5. DISCUSSION 47
6. CONCLUSION 52
7. REFERENCES 53

FOREWORD

I would like to hereby express my deepest gratitude to Victor Casula and Ari Väärälä for their undying patience and wonderful support during the entire thesis process. Their leadership and guidance is what made this thesis a very special and unforgettable experience. The whole project provided me with a big insight into the research of knee osteoarthritis and allowed me to not only deepen my skills in a plethora of areas from image processing up to machine learning, but also work with a great team of researchers and experience the struggles and joys of a long-term research project.

I would like to also dedicate a special thank you to Antti Isosalo for his much appreciated insights and guidance through the world of machine learning and providing his own expertise with various learning algorithms, which were, as a result, applied in this project.

Lastly, a big thank you to everyone who directly or indirectly participated in the thesis, from data collection over cartilage segmentation up to sharing open source tools across the internet or providing help with the text editing. All of those were crucial to finalize this thesis and contributed significantly into making this a highly enjoyable and exciting research experience.

Oulu, December 7th, 2020

Daniel Uher

LIST OF ABBREVIATIONS AND SYMBOLS

2D	two-dimensional
3D	three-dimensional
BCI	bone-cartilage interface
CPU	central processing unit
CTRL	control cohort
DESS	double echo steady state
FISP	fast imaging with steady-state precession
GE	gradient echo
GLCM	grey level co-occurrence matrix
GPU	graphics processing unit
KL	Kellgren-Lawrence
L10	cartilage layer at 10% height of the cartilage thickness
L50	cartilage layer at 50% height of the cartilage thickness
L90	cartilage layer at 90% height of the cartilage thickness
MRI	magnetic resonance imaging
OA	osteoarthritis
OAFI	Osteoarthritis Foundation International
OAI	Osteoarthritis Initiative
PRGS	progressor cohort
PSIF	reverse FISP
RF	radio-frequency
ROI	region of interest
SE	spin echo
SNR	signal-to-noise ratio
SUM	cartilage layer representing the full cartilage thickness
TA	texture analysis
TE	time to echo
TR	repetition time
δ	Cliff's delta
$\sum$	summation
N_g	GLCM size corresponding to the number of quantization bins

1. INTRODUCTION

Knee osteoarthritis (knee OA) is one of the most troubling medical conditions plaguing the world's population. A disease considered to be one of the leading causes of disability, though no cure has yet been developed. Medications and drugs only target the symptoms, however the cause (or causes) of the knee OA remain a mystery. There are plenty of factors known to be contributing to the development of knee OA, including age, weight, metabolic processes, environment, etc., however there is no standardized map outlining the osteoarthritis timeline and the severity of the impact caused by various factors. A significant emphasis is made regarding the prevention by adopting a healthy lifestyle, however some degree of osteoarthritis seems to be inevitable for the vast majority of people throughout their lifetime. [1, 2, 3, 4]

In recent years, researchers have been utilizing Magnetic Resonance Imaging (MRI) to visualize and study the articular cartilage and the subchondral bone. MRI provides yet unsurpassed detail and resolution of the cartilage structure, plus is capable of generating three-dimensional (3D) images. Such images are considered to be theoretically bearing specific markers, which might indicate an inclination towards osteoarthritis, which is yet invisible to the human eye. Therefore, various quantitative methods were established to extract texture information from the MR images and attempt to differentiate between subjects with and without various degrees of osteoarthritis progressions. One of the most well-known methods is the calculation of Grey Level Co-occurrence Matrices (GLCMs) and the resulting GLCM features, which provide a quantitative insight into the spatial distribution of the pixel intensities within the given image. The GLCM approach was originally developed in the 1970's for analyzing aerial photography, however various research projects have shown promising results for utilizing GLCM features for the detection of knee osteoarthritis. [5, 6, 7, 8]

In 2018, a novel method for calculating the GLCM features from 3D isotropic MR data called 3D Texture Analysis was developed by Ari Väärälä at the Research Unit of Medical Imaging, Physics and Technology, University of Oulu. This novel algorithm utilizes the construction of the well-known grey level co-occurrence matrices, however their calculation is based on a unique way of interpolating and extrapolating the 3D MR data. Since the studied software is based on an entirely new methodology, there are no prior roadmaps on what to expect, except of some preliminary results and the general knowledge about the GLCMs provided by the referenced research projects. [9]

The software has 4 input parameters, which are of crucial importance for the GLCM calculation: *minimum grey level, maximum grey level, bin quantization number* and *offset*. The core of this thesis lies within the presumption that the studied 3D Texture Analysis input parameters might have a significant impact on the analysis output. However, the size of the impact is the topic of this thesis. Therefore, in order to gain a clearer image about the software, multiple sets of 3D Texture Analysis with varied input parameters were conducted on both symptomatic and asymptomatic subject images derived from a large longitudinal study and the output features were studied and evaluated. [9, 10]

Although the four parameters are crucial for determining the structure of the grey level co-occurrence matrices, only a handful of researchers actually report their selected values and further discuss them. Several studies pointed out the lack of standardization in terms of the GLCM parameters and call for a study focusing on the resulting differences in the outputs features caused by the changes in the input parameters. [11, 12, 13]

A study sample of 20 subjects including image data from two timepoints (baseline and 36-month follow up) was extracted from the Osteoarthritis Initiative database, which is an ongoing longitudinal study of OA. 10 subjects were extracted from the control cohort, which contains subjects without any sign of osteoarthritis throughout the entire duration of the study. The other 10 subjects came from the incidence and progression cohorts, which contain subjects who actually developed the disease since the baseline screening. [14, 15]

The methodology has been divided into **statistical analysis** and **machine learning analysis**. Statistical analysis aims to provide a better idea about how well the output features can differentiate between the subject cohorts in terms of effect size. The subsequent machine learning analysis, on the other hand, utilizes some of the findings from the previous statistical results. It provides more of a practical evaluation and determines the robustness of the collected 3D TA outputs in terms of the OA prediction. A complex machine learning pipeline including three machine learning algorithms and various additional approaches was utilized. The prediction was based on knee images with no signs of osteoarthritis. Some knees would go on and develop OA and others would not. However, the machine learning approach does not utilize the image data from the subsequent screenings and the prediction is based merely on the baseline image data.

This thesis addresses two questions: 1) is it possible to use the 3D Texture Analysis software to predict knee osteoarthritis purely from the baseline data? 2) Does altering the input parameters affect the predictive performance of the calculated features? Since each knee is of unique anatomy and attributes, a single set of recommended input parameters does not seem viable. Therefore, the goal of this thesis is to provide an idea about how the input parameters affect the output features of the 3D Texture Analysis in terms of osteoarthritis prediction and establish recommendations, which might hopefully provide some additional guidance for the researchers focusing on GLCM-based knee osteoarthritis predictions.

2. THEORETICAL BACKGROUND

2.1. Anatomy of the Knee Joint

The knee is the largest joint within the human body. It is encapsulated within a joint capsule and held together by Collateral ligaments and Patellar ligament, which is an extension of the Quadriceps femoris tendon. Femur, Tibia and Patella are the three bones present within the knee joint. In some literature, Fibula might also be included, however Fibula is not part the joint itself, because it attaches to the proximal part of Tibia. Tibiofemoral and Patellofemoral joints create the knee mechanism (Figure 1). The basis of each joint is the articular cartilage, which serves as gliding surface for the touching bone ends. The bone surface lying directly below the cartilage is referred to as subchondral bone. Joint cavity is the small space between the cartilages and contains the synovial fluid, which ensures a resistance-free joint movement. In order to help the cartilage protect the knees from various pressures, a fibrocartilaginous structure (Meniscus) acts as a shock absorbent by distributing the outer forces towards various directions and basically acts as a cushion for the tibiofemoral joint. [16, 17]

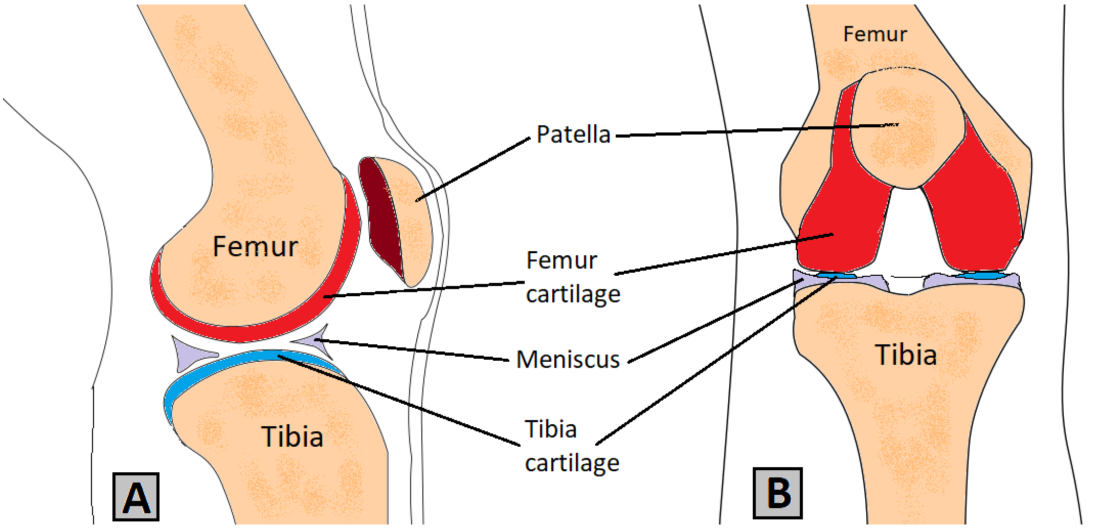


Figure 1. The anatomy of the tibiofemoral joint. A) Side view of the knee; B) Frontal view of the knee.

The cartilage tissue is quite unique due to its complete lack of blood vessels, lymph vessels and nerves. The main determinant of its mechanical properties (durability, stiffness, etc.) is its extracellular matrix (ECM), which is basically the scaffolding of the cartilage. The ECM consists mainly of water (approximately 80%), collagen (10-30%) and proteoglycans (3-15%). Their properties and concentration vary across the cartilage thickness. The amount of water is higher at the surface of the cartilage, meaning that the cartilage gradually becomes more dense towards the subchondral bone. There are specialized cells called chondrocytes distributed across the ECM in no particular order. ECM provides an optimal environment for the chondrocytes, which in return synthesize and regenerate the extracellular matrix. There is a constant fluid movement between the chondrocytes and the synovial fluid in order to ensure the delivery of proper nutrients and the removal of molecular waste. [16, 18, 19]

2.2. Osteoarthritis

Osteoarthritis ("osteo" = bone; "arthritis" = joint disease) is a degenerative joint disease and is considered to be one of the leading causes of disability for the adult population. According to the Osteoarthritis Foundation International (OAFI), over 300 million people worldwide are suffering from some kind of osteoarthritis and as a result experiencing a significantly lower quality of life. The diagnosis and treatment of osteoarthritis has been for decades one of the most researched topics in the medical field. The main symptoms of OA include joint pain and joint stiffness. Therapies like intra-articular injections are common, however, such therapy usually numbs the pain without any sign of helping to regenerate the missing soft tissue or to improve the state of the joint bones. [1, 2, 9, 20]

Two general types of OA are universally recognized. Primary OA is a result of articular degradation of the cartilage. There are no particular or specific reasons to be credited for the continual loss of soft tissue. There are merely known factors, that are believed to influence the rate with which the cartilage dissolves. Primary OA is the most common type of osteoarthritis. Secondary OA is caused by an inflicted trauma, accident, etc. [1, 21]

The OA research has shown that multiple crucial factors are contributing to the development of the disease such as mechanical forces, inflammation, metabolic processes, etc. Apart from the degeneration of the cartilage tissue, bone lumps called osteophytes start to form usually at the subchondral bone. As a result of the reduction of cartilage thickness, osteophytes start to peak through the cartilage and cause a direct contact between the bones themselves. Such bone-on-bone contact is a major pain inflictor and a cause for knee stiffness. In more advanced stages of the OA progression, the remodeling of the bone becomes quite significant along with the narrowing of the cartilage space. [20, 22]

Knee osteoarthritis is the most common type of OA. The probability of developing knee OA increases with age. The significant lifelong stress combined with metabolic, environmental, genetic and inflammatory impacts cause the cartilage and bone to degenerate over time. Even though knee OA is one of the most researched diseases, it is still significantly unknown. There are many question marks in terms of both diagnosis and treatment. The absolute majority of treatment methods are focused on reducing the pain and therefore only delay the probable surgical action. Arthritis foundations put a high emphasis on proper prevention and joint care. Moderate exercise is strongly recommended. Exercise with low joint stress (swimming, yoga, etc.) should be prioritized over heavy lifting. Also, a healthy diet and weight control play important roles and are recommended. The progression is so far understood as irreversible and therefore the research focuses highly on early diagnosis. [1, 3, 4]

Kellgren-Lawrence score (KL score), a scoring system developed in 1957 by J. H. Kellgren and J. S. Lawrence, aims to evaluate the OA progression based on medical images. The range of the KL score is divided into 5 grades; Grade 0 - no visible presence of OA; Grade 1 - debatable narrowing of the joint space; Grade 2 - Formation of osteophytes, joint space probably narrower; Grade 3 - definite decrease in the joint space volume, definite presence of osteophytes; Grade 4 - a severe case of OA, bone deformation with significantly reduced joint space. KL = 2 is officially considered osteoarthritis. [23, 24]

Magnetic Resonance Imaging has been used in recent years to study and visualize the joint structure due to its great ability to differentiate between the bone and cartilaginous tissue in great detail.[6, 25]

2.3. Magnetic Resonance Imaging

Magnetic resonance imaging (MRI) is an imaging modality used in clinical diagnostics to visualize the anatomy and physiology of the subject. At this moment in time, MRI is the best modality for the articular cartilage visualization both in 2D and 3D, and therefore is very suitable for monitoring and diagnosing the knee osteoarthritis. [6, 5, 26]

The history of MRI reaches back to 1920s to the studies of Nils Bohr and Arnold Sommerfeld focusing on the discrete magnetic moments of particles. A decade later in 1938, Isidor Isaac Rabi et al. from Columbia University and Huner College in New York published an article describing a method of measuring the magnetic properties of individual atoms by utilizing a focused electromagnetic beam to re-orient the magnetic moment and nuclear spin.[27, 28, 29, 30]

It was, however, in 1946 when two teams of researchers, one team from Massachusetts Institute of Technology lead by Edward Mills Purcell and the other team from Stanford University lead by Felix Bloch, experimentally described the nuclear resonance phenomenon in both solids and liquids and thus laid the grounding stone for what is known today as Magnetic Resonance Imaging. [30, 31, 32]

Decades later, in 1970's, Paul Lauterbur developed a method to obtain images based on the local magnetic interactions. In 1977, the first MRI machine was built and the first human subject was screened. Over the past decades, MRI has established itself as one of the best screening modalities and is unsurpassed in terms of visualizing soft tissues within body. To sum up the history section with an interesting trivia; in 1988 shortly before he passed away, Isidor Isaac Rabi himself underwent a screening in an early MRI machine, to which he said: "I never thought my work would come to this." [33, 34, 35]

The main part of the MR machine is the magnetic coil wrapped around the subject bed introducing a horizontal magnetic field flowing through the tunnel. The machine size is described by the strength of the magnetic field. Usually, 1.5 Tesla to 3 Tesla machines are available in clinical practice. The quality of a MR machine reflects the homogeneity of the magnetic field. Non-homogeneous field may result in various artefacts. [5]

The basic principle of MRI (Figure 2) lies within the atomic nuclei. In clinical practice, hydrogen protons are the most common, due to the high percentage of

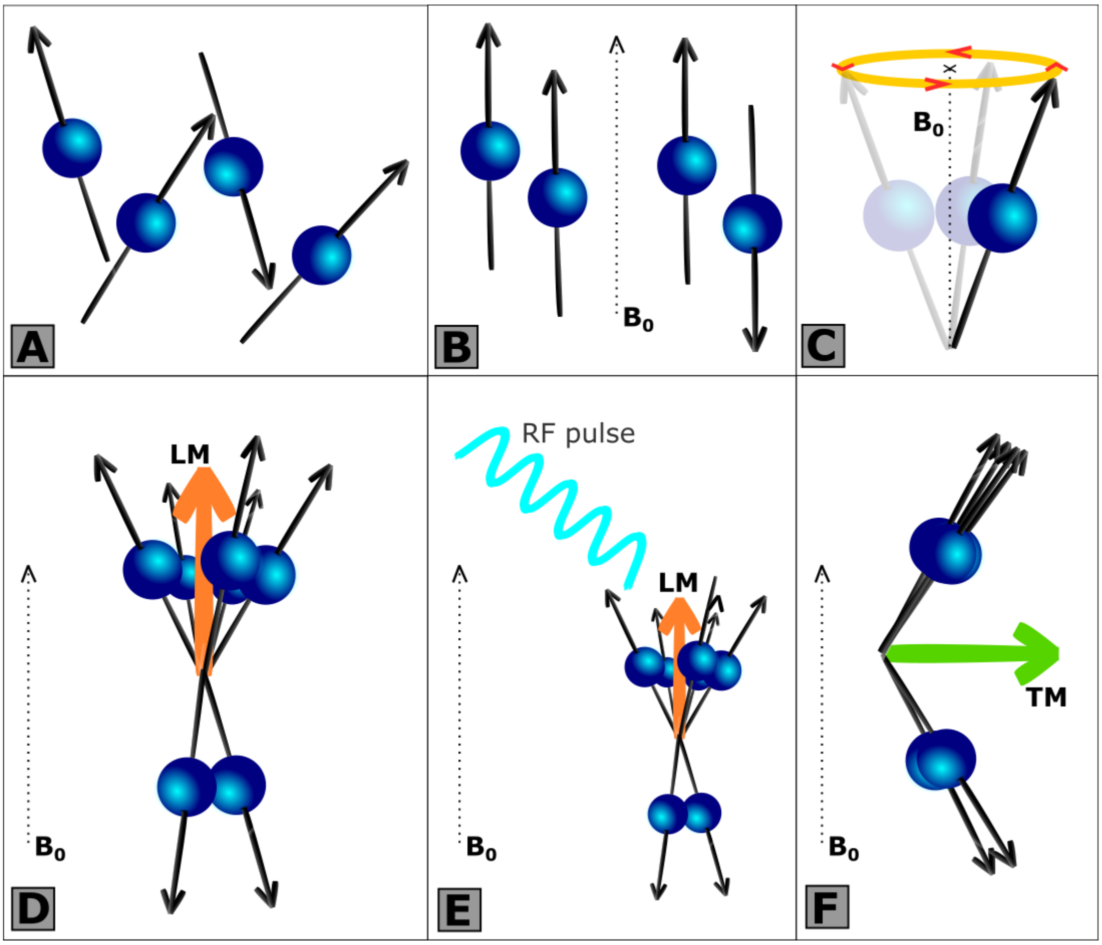


Figure 2. Visualization of the physical phenomena occurring during the MR screening. The blue balls represent hydrogen protons and the black arrows illustrate their magnetic moment. A) randomly scattered hydrogen protons in the absence of external magnetic fields; B) Alignment of the hydrogen protons either against (high-energy protons) or along (low-energy protons) the external magnetic field B_0; C) Analogy of the nuclear spin. The protons, once aligned, precess around the external magnetization; D) For illustrative purposes, the protons were plotted starting from the same base. The orange arrow signifies a longitudinal magnetization - sum of all the aligned magnetic moments; E) The introduction of a radio-frequency pulse (RF pulse) onto the protons. The RF pulse has the same frequency as the proton precession frequency. F) Upon the RF pulse excitation, a transversal magnetization (green arrow) is created as a result of flipping some of the protons into their high-energy states and aligning the phases.

water within a human body. Each hydrogen proton spins[1] around its axis similarly to a spinning top. This constant spin orients the direction of the proton magnetic moment. However, the orientation of each proton is randomly scattered (Figure 2A). By introducing the protons into a magnetic field B_0, the orientations of the magnetic moments align either along (low-energy state) or against (high-energy state) the direction of the external magnetic field and thus creating a clearly oriented *longitudinal magnetization* (2B). The nuclear precession frequency can be obtained by the Larmor equation [5, 9]:

$$\omega_0 = \gamma \times B_0 \tag{1}$$

where ω_0 is the Larmor frequency of the proton, γ corresponds to the gyromagnetic ratio, which is constant specific for any type of nuclei (Hydrogen nuclei have $\gamma = 42{,}58$ MHZ/Tesla), and B_0 refers to the magnet strength.

Up until this point, the nuclei are aligned in terms of the directions, however their precessions are out of phase. To obtain measurable data, a radio-frequency pulse with a frequency equal to the Larmor frequency is introduced (Figure 2D) and disrupts the aligned equilibrium by flipping some of the low-energy protons into their high-energy states and aligning the precessional phases (Figure 2E). As a result, a *transversal magnetization* vector perpendicular to B_0 gets established (Figure 2F). The transverse magnetization can be measured with a receiver coil. The system starts to relax and return to equilibrium after the RF pulse excitation. The time between adjacent RF pulses is marked as Repetition Time (TR). As the transversal magnetization slowly decays, the measurable signal grows smaller in amplitude. The time between the RF excitation and measurement of the signal is called Time to Echo (TE). The return of some protons to low-energy state from high-energy state (rebirth of the LM) results in heat dissipation into the surrounding tissue. The time it takes to annihilate the transversal magnetization is referred to as T_2 relaxation time, or spin-spin relaxation. In return, the time to recover the longitudinal magnetization is called T_1 relaxation time, or spin-lattice relaxation. The lengths of T_1 and T_2 relaxation times depend on the tissue properties. T_1 and T_2 create the contrast differences, however TR and TE can be adjusted to accentuate contrasts between the desired types of tissue. T_1-weighted images best depict anatomical details. On the other hand, T_2-weighted images are best to depict pathologies and are great to differentiate between various tissues which have high water content. [5, 9, 26, 37]

The received signal needs to be spatially located and encoded. Short-term linear inhomogeneities of the magnetic field called *gradients* are created by gradient coils across all three axes in order to localize the echoing signal. The gradient pulses are responsible for the typical loud "bangs" during the MR screening process. All the collected signal responses compose the raw MRI data and the locations and amplitudes are encoded into a 2D k-space matrix, which is a matrix of the same size as the final image. The k-space matrix consists of complex values, where each

[1] This is merely an analogy to describe the behavior of subatomic particles. In reality, the particles are not actually spinning according to the traditional meaning. However, the quantum understanding of a dipole is analogous to a description of spinning object by classic mechanics. [36]

pixel stores information about the spatial frequency and phase and contributes to the whole final image. To generate the real image, a discrete inverse Fourier transform reconstructs the human-readable MR image from the k-space matrix. [9, 5, 37, 38]

Various MR sequences are used to create different contrasts between tissues and fluids. Each sequence consists of carefully placed RF pulses and gradient pulses. All the available sequences are based on two essential sequence families: spin-echo (SE) and gradient-echo (GE or GRE). SE sequences use two RF pulses (usually $90°$ for exciation + $180°$ for refocusing), whereas GRE sequences use a single RF pulse with variable flip angle and image gradients to dephase and rephase the echoed signal. The main benefit of GRE is their very short TR, however they are very susceptible to errors caused by the magnet inhomogeneities. [5, 37]

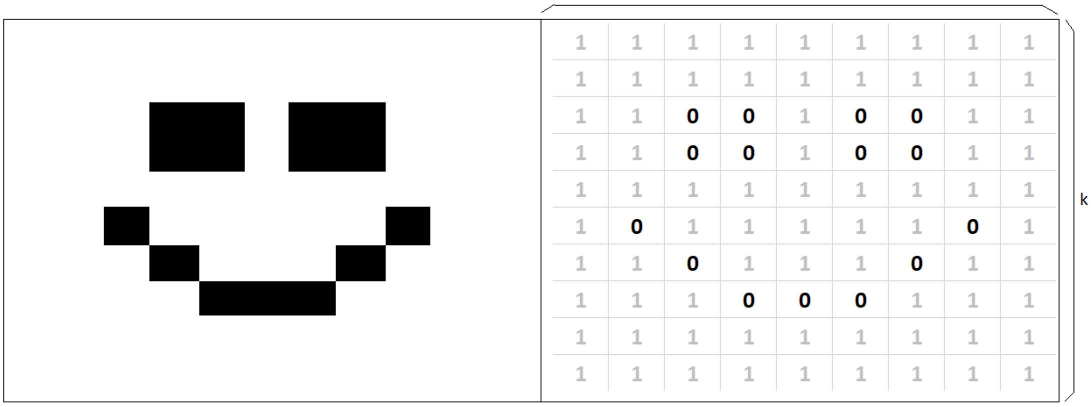
3D DESS (Double Echo Steady State) is a GRE sequence patented by Siemens and is a direct result of the effort by Bruder et. al. from 1988 [39] to produce two separate MR images with different contrasts at the same time and combine them into one. The first part is FISP-like (fast imaging steady precession) with high T_1/T_2 ratio contrast and provides a good morphological detail of the cartilage structure. The second part is based on reversed FISP (PSIF), which is responsible for a good resolution between various fluids due to its T_2-weighted contrast. 3D DESS basically combines the T_1-weighted intra-cartilaginous detail while providing the great T_2-weighted contrast between cartilage and synovial fluid. The sequence is however very sensitive to motion and, as a GRE sequence, struggles with the magnetic inhomogeneities. 3D DESS provides improved SNR over traditional methods, better contrast between the cartilage and synovial fluid and high isotropic resolution of the formed images. [5, 9, 26, 39, 40]

2.4. Digital Image Processing

Every two-dimensional (2D) digital image is merely a $m \times k$ matrix of numerical values, where m is the number of horizontal pixels and k is the number of vertical pixels constructing the image (Figure 3). Each pixel value determines the intensity of the given pixel. The pixel data is then described as the texture. By applying various algorithms onto the texture i.e. pixel values, various texture features can be extracted in order to precisely describe, categorize or classify the image.

A binary image, which means that the pixel values can only be either 0 or 1, is depicted in Figure 3. In this case, 0 indicates black color and 1 indicates white. A binary image can be described as a **1-bit** image. A bit is the smallest carrier of digital information and it can either be 0 or 1, true or false. So, if only 1 bit determines the value of each pixel, such value can only be 0 or 1. Expanding this logic, in a n-bit image, each pixel is encoded by n number of bits and since each bit provides 2 possible pixel values, the color spectrum increases into 2^n tonal values ranging from 0 up to $n - 1$. To visualize this more clearly, Figure 4 shows how the encoding of the grey scale works.

In slightly different words, n-bits divides the color space between black and white into 2^n color values. For example, in a grey scale 3-bit image, there would be $2^3 = 8$ different grey tones encoded. Similarly, for an 8-bit image, $2^8 = 256$ intensity values



1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1
1	1	0	0	1	0	0	1	1
1	1	0	0	1	0	0	1	1
1	1	1	1	1	1	1	1	1
1	0	1	1	1	1	1	0	1
1	1	0	1	1	1	0	1	1
1	1	1	0	0	0	1	1	1
1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1

Figure 3. Numerical representation of a binary digital image

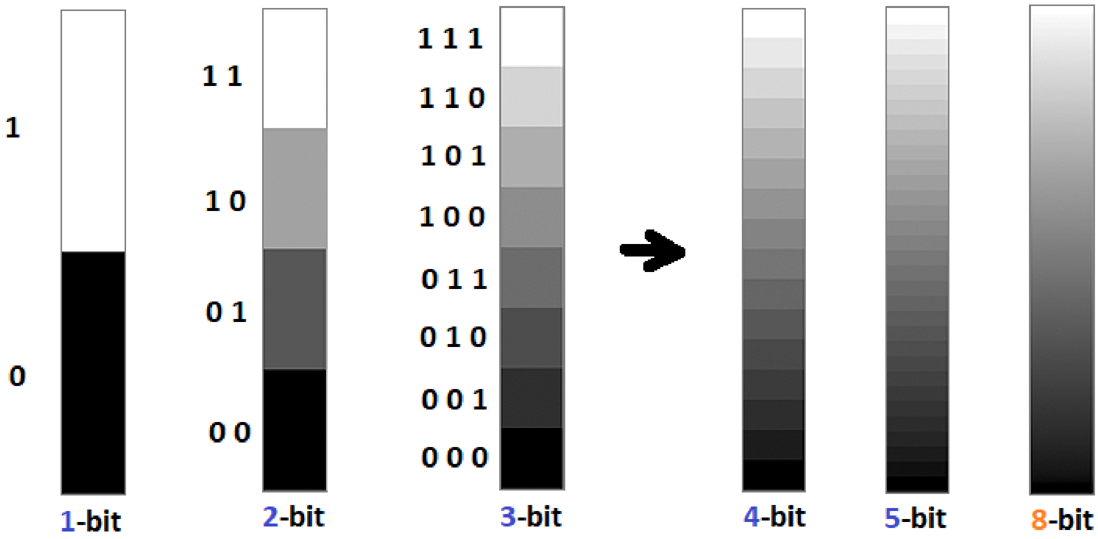


Figure 4. Encoding grey scale images according to the number of bits

would be available and so each pixel value could be anywhere between 0 and 255. Higher bit number ensures better color resolution but the information size of the image increases and more disc space and computational power is required to work with such images. Medical images are usually encoded with 12 to 16 bits per pixel. Such bit depth provides between 4 096 - 65 536 distinguishable grey tones. [41]

This brief introduction into digital computing is included not only to revise the basics, but most importantly to transfer the reader into a slightly more digital mindset and evoke the understanding that studying digital images is nothing but a game of numbers. Every image is a simple matrix of numerical values which can be mathematically described and analyzed. And such analysis, although seemingly abstract, might extract some information from the image which are hidden to the human eye.

2.5. Texture Analysis

Texture analysis could be considered in itself a sub field of the image analysis research branch, which has been linked to a great diagnostic success in the field of medicine. The term texture analysis can be understood as a method of describing images or regions of images by extracting their specific texture features based on the spatial distribution of the pixel data. [42]

To fully understand what exactly is meant by texture analysis and why it is done at all, let us consider a situation, where we have visited an art exhibition and afterwards we want to describe the paintings to our friends. In order to describe the memorized image, we put the memory in front of our eyes and start describing the attributes of the image, which we believe are important and clearly recognizable. In other words, we are trying to describe the texture of the image by extracting its texture features, most probably the ones we believe are the most significant and will result in a quick and easy recognition by our friends. In the case of a painting, the significant features might include the colors, size, style, technique, etc. Another example might be a description of a human face. In such case, the features might include eye color, beard, hair length, smile, etc. Essentially, texture analysis helps us to summarize an image in terms of its apparent features so it is as recognizable as possible without the need to describe every single detail.

In computing, the idea of texture analysis stays the same, however the extracted features might not be as clearly representable visually, but merely statistically and numerically. This also gives the opportunity to study and compare various texture features, that might have significant impact on the classification process. Computers cannot see and recognize their surroundings the same way humans do, therefore mathematical representation is necessary for the computer to classify the image.

If we transpose this logic into the world of computational medicine, texture analysis has a tremendous potential to assist with the diagnostic process and most importantly uncover aspects of the subject images that might be simply invisible to human perception. Texture analysis has had a substantial success with not only segmentation of anatomical structures, but also with diagnosis of lesions, differentiating suspicious tissues and many more. [42]

One of the well-established methods for texture analysis is a **Grey-Level Co-occurrence Matrix (GLCM)** proposed by Robert M. Haralick in 1973 [7]. A **GLCM contains counts of how many times a pair of pixel intensities appeared within a quantized image**. As such, GLCMs provide a quantitative information about the spatial distribution of pixel intensities. [8]

A GLCM is a simple $N_g \times N_g$ square matrix, where N_g is the number of grey levels represented in the quantized image, i.e. the bin quantization number. The calculation of GLCMs has a few input parameters, which significantly influence the output. The first step to construct the GLCM matrix is the quantization (or "binning") of the image. The image quantization refers to a process, where all the intensities present within the image are put (i.e. quantized) into a selected number of bins. If, for example, we choose to do an 8-bin quantization, then all the pixels will be assigned values between

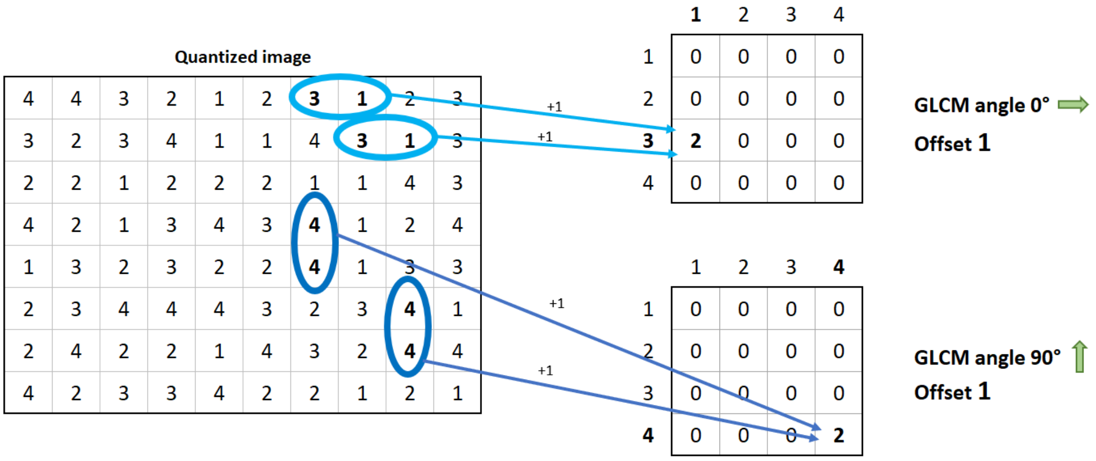
1 - 8 according to their original pixel intensities. An $m \times k$ image with pixel intensities ranging from 0 to 255 (8-bit image) would result in $256/8 = 32$ intensity values per each bin. As a result, pixels with intensities 0 - 31 would be assigned intensity 1 in the quantized image; pixels 32 - 63 would get the value 2 and so on. It is possible to determine the grey level range, in which the quantization is being calculated. This is done by selecting the minimum and maximum grey level. Intensities with values outside of the minimum-maximum grey area can be either omitted or assigned to a certain bin as well.

GLCMs can be calculated for various angles (or "directions") and offsets. Each pixel has 8 neighbouring pixels, which results in 8 possible angles for the GLCM calculation (Figure 5). The only exceptions are the edge and corner pixels.

135°	90°	45°
180°	*	0°
225°	270°	315°

Figure 5. Possible directions described as angles for the GLCM calculation. The * represents the root pixel.

Figure 6 shows an example of the construction of GLCMs for angles 0° and 90° out of a 4-bin quantized image. The GLCM is always a square matrix and its dimension is determined by the number of bins. To fill the matrix, an algorithm looks for pixel pairs, counts how many times they appear within the quantized image and adds the count to the GLCM. The corresponding row in the GLCM is determined by the root pixel value and the column is determined by the value of the selected neighbour.



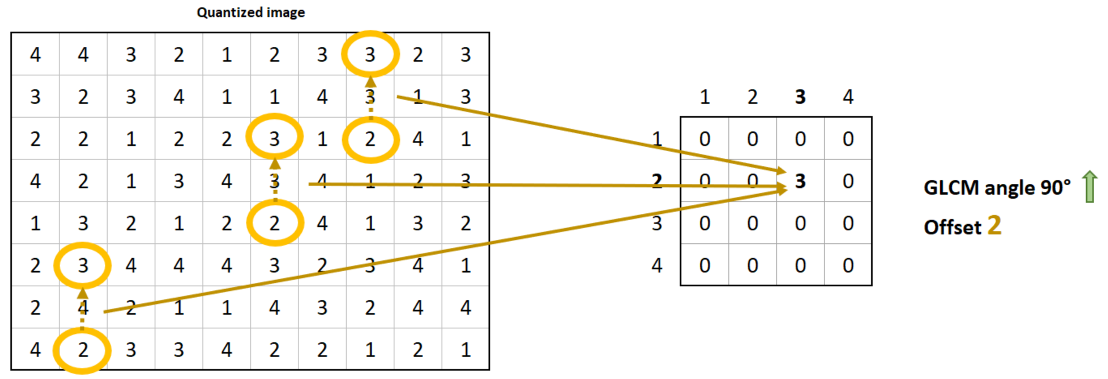
4	4	3	2	1	2	3	1	2	3
3	2	3	4	1	1	4	3	1	3
2	2	1	2	2	2	1	1	4	3
4	2	1	3	4	3	4	1	2	4
1	3	2	3	2	2	4	1	3	3
2	3	4	4	4	3	2	3	4	1
2	4	2	2	1	4	3	2	4	4
4	2	3	3	4	2	2	1	2	1

	1	2	3	4
1	0	0	0	0
2	0	0	0	0
3	2	0	0	0
4	0	0	0	0

	1	2	3	4
1	0	0	0	0
2	0	0	0	0
3	0	0	0	0
4	0	0	0	2

Figure 6. Visual representation of constructing the GLCM for offset 1 and angles 0° and 90° from a 4-bin quantized image.

The offset determines the spacing of the considered pixels. In other words, the offset number means which neighbouring pixel should be considered along the selected angle. Offset 1 takes the nearest neighbouring pixel; Offset 2 takes the second pixel behind the nearest one and so on. Figure 7 depicts the impact of offset 2 on the overall construction of the matrix.



4	4	3	2	1	2	3	3	2	3
3	2	3	4	1	1	4	3	1	3
2	2	1	2	2	3	1	2	4	1
4	2	1	3	4	3	4	1	2	3
1	3	2	1	2	2	4	1	3	2
2	3	4	4	4	3	2	3	4	1
2	4	2	1	1	4	3	2	4	4
4	2	3	3	4	2	2	1	2	1

	1	2	**3**	4
1	0	0	0	0
2	0	0	**3**	0
3	0	0	0	0
4	0	0	0	0

Figure 7. Example of a GLCM construction with offset 2.

Haralick et. al. [7] defined 14 texture features, which can be extracted from the GLCMs. The texture features are calculated from a normalized GLCM, which is the original GLCM divided by the summation of its elements. This means, that the normalized GLCM can be viewed as a matrix of probabilities, i.e. how likely it is that such grey pixel pair is to appear within the image. All the features describe the image texture in a quantitative way and it is presumed that they might have an indicative value in terms of detection of early OA [8, 9, 10, 11].

There are many research projects ([10, 11, 12, 13, 43, 44, 45, 46, 47, 48, 49] and many more) utilizing the GLCM features in the biomedical field. More specifically, various knee osteoarthritis studies utilized GLCMs to a great success [12, 45, 46, 49]. In 2014, Schooler et. al. [10] directly showed good indicative abilities of GLCMs for cartilage degradation. Unfortunately, only a fraction of the studies actually report the GLCM input parameters used during their research and even less underline their importance and how they might affect the output texture features.

Gomez et. al. [13] studied breast ultrasound images and used GLCM features to analyze segmented lesions. Gomez's study is the only one found which details the approach of minimum and maximum grey level assignment. Each lesion was normalized between grey values 0 and 255 in order to stretch (or reduce) the grey scale of the image within the same boundaries. This approach allowed them to use 0 as the minimum grey level and 255 as the maximum grey level and thus cover all the grey tones within all the images. In terms of bins, 8, 16, 32, 64, 128, and 256 were tested and the team reported, that bin quantization does not improve nor worsen the results.

Brynolfsson et. al. [11] considered the importance of the input parameters by testing bin counts of 4, 8, 16, 32, 64, 128 and 256 on ADC MR images. They concluded to a suggestion to keep the bin number static and reported large changes in the quantitative results based on the different bin quantizations. Brynolfsson's study is also the only one found that reports a conclusion in terms of the grey level boundaries and suggests static minimum and maximum grey level, which would encapsulate all the possible intensities found across the cohort. Brynolfsson's team also discussed the lack of standardization of the input parameters and urges researches to report their chosen parameters.

Peuna et. al. [12] also discussed the lack of information about the GLCM parameters and how exactly they might affect the outcome. Peuna's research implemented offsets 1-4 and 8-bin quantization in their GLCM calculations for evaluating a novel method of GLCM calculation by cartilage flattening. The results from various offsets were found to be similar. In terms of the bin number, in spite of statistically significant results, a call was made for a further optimization of the bin quantization scheme in future studies.

Several publications [10, 45, 47, 48] were found to be using exclusively offset 1 (the nearest neighbour co-occurrence) for the GLCM calculations, however they did not conduct any further offset investigation.

Blumenkrantz et. al. [49] studied 3D spoiled GE images of the knee and utilized GLCM calculations to evaluate cartilage T_2 values. The team implemented offsets 1-3 based on the approximate cartilage thickness being 3-4 pixels. Their conclusion showed equally good results for all of the studied offset settings and demonstrated strong positive correlations between them. The study reports a shortcoming of a low image resolution, which could have an impact in terms of the offset setting.

Li et. al. [50] studied the relationship of MR relaxation times and knee osteoarthritis. Their study included GLCM calculations from various 3T MR images with offsets 1-3 due to the cartilage thickness being 3-4 pixels. The paper reports all offsets showing similar results, which supports findings by Blumenkrantz et. al. [49].

Williams et. al. [46] showed that short-term (6 months) evaluation of T_2 mappings and selected texture features may provide an early anticipation of the cartilage degeneration. In their paper, offsets 1, 3 and 5 were used to study *contrast, homogeneity, correlation* and *energy*, however their results remain inconclusive in terms of the offset setting.

Materka et. al. [51] demonstrated that the GLCM texture features might be sensitive to inhomogeneities in MRI and therefore recommend artifact removal and/or image normalization prior to the texture analysis. They also point out a good resolution and large volumetric data from the ROI should have a positive impact on the analysis.

2.6. Machine Learning

Machine learning (ML) is a computing approach to learn the inner patterns and relationships based on empirical data and as a result establish decisions based on the learnt patterns. It can be understood as a computational approximation of the human learning process and the subsequent application of the learnt knowledge. Machine learning is concerned with utilizing algorithms to make predictions based on collected data. [52]

The machine learning process generally consists of three phases: **Pre-processing**, **training**, **validation** and **testing**. [52]

- **Pre-processing** refers to data preparation, feature extraction and feature selection. The data preparation involves 1) filtering incomplete, missing or noisy samples; 2) merging all the multiple sources if applicable; and 3) data normalization, i.e. fitting the feature values into a given range. Feature extraction is the process of extracting the features from the collected data. The raw data can serve as a training set or there is a further analysis applied in order to extract additional information about the dataset. Once the features are extracted, feature selection methodology takes into account either all of the features or merely their subset. The selection can be justified by various means such as statistical analysis, or by utilizing a more complex methods for dimensionality reduction, such as Principal Component Analysis (PCA). The resulting feature vectors are divided into a training set and a testing set. [52, 53]

- **Training** is the learning phase. Essentially, training a model means to establish a function, which is responsible for producing the desired outputs based on the available inputs. Theoretically, there exists a function, which predicts the labels flawlessly based on the input data and the machine learning algorithm simply attempts to approximate it as closely as possible. Once the training set is ready, a machine learning model is applied onto the training set and adjusts its properties accordingly in order to understand its patterns. The learning can follow three main paradigms: *supervised*, *unsupervised* or *reinforcement* approach.
  - *Supervised learning* relies on prior knowledge of the correct output. The goal of the algorithm is to adjust its properties so that the output of the algorithm matches the desired labels. Classification and regression are the two main types of supervised learning. Classification provides categorical outputs (for instance, differentiating if an image is a cat or a dog) while regression approximates the input data in order to predict a real value (for example, temperature prediction). [52, 53, 54]
  - *Unsupervised learning* has no prior knowledge about the labels and its goal is to identify the structural patterns of the input data. A typical example of unsupervised learning is clustering. The algorithm attempts to divide the dataset into a number of groups (clusters) based on the data distribution. [52, 53, 54]
  - *Reinforcement learning* is based on trial and error methodology by assigning a reward or a penalty based on the learnt outcome [52]. A great example of reinforcement learning are the open-source AI Pac-Man projects provided by the UC Berkeley [55]. The projects are based on Pacman independently learning how to get to the desired dot through the complicated maze while avoiding the evil ghosts.
- **Validation** is utilized to improve the trained model. *K-fold cross-validation* is a great example of validation. The k-fold method takes the training dataset and splits it into *k* number of chunks. For each learning epoch, one of the chunks is assigned as a training set and the rest of them are combined into a new training set. This process repeats until all of the small chunks have been assigned as a testing set. In the end, the results are combined based on their evaluation scores and used to update the model. [56]

- **Testing** phase applies a testing dataset as an input for the trained model. The testing accuracy is then recorded and presented. It is crucial to have the testing

> dataset separated from the training dataset, otherwise the testing results might be misleading. The difference between a validation set and a testing set is that the validation set is utilizing during the training, however the testing set serves merely for the evaluation of the tuned model. [54]

Ideally, a learnt model should be applicable to a large variety of testing data and have no problem with correct predictions. In other words, the more generalized a model is, the better. Overfitting and underfitting are two terms related to model generalization. Overfitting refers to a situation, where a model is fit very tightly to the training data and therefore has difficulties classifying any other inputs. Underfitting is the opposite case and refers to a model which is too loosely fit and therefore outputs inaccurate predictions. The ultimate goal of machine learning is to strike the perfect balance between overfitting and underfitting and achieve a generalized model which remains applicable for future predictions. [54, 57]

The data for machine learning usually comes in a form of a $N_S \times N_F$ table, where N_S is the number of samples and N_F is the number of features. Each row represents a feature vector, a $1 \times N_F$ array of the individual feature values describing a specific sample. The features indicate the state of the samples in relevance to the study. For example, features such as age, Body Mass Index (BMI), KL score, etc. are valuable for subjects involved in knee OA studies.

Artificial neural networks (ANNs or NNs) are a very popular machine learning methodology. Basically, they are constructed to simulate the human brain. The fundamental piece of any neural network is a neuron. Just like in the human brain, neurons can receive an input, process it and send out an output. The idea is exactly the same for the artificial neuron. A neural network can consist of multiple neurons organized into consecutive layers. With each added neuron, the complexity of the network increases. The simplest NN is called a perceptron and it consists of a single neuron. A neural network with a number of perceptrons organized into multiple layers is called a multilayer perceptron (MLP). MLP consists of an input layer, hidden layers and output layer. The feature values are passed from the input layer to the neurons within the hidden layers. The hidden layers process the values and send them to the final output layer, which produces the network output. [58]

In order for a Multilayer Perceptron to learn, a backpropagation learning is often utilized. Firstly, the backpropagation algorithm calculates the error between the predicted value and the desired label and, secondly, travels back through the network and initiates a change in the internal parameters (weights and biases for each neuron) based on the calculated error. The backpropagation algorithm requires at least one layer of neurons fully connected to another layer. [59]

Naïve Bayes (NB) is another machine learning method however different from the neural networks. It is based on the Bayes theorem and it determines the most probable output based on calculating the prior probabilities from the prior knowledge about the data. Naïve Bayes is called naive, because it assumes independence of the attributes, which is a very rare condition in real life. However, in practice, Naïve Bayes is a very simple algorithm and as such allows for a very fast processing. [59]

Support Vector Machines (SVM) are based on calculating a hyperplane or hyperplanes to separate the input data according to their class. The hyperplane is fit in between the data by maximizing the distances between the hyperplane and the closest points i.e. maximizing the margin. In contrast to Naive Bayes, SVM methodology is not described with probabilities and only classifies the samples by either being on one side or the other from the hyperplane. SVM achieve their accuracy by transforming the data into higher dimensions based on a chosen kernel function and as a result, achieve a distribution that is easier to separate. [54, 60]

Machine learning has been substantially utilized in knee osteoarthritis studies throughout the past couple of years. In September 2020, Kokkotis et. al. [52] published a review of the current state of machine learning in knee OA studies. They found that Support Vector Machines seemed to be the most utilized algorithm due its good generalizability. Neural Networks were the second most popular choice. Furthermore, they found GLCMs to be one of the most popular approaches for feature extraction in studies utilizing either MR or X-ray images.

Deokar et. al. [61] applied a Multilayer Perceptron with Back Propagation learning method to differentiate between subjects with and without osteoarthritis. In their methodology, GLCM features *Contrast, Correlation, Energy, Homogeneity and Entropy* were used for the training. Their results show a 92% testing accuracy, however the paper does not provide any further information about the subject data nor about the GLCM input parameters.

Du et. al [60] applied a Multilayer Perceptron (one hidden layer), SVM, Naïve Bayes and a Random Forest for knee osteoarthritis prediction based on a novel texture feature extraction methodology. Although their study did not utilize GLCM features, they provide a good insight into the machine learning used for osteoarthritis studies. The best overall results were reported for the MLP, however very competitive results were reported in terms of Naive Bayes as well.

3. METHODS

3.1. Subject Dataset

The study sample was extracted from the Osteoarthritis Initiative (OAI), a longitudinal, observational study consisting of $4\,796$ (information valid for July 2020) participants being both men and women aged 45-79 at recruitment with, or at risk of, primary knee OA. All participants underwent MRI screenings at baseline and then after 12 months, 24 months, 36 months, 48 months, 72 months and 96 months. Such longitudinal research protocol provides highly valuable information about the progression of the knee OA disease over time. The collected images were analyzed by medical professionals and each subject was assigned a KL score according to the state of their cartilage at each screening timepoint. Based on the acquired KL grade, the entire dataset is divided into three separate cohorts. The *control cohort* refers to a group of asymptomatic subjects (KL = 0 at all screenings); *incidence cohort* is categorized by an increased risk of OA and longitudinal data show slow development of the disease; and *progression cohort*, consisting of subjects with both osteophytes and frequent symptoms. [14, 15]

For the purposes of this study, a dataset of 20 subjects was derived from the OAI database. For each selected subject, image data from the baseline (00m) and from the 36-month follow-up (36m) screening were included. 10 subjects were extracted from the control cohort (CTRL) and the other 10 subjects came from the incidence and progression cohort (PRGS). The selected PRGS subjects showed KL = 0 at baseline but progressed rapidly and scored KL $\geq$ 2 at the 36 month visit. Both subject cohorts showed no signs of the disease at the baseline screening. The study sample was constructed in a pair-wise manner. Every CTRL subject had a matching PRGS subject, with whom they shared same sex, same age ($\pm$ 2 years) and the difference in Body-Mass Index (BMI) over the 36-month long period was ± 2 kg/m^2. This subject matching was done to rule out the age, sex and BMI as confounding factors. The longitudinal relative BMI variation for each subject was within 10%. The samples and their attributes can be found in Table 2.

This dataset provided altogether 80 cartilages for analysis, 40$\times$femur and 40$\times$tibia. All cartilages were be divided into 8 subgroups by 10, according to the subject cohort, cartilage type and the screening time. The list of all the cartilages can be found in Table 1.

The image data itself included isotropic DICOM images collected using 3T clinical MR system with a 3D DESS MR sequence. The slice thickness is $0,7$ mm and all the data was encoded into 16-bit grey scale with a resolution of 384$\times$384 pixels per slice. There are 160 slices per subject. The cartilages were automatically segmented by a novel deep-learning software developed at the Research Unit of Medical Imaging, Physics and Technology at the University of Oulu by Panfilov et. al. [62]. An example of the cartilage segmentation is illustrated in Figure 8.

Table 1. The subject dataset described by the various cartilages

Number of Cartilages	subject cohort	Cartilage type	Timepoint
10	CTRL	Femur	00m
10	CTRL	Femur	36m
10	CTRL	Tibia	00m
10	CTRL	Tibia	36m
10	PRGS	Femur	00m
10	PRGS	Femur	36m
10	PRGS	Tibia	00m
10	PRGS	Tibia	36m

Table 2. subject dataset

Pair	ID	Group*	Sex	Age [years]	BMI 00m** [kg/m2]	KL 00m	BMI 36m [kg/m2]	KL 36m	Difference BMI 00m-36m
1	9892736	**CTRL**	male	48	23,0	**0**	23,0	**0**	0,0 %
	9509294	**PRGS**	male	48	24,3	**0**	23,9	**2**	1,6 %
2	9893729	**CTRL**	male	49	24,7	**0**	24,8	**0**	0,4 %
	9723972	**PRGS**	male	49	25,5	**0**	24,5	**2**	3,9 %
3	9093584	**CTRL**	male	61	25,7	**0**	25,9	**0**	0,8 %
	9656912	**PRGS**	male	61	25,7	**0**	25,9	**2**	0,8 %
4	9256066	**CTRL**	male	65	27,0	**0**	26,1	**0**	3,3 %
	9086407	**PRGS**	male	64	26,1	**0**	28,1	**3**	7,7 %
5	9931342	**CTRL**	female	46	22,3	**0**	23,0	**0**	3,1 %
	9623707	**PRGS**	female	45	24,1	**0**	21,8	**2**	9,5 %
6	9900690	**CTRL**	female	48	25,1	**0**	26,8	**0**	6,8 %
	9624154	**PRGS**	female	46	26,9	**0**	26,5	**2**	1,5 %
7	9915764	**CTRL**	female	51	26,2	**0**	26,6	**0**	1,5 %
	9271853	**PRGS**	female	51	25,6	**0**	25,6	**3**	0,0 %
8	9276291	**CTRL**	female	57	22,6	**0**	24,1	**0**	6,6 %
	9828518	**PRGS**	female	57	24,6	**0**	22,6	**2**	8,1 %
9	9254514	**CTRL**	female	56	23,9	**0**	23,4	**0**	2,1 %
	9545340	**PRGS**	female	57	24,8	**0**	25,5	**2**	2,8 %
10	9907767	**CTRL**	female	59	22,1	**0**	23,0	**0**	4,1 %
	9412037	**PRGS**	female	59	22,5	**0**	22,6	**2**	0,4 %

*CTRL - subjects from the control cohort; PRGS - subjects from the incidence or progressive cohort.
**00m is the baseline visit (initial screening); 36m is the 36 month visit (3 year follow up screening)

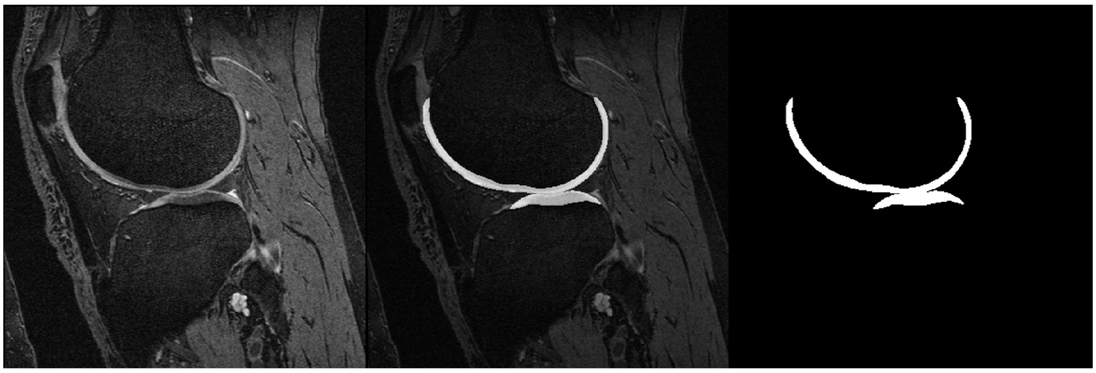


Figure 8. An example of the image data with the pre-calculated mask segmentation using the automatic cartilage segmentation tool.

3.2. Cartilage Histogram Analysis

The pixel distributions of the cartilages can provide a quantitative insight into the physiological changes within both symptomatic and asymptomatic cartilages. In order to establish initial understanding of the studied cartilages, a histogram analysis was performed. Each subject's cartilage data, both femur and tibia, was studied both individually and longitudinally. Particular focus was put on finding some clear differences between the cartilage at baseline and 36-month-follow-up screenings for the subject cohorts.

Upon extracting the cartilage pixels based on the segmented masks (Figure 9), each 2D cartilage matrix was flattened into a 1-dimensional vector v_c and used for the histrogram calculations. Each subject provided a histogram of: 1) Femur from 00m; Femur from 36m; Tibia from 00m; Tibia from 36m. The focus was kept on both longitudinal changes but also the differences between the subjects at a single timepoint. All histograms were visually inspected.

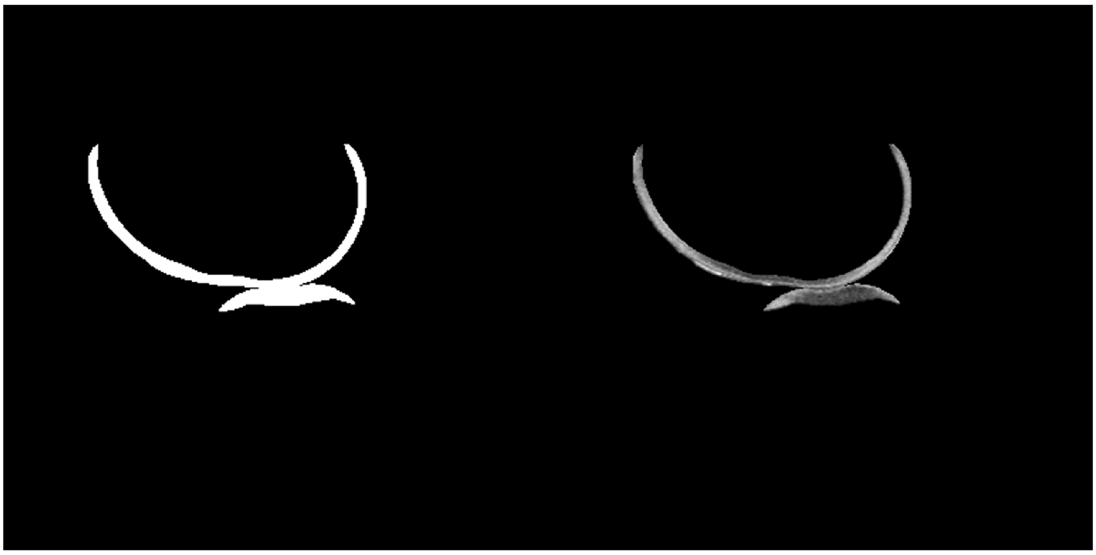


Figure 9. Acquiring the cartilage pixels.

To extract an overall picture about all the subject cartilages, the occurrences of every detected pixel intensity were counted across all 80 cartilages and a cumulative histogram H_{total} was calculated. Such cumulative histogram provided information about all the pixel intensities across the entire subject dataset and provided indications about the possible choices of the input parameters for the subsequent 3D Texture Analysis.

The minimum pixel intensity and maximum pixel intensity were collected from each femur and tibia. In order to remove extreme points from the cartilage histograms before applying the 3D Texture Analysis, a threshold algorithm dubbed *Pixel threshold* was utilized. The algorithm follows these steps: 1) The flattened cartilage vector v_c is sorted in ascending order; 2) The non-cartilaginous (black) pixels are discarded;

3) A given percentage of the pixel amount was cut off from each side of the sorted sorted v_c vector. For example, if a femur cartilage consisted of 10×10^4 pixels and the cutoff percentage was selected to be 2%, then $0,2 \times 10^4$ pixels would be removed from each side of the flattened cartilage resulting in a output cartilage vector v_{cut} of $9,6 \times 10^4$ pixels. The first and last element of the shortened sorted cartilage vector v_{cut} marked the updated minimum and maximum pixel intensities respectively. Updated minimum and maximum pixel intensities were collected for 2%, 5% and 10% cutoff percentages. The collected minimum and maximum pixel intensities were further averaged across 1) all the cartilages; 2) the subject cohorts; 3) the timepoints; and 4) subject cohorts at different timepoints. These average pixel intensities were further utilized for the application of the 3D Texture analysis.

3.3. 3D Texture Analysis Tool

The 3D Texture Analysis Tool (3D TA) was developed in 2018 by Ari Väärälä[9] as a novel method for the extraction of GLCM-based features from the knee cartilage texture. The unique aspect of this type of analysis is, as the name suggests, that the data is analyzed in three dimensions. This novel approach has a potential for improved recognition of cartilage degradation and early prediction of OA.[9]

The tool provides a unique way to quantitatively assess the cartilage in various thickness layers. The cartilage layers are 1-pixel thick and calculations for four layers are currently implemented: 1) **L10**, accounting for layer found at 10% thickness height; 2) **L50**, found in the middle of the cartilage; 3) **L90**, providing information about a cartilage layer found at the 90% cartilage height; and 4) **SUM**, which represents the full cartilage thickness. Figure 10 provides a visual representation of the cartilage layers.

The 3D Texture Analysis was developed with Matlab®(MathWorks Inc., MA, USA) at the Medical Imaging, Physics and Technology Research Unit at the University of Oulu. In order to extract the texture features, the software first calculates the pixels at the bone-cartilage interface (BCI), which is the edge where the cartilage and subchondral bone meet. Next, the segmented 3D cartilage is anatomically normalized into a number of tiny overlapping 3D rectangles, which carry the information about the neighbouring pixels for each BCI pixel. The layers of the 3D rectangle correspond to the cartilage 1-pixel thick layers. The grey level co-occurrence matrices are calculated from the neighbouring pixels in the 3D rectangles. Each studied layer (L10, L50 and L90) has its own calculated GLCMs. The SUM layer provides information about the full thicknes by summarizing the GLCMs altogether. 19 GLCM features defined by Haralick et. al. [7], Clausi et. al. [63] and Soh et. al. [64] are extracted by the 3D Texture Analysis for each layer. Table 3 lists all the extracted features and their corresponding reference. [9, 11]

The tool has 12 input parameters altogether, out of which the first 4 are of crucial importance for this thesis:

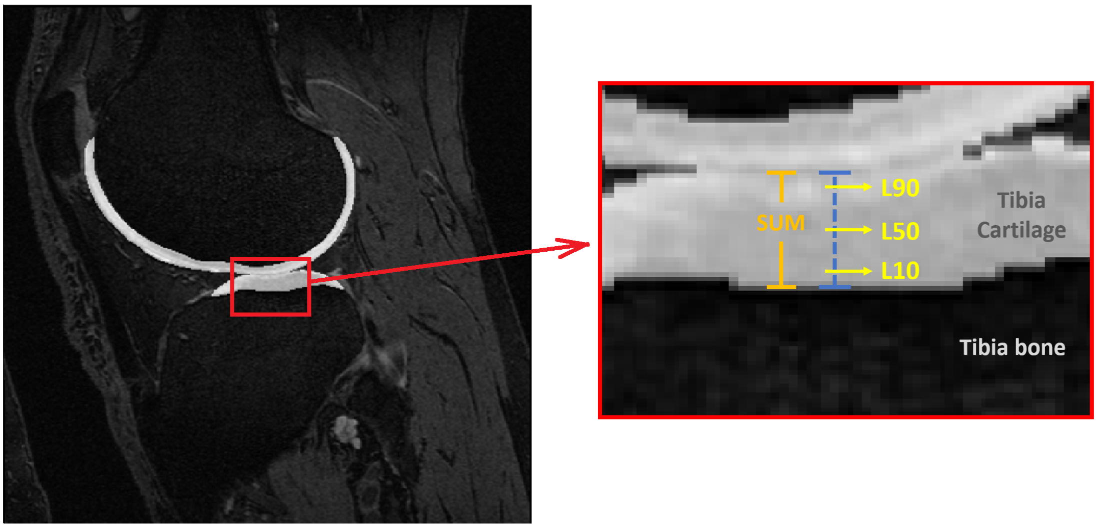


Figure 10. Cartilage layers. Blue dashed line represents the thickness of the cartilage. L10, L50, L90 show the layer heights, in which the cartilage was analyzed. SUM represents the summation of all layers (full cartilage thickness)

Table 3. List of the output GLCM texture features with their corresponding reference

	Feature name	Reference
1	Autocorrelation	Soh et. al. [64]
2	Cluster prominence	Haralick et. al. [7]
3	Cluster shade	Haralick et. al. [7]
4	Contrast	Haralick et. al. [7]
5	Correlation	Haralick et. al. [7]
6	Difference entropy	Haralick et. al. [7]
7	Difference variance	Haralick et. al. [7]
8	Dissimilarity	Soh et. al. [64]
9	Energy	Haralick et. al. [7]
10	Entropy	Haralick et. al. [7]
11	Homogeneity	Soh et. al. [64]
12	Information measure of correlation 1	Haralick et. al. [7]
13	Information measure of correlation 2	Haralick et. al. [7]
14	Inverse difference	Clausi et. al. [63]
15	Maximum probability	Soh et. al. [64]
16	Sum average	Haralick et. al. [7]
17	Sum entropy	Haralick et. al. [7]
18	Sum of squares (variance)	Haralick et. al. [7]
19	Sum variance	Haralick et. al. [7]

1. **Minimum grey level** - determines the minimum pixel intensity and sets the lower boundary for the bin quantization. Figure 11 shows an example of choosing the pixel intensity of 50 as the minimum grey level on a randomly selected tibia histogram.
2. **Maximum grey level** - determines the maximum pixel intensity and sets the upper boundary for the bin quantization. Figure 11 shows an example of choosing the pixel intensity of 300 as the maximum grey level on a randomly selected tibia histogram.
3. **Bin quantization number** - determines the quantization range. Figure 11 illustrates an example of an 8-bin quantization. The algorithm keeps the grey values located outside of the selected grey level boundaries and assigns them to the first and last bin. In this example, the quantized image would consist merely of pixel values 1 to 8 and all the calculated GLCMs would be 8×8 in size. The number of bins is also directly proportional to the computational time. Since the bin quantization merges several pixel intensities into one, the quantization also works as a simple noise filter. As a result, the higher the bin number, the more noise will pass through into the analysis. At this point, only linear quantization (= equal bin size for all the bins within the grey level boundaries) is supported. [9, 13]
4. **Offset** - refers to the spacing between the root pixel and the selected neighbour while constructing the GLCM. An example of the change in offset number and its impact on the GLCM calculation can be found in section 2.5.

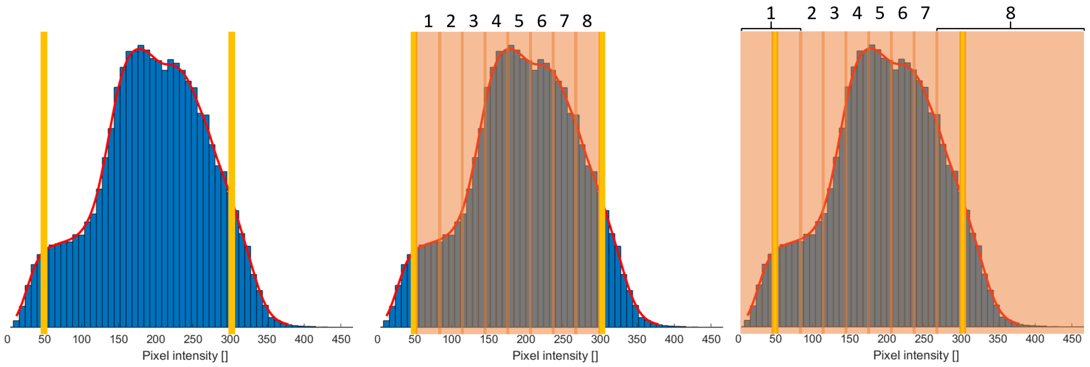


Figure 11. Example of an 8-bin quantization based on 50 minimum grey level and 300 maximum grey level. Pixel intensities found below the minimum grey level are added to the first bin. Pixel intensities found above the maximum grey level are added to the last bin.

3.4. Applied 3D Texture Analysis

The input parameters for the GLCM calculation are rarely studied and multiple researchers have raised awareness about the lack of the parameter standardization [49][12]. Therefore, to achieve a clearer image about their effect, various combinations of *minimum grey level, maximum grey level, bins* and *offset* were studied and analyzed.

The rest of the 3D TA input parameters were kept fixed.

In this thesis, **bin quantization schemes 4, 8, 12, 16, 32, 64** and **offset settings 1, 2, 3, 4** were studied. The input parameters were divided into five subcategories based on the analyzed grey level quantization schemes:

- **dynamic - dynamic** approach assigns a specific grey range based on the pixel intensity distribution of the analyzed cartilage. Minimum grey level is assigned based on the minimum intensity of the current cartilage. Similarly, the maximum grey level is assigned based on the maximum pixel intensity of the cartilage. This approach corresponds to the grey level quantization methodology utilized by Gomez et. al. [13].
- **dynamic - static** uses the dynamic grey level assignment only for the minimum grey level value. The maximum grey levels are statically assigned.
- **static - dynamic** uses the dynamic grey level assignment only for the maximum grey level value. The minimum grey level is constant.
- **static - static** approach has both grey level boundaries fixed for all the subjects. Encapsulating all the possible intensities present within the dataset is a grey level quantization method recommended by Brynolfsson et. al. [11].
- **special** approach is based on averaging the minimum and maximum pixel intensities of the cartilages across to the cartilage type (femur or tibia), time of acquisition (00m or 36m) and subject cohort (CTRL or PRGS). The minimum and maximum grey levels were selected based on the average values and assigned to the corresponding cartilage subgroups. The special approach can be viewed as a subtype of the static - static approach, however with mathematically defined grey level boundaries. There were 3 different setups:

  1. Average grey levels for **femur** and **tibia**. This creates **2** different sets of input values for each subject, one for femur and one for tibia.
  2. Average grey levels for **femur** at **00m**, **femur** at **36m**, **tibia** at **00m** and **tibia** at **36m**. This approach creates **4** different sets of input parameters for each analyzed subject.
  3. Lastly, separate static grey levels for every possible cartilage categorization, which means a specific input parameter set for a **femur** from a **CTRL** subject at **00m**, **femur** from **CTRL** at **36m** and so on. This approach creates **8** different sets of input parameters for each subject.

  The special approach was included to study, if a customized grey level quantization schemes adjusted for separate cartilages, timepoints and subject cohorts could yield superior results.

The nomenclature of the analysis outputs is based on the order of the input parameters. For special cases, the naming varies according to the type of the special approach. Table 4 summarizes the naming of the various outputs. The cutoff percentages within the examples in Table 4 are $0,05$, i.e. 5%.

The original 3D Texture Analysis script allowed only a single set of grey levels and bins to be calculated at a time. However, the algorithm was optimized to accommodate a calculation of multiple grey level ranges and bin quantizations at the same time,

Table 4. Nomenclature based on the various grey level quantization schemes

Grey level quantization scheme	**Nomenclature**	**Example**
dynamic - dynamic	-1_-1_binNumber	-1_-1_8
dynamic - static	-1_maximumGreyLevel_binNumber	-1_300_16
static - dynamic	minimumGreyLevel_-1_binNumber	0_-1_4
static - static	minimumGreyLevel_maximumGreyLevel_binNumber	0_400_32
special 1	cutoff_f_t_binNumber*	0-05_f_t_12
special 2	cutoff_f00_t00_f36_t36_binNumber	0-05_f00_t00_f36_t36_64
special 3	cutoff_all_different_binNumber	0-05_all_different_4

*cutoff refers to the cutoff percentage used to shorten the original cartilage vector before extracting the grey levels

since *grey level limits* and *bins* are **not** bound to the interpolation and extrapolation tasks. This update should provide a substantial reduction in the computation time and as a result, more combinations of input parameters can be collected and subsequently evaluated.

Altogether, 528 3D Texture Analysis outputs were collected. The outputs were collected separately for each offset, i.e. 132 results for offsets 1-4. Due to this type of collection, majority of the methods were separated by offset as well to make the analysis more feasible.

3.5. Statistical Analysis

In order to collect the statistical differences between CTRL and PRGS subjects, the effect size was calculated between the corresponding output texture features. For example, CTRL Autocorrelation values calculated from femur at 00m were statistically compared to the PRGS Autocorrelation values calculated from femur at 00m and calculated effect size was collected.

Normality of the output texture features was tested with Lilliefors test, Kolmogorov-Smirnov test and Jarque-Bera test. Lilliefors and Jarque-Bera indicated, that the features come from normal distribution, however Kolmogorv-Smirnov test rejected that hypothesis. Additionally, the small sample size in this study (10 subjects per cohort) should not be overlooked and, although two out of three tests indicated normality, a non-parametric analysis was considered to be the most appropriate solution. [65, 66]

Therefore, Cliff's delta (δ) was utilized to measure the differences between CTRL and PRGS cohorts based on the output texture features. Cliff's δ was introduced in 1993 as a non-parametric measure of effect size and has been utilized ever since in order to remove the condition of normality from the statistical analysis. Cliff's method has not only been shown to be more robust and therefore might be more suitable, but also seems to strongly correlate with Cohen's d, a well-established powerful parametric measure of effect size. [67, 68, 65]

Table 5. Table of effectiveness level based on Cliff's delta

Cliff's δ	Level of effect
$0,474 \leq \lvert\delta\rvert$	Large
$0,33 \leq \lvert\delta\rvert < 0,474$	Medium
$0,147 \leq \lvert\delta\rvert < 0,33$	Small
$\lvert\delta\rvert < 0,147$	Negligible

To calculate the δ values, a Matlab toolbox 'Measures of Effect Size' by Hentschke and Stüttgen [68] was utilized. The algorithm for Cliff's δ per se is not implemented in the toolbox, however the δ is linearly proportional to the Area Under the Receiver Operating Characteristic Curve (AUROC), which is a widely used non-parametric measure of effect size and happens to be part of the toolbox. Based on the linear relationship, the δ values can be extracted from the AUROC values fairly easily with the following equation[68]:

$$\delta = 2 \times AUROC - 1 \tag{2}$$

The range of Cliff's δ is from -1 to 1, where 0 indicates no detectable effect between the studied groups. In 1988, Cohen [69] established a method, how to evaluate the effect of his d. Those principles can be re-interpreted for the non-parametric δ into a set of evaluation guidelines for the absolute δ values, which can be found in Table 5. [68, 70]

To statistically evaluate which combination of 3D Texture Analysis input parameters might be the most beneficial for differentiating between CTRL and PRGS subjects, the total number of small, medium and large effects found per the studied combinations of grey level ranges, bins and offsets were counted and tabulated.

The goal of the statistical analysis is: 1) evaluate the ability of individual 3D Texture Analysis outputs to measure differences between the subject cohorts; and 2) to identify features which are responsible for localizing above-average number of effects between the studied subject cohorts and thus create a selected feature subset for the subsequent machine learning analysis.

3.6. Machine Learning Analysis

To assess the predictive abilities of the calculated features, a machine learning (ML) pipeline for the knee OA prediction was established. The ML pipeline has 5 parameters and the entire flowchart is depicted in Figure 12. The collected 3D Texture Analysis output features served as an input for the machine learning algorithms. Only data from **baseline** was utilized for the machine learning analysis. The goal was to see the predictive capabilities based upon measurements at the beginning of the longitudinal study before any OA symptoms were registered in the progressive cohort.

The machine learning analysis consists of these steps:

1. **Select the ML pipeline parameters.**
   - **Cartilage selection.** Firstly, it is important to select which cartilage will be utilized to train the classifiers.
   - **Feature selection.** The learning phase can utilize either all 19 available features or only a selected feature subset determined upon the results from the statistical analysis.
   - **Classifier selection.** *Multilayer Perceptron (MLP)*, *Naïve Bayes (NB)* and *Support Vector Machines (SVM)* were utilized in order to collect the ML results.
     - *Multilayer Perceptron* is a type of feed-forward artificial neural network. A predefined implementation of MLP with backpropagation learning developed in 2016 for Matlab by Hesham Eraqi [71] was utilized. The network consists of two hidden layers, each containing 11 neurons, and 2 output neurons (one for each class).
     - *Naïve Bayes (NB) classifier* was implemented with the native Matlab function `fitcnb()`.
     - *Support Vector Machine (SVM) classifier* was implemented with the native Matlab function `fitsvm()` with no kernel function assigned to it.
   - **Data split type.** Two possible ways of creating the training and testing sets were utilized: **pair-wise** and **random**. In the *pair-wise* scenario, a single CTRL-PRGS pair was used as the testing set and the remaining 18 samples (9 CTRL samples and 9 PRGS samples) were used for training. In the *random* scenario, either 1, 2 or 3 randomly selected subjects were used as testing samples. The rest of the samples were used for training.
   - **Bootstrapping.** Bootstrapping was introduced to further challenge the robustness of the input data. If m is the number of testing data, then the bootstrapping takes m random training samples and adds their copies into the training dataset. With the original dataset consisting of 20 samples, the bootstrapping makes sure that the training set is scaled up to 20 samples as well, no matter the data split type.
   - **Cross-validation.** Before testing the trained model, Naïve Bayes and Support Vector Machines were either cross-validated with *10-fold cross-validation* or left untouched. The MLP remained without cross-validation, because the idea was to test the implemented MLP algorithm as is without any adjustments.
2. **Apply the established pipeline**. In order to evaluate any stochastic machine learning, Brownlee [72] recommends to repeat the algorithm anywhere between 30 up to 1000 times. The number of repetitions purely depends on the hardware and the time necessary to finish the task. Each repetition, the output labels are recorded and stored. For the purposes of this thesis, 200 was chosen as the repetition number due to the time limitations. In the end, each Texture Analysis output is described by its 1×200 vector containing the individual classification accuracies from each repetition.
3. **Collect the tables Y_1, Y_2, Y_3 and Y_4.** Each Y_i table is $N_{R/O} \times 200$, where $N_{R/O}$ refers to the number of 3D Texture Analysis outputs per offset ($N_{R/O} = 132$).

For example, table Y_1 contains accuracies only for the Texture Analysis outputs obtained with offset 1, Y_2 with offset 2, etc. Such approach should provide easier evaluation due to the smaller table size, plus it allows us to immediately compare the robustness values merely for the grey level and bin quantization schemes.

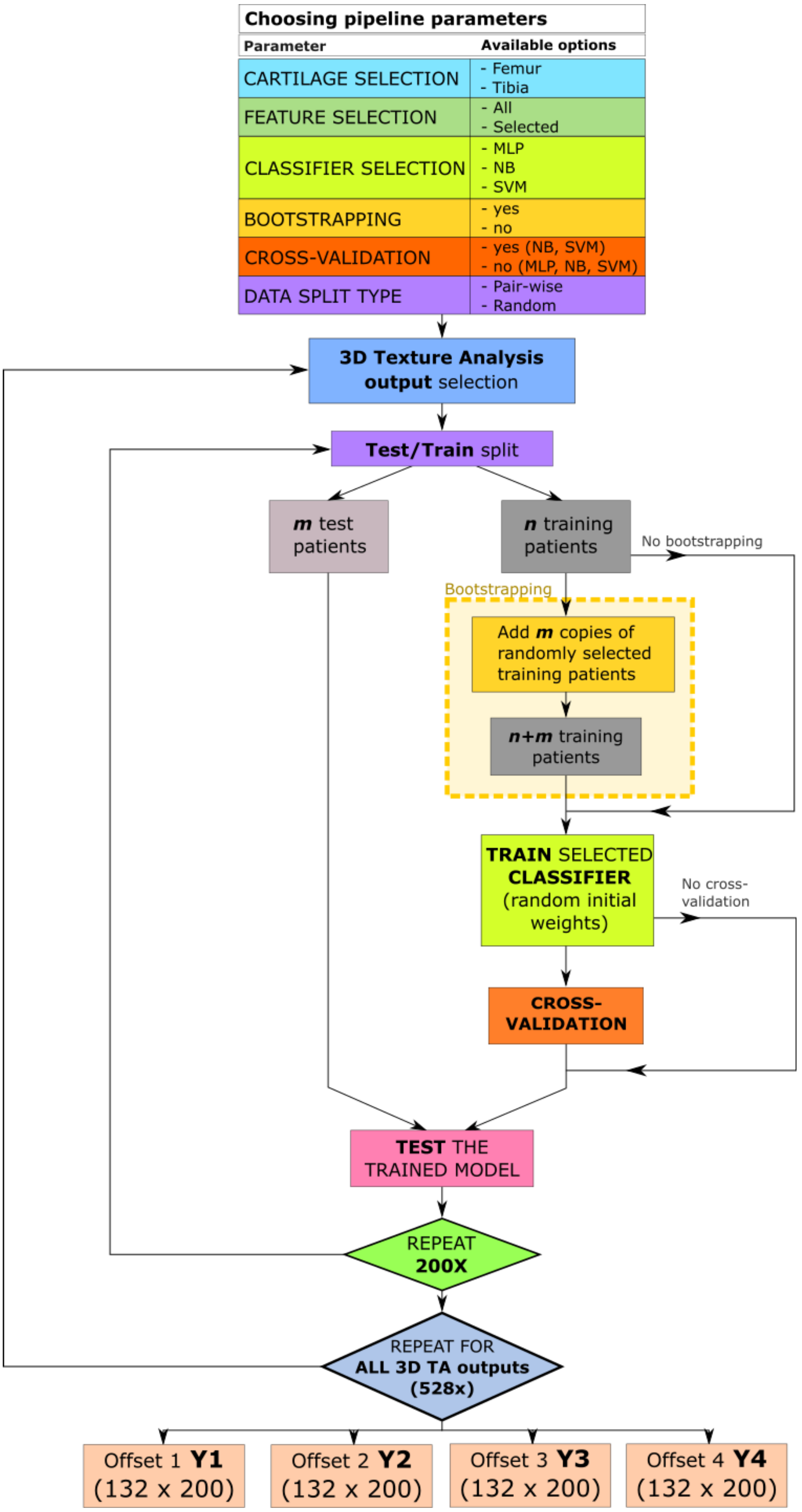
4. **Calculate robustness**. The robustness refers to the average accuracy for each row of a Y_i matrix, i.e. the average accuracy of each 3D TA output across the 200 repetitions. Those Texture Analysis outputs which yield consistently favorable results (i.e. high robustness) will be considered superior. The measure of robustness simply indicates how many times the algorithm was able to classify the samples correctly into their corresponding cohorts (for example, 60% robustness means 60 times out of 100 were the subjects classified correctly as CTRL or PRGS by a model trained on features from the corresponding 3D Texture Analysis output).
5. Extract tables B_1, B_2, B_3 and B_4. Each B_i table contains only the top 10 most robust performers from the corresponding Y_i table, where $i = \{1, 2, 3, 4\}$. Each B_i table can be understood as a shrunk down version of the Y_i table with only the best performers.
6. Repeat the entire pipeline j number of times with various combinations of the pipeline parameters. Therefore, j number of the pipeline runthroughs will yield j number of $B_{1,2,3,4}$ tables, each containing the best performers for the corresponding pipeline parameter combination. Only a subset of all the possible pipeline parameter combinations were applied due to the time constraints.
7. Find the best performing 3D Texture Analysis outputs. It was considered how many times a certain 3D Texture Analysis output appeared within the $B_{i,j}$ matrices. In other words, how many times has a certain 3D TA output appeared amongst the top 10 performers across the results from all the applied pipeline parameter combinations. The more times a 3D TA output performed the top 10, the higher it scored. Finally, a sorted B_{all} table is created, containing $N_{R/O}$ of 3D Texture Analysis outputs with their corresponding number of times each one appeared amongst the top 10 performers. The 20 best overall performers as well as 20 overall worst performers were tabulated.

In order to measure the average impact of the 3D Texture Analysis input parameters, the differences between the 3D TA outputs with the best robustness and worst robustness from each collected $Y_{i,j}$ matrix were calculated and subsequently the maximum measured difference and the average difference were extracted.

The impact of feature selection was studied by comparing the average robustness scores across all the $B_{i,j}$ tables between all-feature-based and selected-feature-based results.

The bin quantization number of each 3D Texture Analysis output was extracted from the B_{all} table. Since the B_{all} table contains the sorted best performers, extracting their bin quantization numbers into a $1 \times N_{R/O}$ vector provided a bin distribution across the best performers.

The impact of varying offset was evaluated by averaging the robustness scores for each 3D Texture Analysis output across the studied offsets and collecting the percentual increase or decrease in robustness for offset 2,3 and 4 compared to the average robustness scores for offset 1.



Choosing pipeline parameters	
Parameter	Available options
CARTILAGE SELECTION	- Femur - Tibia
FEATURE SELECTION	- All - Selected
CLASSIFIER SELECTION	- MLP - NB - SVM
BOOTSTRAPPING	- yes - no
CROSS-VALIDATION	- yes (NB, SVM) - no (MLP, NB, SVM)
DATA SPLIT TYPE	- Pair-wise - Random

Figure 12. Flowchart of the machine learning pipeline. Y1, Y2, Y3 and Y4 are the output matrices of accuracies from individual offsets.

4. RESULTS

All results were collected on a Windows 10 PC with 8 GB of RAM using Intel®Core™i5-8400T CPU with 6 cores, 6 threads and a base clock speed 1,70 GHz.

4.1. Cartilage Histogram Analysis

The number of pixels found in femur was approximately three times larger than in tibia. On average, femur cartilage contained around $(13,7 \pm 3,0) \times 10^4$ pixels (mean ± SD). Tibia cartilage, on the other hand, consisted of about $(4,9 \pm 1,4) \times 10^4$ pixels.

Femur showed a slightly negatively skewed unimodal distribution (see Figure 13). On the other hand, tibia showed a predominantly bimodal distribution. The y-axis values varied due to the different pixel amounts. Both cartilages showed similar intensity levels across their x-axes at baseline, however longitudinal comparison showed an increase in pixel intensity range for **all** subjects over time. After 36 months, maximum pixel intensities exhibited an average increase of 31,5% for femurs an 30,5% for tibiae. Histograms from a single randomly selected subject are plotted in Figure 13 to demonstrate the longitudinal change in pixel intensity range.

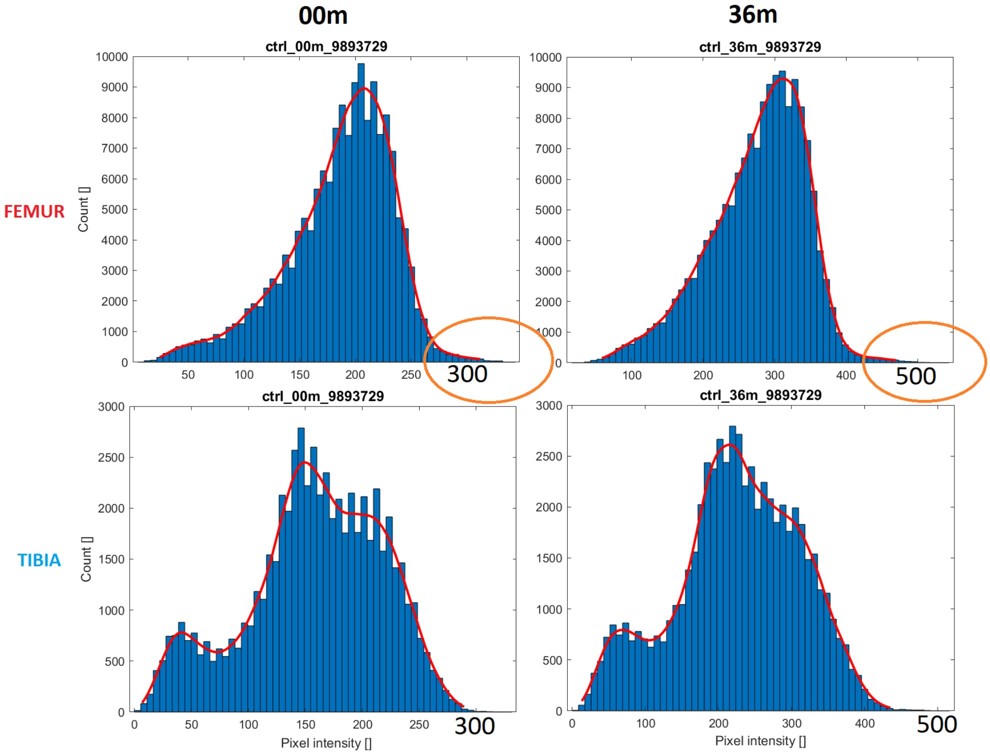


Figure 13. Histograms of a femur (upper row) and tibia (lower row) at 00m (left column) and 36m (right column) from a randomly selected subject. The x-axes denote the pixel intensity distributions. The y-axes mark the number of occurrences. The orange circles highlight the longitudinal increase in pixel intensity.

The number of occurrences of every pixel intensity present within the 80 studied cartilages was counted and a cumulative histogram H_{total} was calculated and plotted in Figure (14). The 90% cumulation was measured right after the 300 pixel intensity, the 99% cumulation at 399 pixel intensity and finally 100% cumulation was recorded at 788 pixel intensity (denoted by a red line in Figure 14). The 788 pixel intensity was found only once in a single progressive subject in their 36m screening data. In contrast, the most represented pixel intensity across all the cartilages was 203 ($4{,}7 \times 10^4$ occurences).

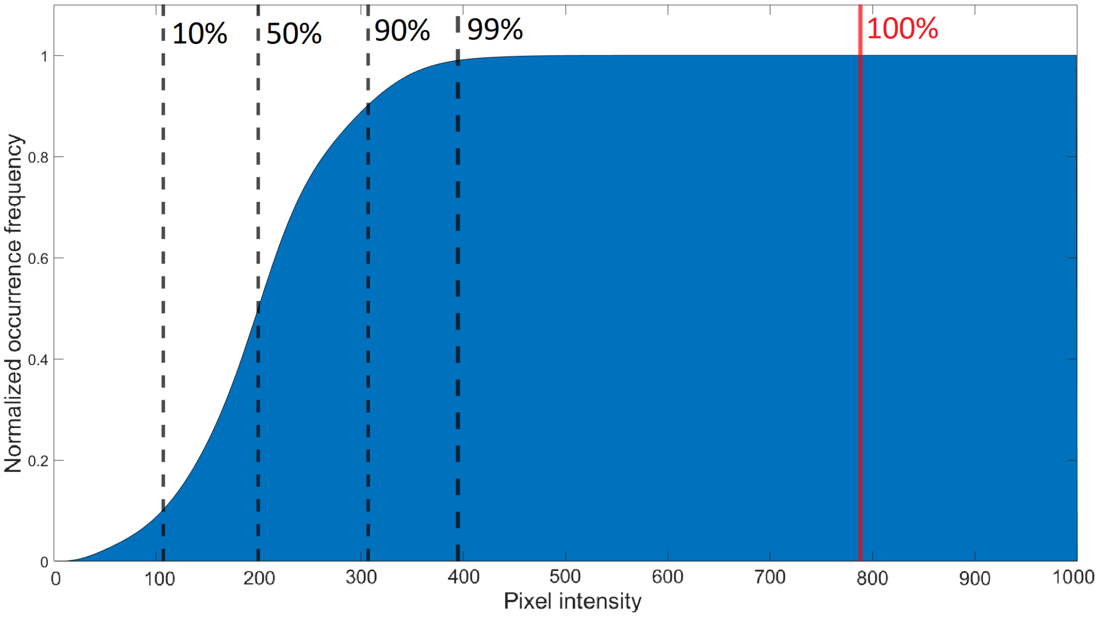


Figure 14. Cumulative histogram of all the pixel intensities collected from all studied cartilages. X-axis marks the pixel intensity distribution. Y-axis marks the normalized number of occurrences. Vertical dashed lines denote the percentual cumulation relative to the pixel intensities. Pixel intensity 107 marks 10% cumulation; pixel intensity 200 marks 50% cumulation; pixel intensity 307 marks the 90% cumulation; pixel intensity 399 marks the 99% cumulation. Red line denotes the full 100% cumulation at 788 pixel intensity.

The maximum and minimum pixel intensities were derived from each cartilage. The minimum and maximum pixel intensities were averaged across all subjects (blue row in Table 6), subject cohorts (red rows in Table 6), timepoints (yellow rows in Table 6) and subject cohorts at separate timepoints (green rows in in Table 6). The table shows minimum and/or maximum average pixel intensities after a particular cutoff. The difference between 00m and 36m maximum grey level further illustrates the increase in pixel intensity over time. The average difference between the actual maximum pixel intensities (0%) and the maximum pixel intensities from the 2% shortened cartilage vectors was $120,7$ intensity values for femur and $97,1$ intensity values for tibia.

These results provide a solid ground for establishing the static grey level limits for the texture analysis algorithm. 2% cutoff percentage was selected for the dynamic grey level quantizations. Based on the total cumulative histogram in Figure 14), static minimum grey level 0 and maximum grey levels $300, 400, 500, 600, 700, 800$ were chosen for the static grey level quantization. The pixel intensity values from Table 6

Table 6. Average minimum and maximum pixel intensities across the available cartilage classes derived from either original or shortened cartilage vectors

Average across	Femur								Tibia							
	Minimum pixel intensity				Maximum pixel intensity				Minimum pixel intensity				Maximum pixel intensity			
	0%	**2 %**	**5 %**	**10 %**	**0%***	**2 %**	**5 %**	**10 %**	**0%**	**2 %**	**5 %**	**10 %**	**0%**	**2 %**	**5 %**	**10 %**
all	8,3	72,1	93,0	107,1	425,6	305,2	293,4	286,6	2,0	31,9	42,0	50,8	395,0	297,9	286,2	278,0
ctrl	7,5	69,9	90,7	104,9	421,9	303,8	292,8	286,4	1,8	31,0	41,2	50,0	399,5	297,5	285,8	277,7
prgs	9,1	74,3	95,4	109,4	429,4	306,7	294,0	286,9	2,2	32,7	42,9	51,6	390,5	298,3	286,6	278,2
ctrl_00m	4,9	57,6	76,1	88,9	349,8	257,9	248,8	243,4	1,4	24,7	33,5	41,4	326,5	252,8	243,3	237,0
ctrl_36m	10,1	82,2	105,2	120,9	493,9	349,7	336,8	329,4	2,2	37,3	48,8	58,6	472,4	342,1	328,3	318,4
prgs_00m	5,5	60,4	79,2	91,3	346,3	254,4	245,4	240,1	1,1	25,8	34,6	41,9	322,2	248,8	239,7	233,3
prgs_36m	12,7	88,2	111,6	127,4	512,4	358,9	342,5	333,6	3,3	39,6	51,1	61,2	458,8	347,7	333,5	323,1
00m	5,2	59,0	77,7	90,1	348,1	256,2	247,1	241,8	1,3	25,3	34,1	41,7	324,4	250,8	241,5	235,2
36m	11,4	85,2	108,4	124,2	503,2	354,3	339,7	331,5	2,8	38,5	50,0	59,9	465,6	344,9	330,9	320,8

* 2%, 5% and 10% denote the percentage pixel cutoff used to shorten the cartilage vector before the extraction of minimum and maximum grey levels. For example, minimum pixel intensity 10% column shows the average minimum intensities derived from the flattened cartilage vectors shortened by 10% from each side.
* 0% percentage means the minimum and maximum pixel intensities were derived from the original cartilage vector.

were used for the special approach of static grey level quantization.

4.2. Applied 3D Texture Analysis

Altogether, 528 3D Texture Analysis outputs were collected for. Six bin quantization schemes were calculated at a time during each 3D Texture Analysis run. The number of simultaneously calculated grey level quantization schemes varied between one to three. The **average acquisition time** for one grey level quantization scheme and six bin quantization schemes was approximately **49 hours and 50 minutes**; for two grey level quantization schemes **59 hours and 45 minutes**; for three grey level quantization schemes **68 hours and 42 minutes**. Each grey level calculation added approximately 10 hours to the overall computation time. The 3D Texture Analysis outputs were collected over a period of four months and list of the collected outputs is shown in Table 7.

4.3. Statistical Analysis

Cliff's delta (δ) effect sizes between the control and progressive subjects were calculated based on the output texture features. Altogether, $58\,313$ small δ effect sizes (0,147$\leq$ δ <0,33), $18\,193$ medium δ effect sizes (0,33$\leq$ δ <0,474) and $11\,690$ large δ effect sizes (0,474$\leq$ δ) were observed across the collected outputs. The identified effect sizes between the subject cohorts were summarized across the cartilage layers, cartilage types and timepoints, grey level quantization schemes, bin quantization schemes and offsets in order to accentuate the individual importance of each variable.

The amount of the δ effect sizes across the cartilage layers are shown in Table 10. The highest amount of small and medium effect sizes was found in L10 and the highest amount of large effects in L50. In contrast, the least amount of the studied effects was observed in L90.

Table 7. List of analyzed combinations of input parameters

Cutoff percentage	Minimum grey level	Maximum grey level	Bin number	Offset
2 %[*]	-1[**]	-1	4, 8, 12, 16, 32, 64	1, 2, 3, 4
2 %	-1	300	4, 8, 12, 16, 32, 64	1, 2, 3, 4
2 %	-1	400	4, 8, 12, 16, 32, 64	1, 2, 3, 4
2 %	-1	500	4, 8, 12, 16, 32, 64	1, 2, 3, 4
2 %	-1	600	4, 8, 12, 16, 32, 64	1, 2, 3, 4
2 %	-1	700	4, 8, 12, 16, 32, 64	1, 2, 3, 4
2 %	0	-1	4, 8, 12, 16, 32, 64	1, 2, 3, 4
-	0	300	4, 8, 12, 16, 32, 64	1, 2, 3, 4
-	0	300	4, 8, 12, 16, 32, 64	1, 2, 3, 4
-	0	400	4, 8, 12, 16, 32, 64	1, 2, 3, 4
-	0	500	4, 8, 12, 16, 32, 64	1, 2, 3, 4
-	0	600	4, 8, 12, 16, 32, 64	1, 2, 3, 4
-	0	700	4, 8, 12, 16, 32, 64	1, 2, 3, 4
-	0	800	4, 8, 12, 16, 32, 64	1, 2, 3, 4
2 %	f_t[***]		4, 8, 12, 16, 32, 64	1, 2, 3, 4
5 %	f_t		4, 8, 12, 16, 32, 64	1, 2, 3, 4
10 %	f_t		4, 8, 12, 16, 32, 64	1, 2, 3, 4
2 %	f00_t00_f36_t36		4, 8, 12, 16, 32, 64	1, 2, 3, 4
5 %	f00_t00_f36_t36		4, 8, 12, 16, 32, 64	1, 2, 3, 4
10 %	f00_t00_f36_t36		4, 8, 12, 16, 32, 64	1, 2, 3, 4
2 %	all_different		4, 8, 12, 16, 32, 64	1, 2, 3, 4
5 %	all_different		4, 8, 12, 16, 32, 64	1, 2, 3, 4
10 %	all_different		4, 8, 12, 16, 32, 64	1, 2, 3, 4

* Cutoff percentage indicates the percentage used to shorten the cartilage vector and assign the grey levels.

** -1 indicates dynamic grey level assignment.

*** f_t, f00_t00_f36_t36 and all_different utilize average pixel intensities for the minimum and maximum grey level.

Table 8. The amount of δ effect sizes found per each analyzed cartilage layer

Layer	Amount of effect sizes		
	small δ effects ($0{,}147 \leq \delta < 0{,}33$)	**medium δ effects ($0{,}33 \leq \delta < 0{,}474$)**	**large δ effects ($0{,}474 \leq \delta$)**
L10	15702[*]	5514	2957
L50	14358	4408	4335
L90	14823	3630	1511
SUM	13430	4641	2887

* The table is vertically color-coded. The cell color saturation is directly proportional to the cell value.

Table 9. The amount of δ effect sizes found per each cartilage and timepoint

Cartilage	**Amount of effect sizes**		
	small δ effects	**medium δ effects**	**large δ effects**
	(0,147$\leq \delta <$0,33)	**(0,33$\leq \delta <$0,474)**	**(0,474$\leq \delta$)**
00m Femur	12162*	2065	502
00m Tibia	14617	10559	8890
36m Femur	16023	3033	1311
36m Tibia	15511	2536	987

* The table is vertically color-coded. The cell color saturation is directly proportional to the cell value.

In terms of the results from cartilages and timepoints (Table 9), 00m Tibia showed the uppermost amount of all three effect sizes, most notably accounting for approximately 76% of all observed large effect sizes.

4.3.1. Grey Levels

The amount of non-negligible effect sizes obtained with all texture features for each grey level quantization scheme is reported in Table 10.

Dynamic - dynamic quantization (**-1_-1**) exhibited the highest amount of small effect sizes and medium effect sizes. However, *dynamic - static* -1_500 quantization found the top amount of large effect sizes, which is more than 10% higher than the second highest amount observed with -1_-1. Additionally, *static - dynamic* quantization (0_-1) was second best in terms of the amount of medium effects. *Special* grey level quantization schemes show competitive results, especially 0-05_f00_t00_f36_t36 in terms of large effect sizes and 0-10_all_different in terms of small effect sizes.

4.3.2. Bins

The total amount of non-negligible effect sizes found per each bin quantization schemes is reported in Table 11. The highest amount of small and medium effect sizes were identified using a **4-bin** quantization. However, the amount of large effect sizes peaked with **8-bin** quantization and subsequently proceeded decrease with higher bin quantization schemes.

The amount of large effect sizes per single bin quantization scheme were observed for each texture feature and plotted in Figure 15. Using *Cluster Prominence, Cluster shade, Correlation, Information Measure of Correlation 1, Information Measure of Correlation 2, Sum of square variance* and *Sum variance* found more than 100 large effects for at least one bin count. *Cluster prominence, Cluster shade, Maximum probability, Sum of square variance* and *Sum variance* showed an overall increase while *Information measure of correlation 1* showed an overall decrease in the amount of large effect sizes with increasing bin quantization number.

Table 10. The amount of δ effect sizes per each grey level quantization scheme

Grey level quantization schemes	**Totals of effects sizes**		
	small δ effects **(0,147$\leq$ δ $<$0,33)**	**medium δ effects** **(0,33$\leq$ δ $<$0,474)**	**large δ effects** **(0,474$\leq$ δ)**
-1_-1[1]	2871[3]	1607	601
-1_300	2428	937	568
-1_400	2703	921	569
-1_500	2711	913	687
-1_600	2890	969	597
-1_700	3028	831	613
0_-1	2420	1334	611
0_300	2711	751	423
0_400	2628	717	397
0_500	2343	744	516
0_600	2648	652	511
0_700	2888	642	361
0_800	2810	612	318
0-02_f_t[2]	2369	675	437
0-05_f_t	2378	763	451
0-10_f_t	2547	829	458
0-02_f00_t00_f36_t36	2246	639	580
0-05_f00_t00_f36_t36	2226	629	664
0-10_f00_t00_f36_t36	2227	693	638
0-02_all_different	2926	710	468
0-05_all_different	3172	799	599
0-10_all_different	3142	826	623
average	2650	827	531

[1] Using bold-font schemes yielded above-average totals for all three effect sizes. -1 indicates dynamic grey level assignment.

[2] 0_02, 0_05, 0_10 indicate cutoff percentages used to shorten the cartilage vector and assign the grey levels. f_t, f00_t00_f36_t36 and all_different utilize averaged grey ranges.

[3] The table is vertically color-coded. The cell color saturation is directly proportional to the cell value.

Table 11. The amount of δ effect sizes observed for each bin quantization scheme

Bin quantization scheme	**Amount of effect sizes**		
	small δ effects **(0,147$\leq$ δ $<$0,33)**	**medium δ effects** **(0,33$\leq$ δ $<$0,474)**	**large δ effects** **(0,474$\leq$ δ)**
4	10302*	3651	1907
8	9870	3224	2136
12	9502	2864	2129
16	9548	2740	2054
32	9555	2844	1765
64	9535	2870	1699

* The table is vertically color-coded. The cell color saturation is directly proportional to the cell value.

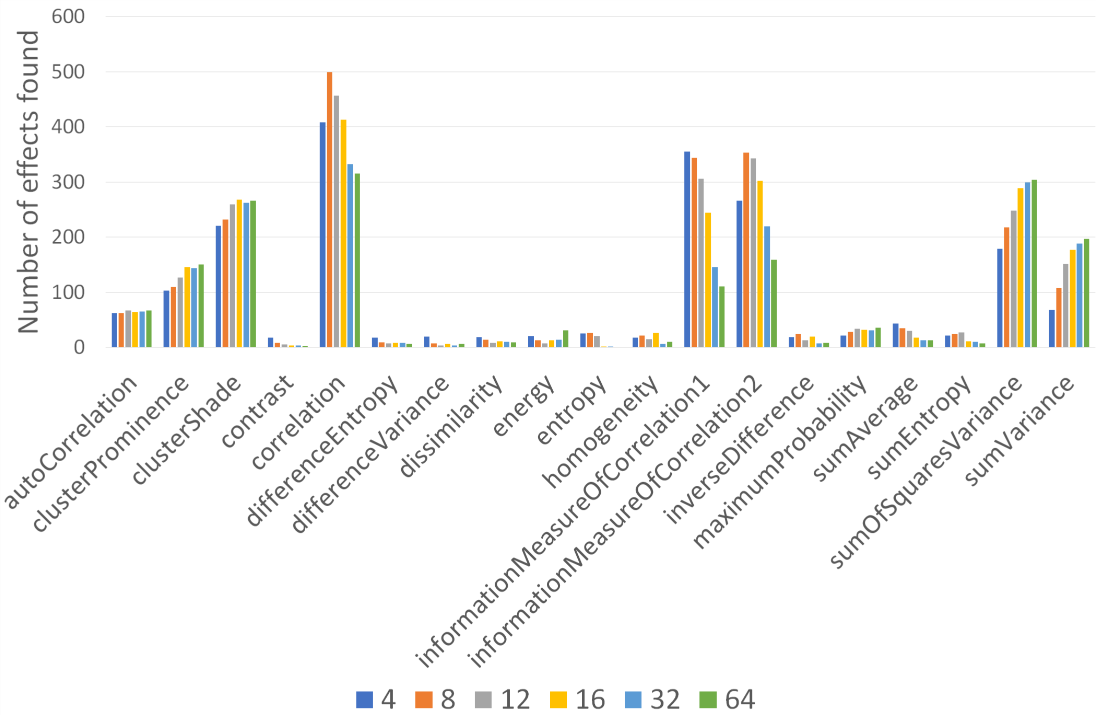


Figure 15. The amount of large effect sizes for each texture feature using different bin quantization schemes. The x-axis shows every bin quantization scheme for each texture feature. Y-axis marks the amount of observed large effect sizes.

4.3.3. Offset

The total amount of non-negligible effect sizes found per each bin quantization schemes is reported in Table 12. The highest amounts of small and medium effect sizes were observed with offset 4. The highest amounts of large effects were found with offset 2, then with offset 4 as the second best, then offset 3 and lastly offset 1.

The amount of large effect sizes observed for each feature and separate offset setting is shown in Table 13. Using *correlation* and offset 2 yielded the highest amount of large effect sizes. The highlighted rows in Table 13 mark those features of which their total value exceeded the average total. Figure 16 visualizes the amount of effect sizes found across the texture features with different offsets.

Those features with totals exceeding the average total were used for the selected training feature set for the subsequent machine learning analysis. Based on the results from Table 13 and Figure 16C, features *Autocorrelation*, *Cluster prominence*, *Cluster shade*, *Correlation*, *Information measure of correlation 1*, *Information measure of correlation 2*, *Sum of square variance* and *Sum variance* were selected. Although the total for *Autocorrelation* (387) did not exceed the average total, the second best feature after *Autocorrelation* was the *Maximum probability* with the total value of 183 across offsets, which is less than a half below the *Autocorrelation* result.

Table 12. The amounts of δ effect sizes per each offset setting

Offset setting	**Amount of effects sizes**		
	small δ effects (0,147$\leq$ δ <0,33)	**medium δ effects (0,33$\leq$ δ <0,474)**	**large δ effects (0,474$\leq$ δ)**
offset 1	14519[*]	4577	2377
offset 2	13766	4241	3343
offset 3	14677	4482	2782
offset 4	15350	4893	3188

* The table is vertically color-coded. The cell color saturation is directly proportional to the cell value.

Table 13. The amounts of large δ effect sizes found using individual texture features per offset setting

Texture feature	**Sum of large δ effect sizes (0,474$\leq$ δ) per offset setting**				
	offset 1	**offset 2**	**offset 3**	**offset 4**	**Total**
autoCorrelation[*]	59	140	97	91	387
clusterProminence	80	249	210	241	780[**]
clusterShade	237	446	398	427	1508
contrast	10	6	11	15	42
correlation	474	741	601	608	2424
differenceEntropy	22	7	13	14	56
differenceVariance	10	6	8	23	47
dissimilarity	27	15	13	16	71
energy	44	23	15	17	99
entropy	19	31	12	15	77
homogeneity	32	14	19	32	97
informationMeasureOfCorrelation1	266	397	367	476	1506
informationMeasureOfCorrelation2	375	415	364	489	1643
inverseDifference	33	12	21	25	91
maximumProbability	66	47	35	35	183
sumAverage	43	53	33	23	152
sumEntropy	37	30	11	23	101
sumOfSquaresVariance	272	416	422	427	1537
sumVariance	271	295	132	191	889
average	125,1	175,9	146,4	167,8	615,2

* Bold italic font indicates features selected to be part of the selected feature subgroup for the machine learning analysis.

** Green marked fields indicate features with the totals above the average total.

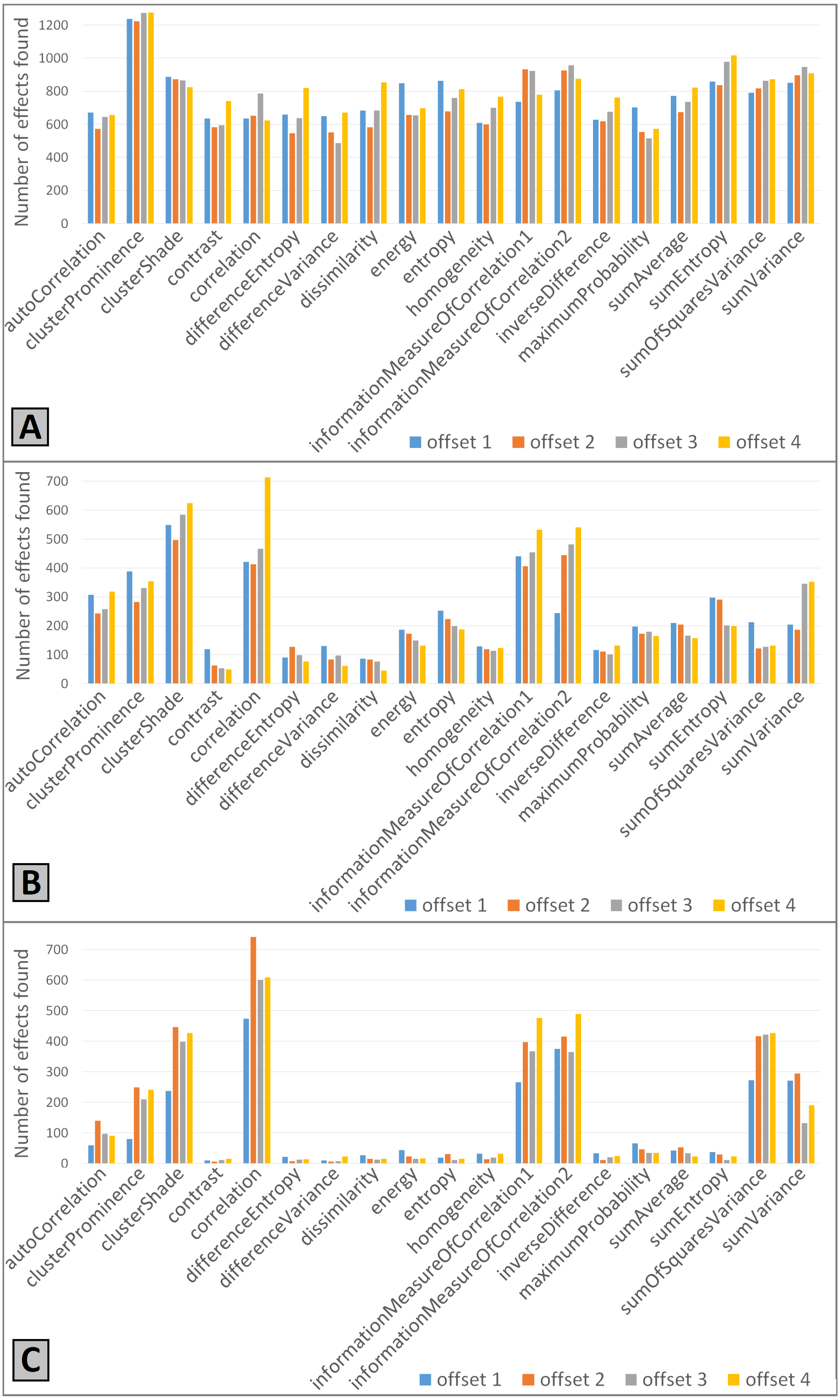


Figure 16. The amount of effects found with each offset setting for individual texture features. Figure A depicts the amount of small δ effects. Figure B shows the amount of medium δ effects. Figure C shows the amount of large δ effects.

4.4. Machine Learning Analysis

Altogether, 200 machine learning pipeline outputs were collected, 50 for each offset. 96 results were calculated using the Naive Bayes (NB) algorithm. Other 96 were achieved by utilizing Suport Vector Machines (SVM). Finally, 8 results were extracted using a Multi-layer Perceptron (MLP). Table 14 describes the collected results and the varying configuration parameters between the individual results. NB and SVM were both collected with and without cross-validation. MLP was the most time consuming, taking over 24 hours to finish the results. NB and SVM were much faster, rendering the output in approximately 12 hours each.

Table 14. List of the collected outputs from the machine learning pipeline

Classifier	**Analyzed cartilages**	**Features utilized**	**Boot-strapping**	**Data split utilized**	**Cross-validation (10-folds)**
Multilayer Perceptron	femur; tibia	selected[1]	yes	pair-wise[2]	no
Naive Bayes	femur; tibia	all; selected	yes; no	pair-wise; random[3]	yes; no
Support Vector Machines	femur; tibia	all; selected	yes; no	pair-wise; random	yes; no

[1] Selected features include *Autocorrelation, Cluster prominence, Cluster shade, Correlation, Information measure of correlation 1, Information measure of correlation 2, Sum of square (Variance)* and *Sum variance*.
[2] Pair-wise method selects a single CTRL-PRGS pair as the testing set.
[3] Random method selects 1-3 random samples as the testing set.

A list of best performers according to their robustness score was drawn for each machine learning pipeline output. This yielded altogether 200 small tables, each containing all 3D Texture Analysis outputs with their robustness scores for each layer. The tables were sorted and only ten best performers were kept. The highest robustness score (87%) was achieved by a *dynamic - static* -1_500 with a 4-bin quantization, offset 2, using Naive Bayes with only selected features from tibia, no bootstrapping and pair-wise data split. The complete results are shown in Table 15.

Differences between between the best performing and worst performing 3D Texture Analysis outputs were collected from each ML pipeline output. The average difference between the best and worst performers was $25,42 \pm 10,23$ %. The maximum measured difference was $62,92\%$.

To evaluate the proposed feature selection, average robustness scores were calculated separately for results collected with selected features and all features. The differences between the average robustness scores from selected-based results vs all-based results are shown in Table 16. Using selected features showed a decrease in robustness for femur-based predictions in layers L10, L50 and SUM. A positive impact of using selected features was observed for tibia-based predictions from all layers.

MLP provided the highest femur robustness score ($76{,}5\%$; offset 2; selected features; bootstrapped; pair-wise), however tibia showed superior results overall. In terms of the average robustness score, tibia outperformed femur by: $10{,}72\%$ for L10; $19{,}74\%$ for L50; $17{,}41\%$ for L90; and $21{,}48\%$ for SUM.

Table 15. Robustness scores containing the highest achieved robustness obtained with Naive Bayes, 3D Texture Analysis outputs based on offset 2, only selected features from tibia, pair-wise data split, 2-sample bootstrapping and no cross-validation

L10		**L50**	
3D Texture Analysis output	**Robustness r [%]**	**3D Texture Analysis output**	**Robustness r [%]**
’0_700_12’	77,3 %*	’-1_500_4’	87,0 %
’-1_-1_12’	75,8 %	’-1_300_64’	86,5 %
’0-05_all_different_8’	75,8 %	’0_300_8’	86,3 %
’0_-1_12’	74,5 %	’-1_400_32’	83,0 %
’-1_600_16’	74,0 %	’0-10_all_different_64’	83,0 %
’0-10_f00_t00_f36_t36_12’	73,8 %	’0_600_16’	83,0 %
’0-02_f_t_12’	73,5 %	’-1_700_8’	82,5 %
’0_-1_4’	73,5 %	’0-10_f_t_8’	82,3 %
’0-02_all_different_8’	72,5 %	’0-05_f_t_32’	81,3 %
’0-10_f_t_16’	72,5 %	’0-10_f_t_64’	81,0 %

L90		**SUM**	
3D Texture Analysis output	**Robustness r [%]**	**3D Texture Analysis output**	**Robustness r [%]**
’0_600_12’	78,8 %	’0_500_8’	83,8 %
’0_700_12’	78,5 %	’0_800_64’	82,0 %
’0-02_f_t_64’	78,3 %	’-1_500_4’	81,3 %
’-1_-1_8’	77,8 %	’0-02_f00_t00_f36_t36_8’	80,5 %
’0_600_4’	76,8 %	’0-02_f_t_8’	80,3 %
’0-10_f_t_64’	76,5 %	’0_600_4’	80,3 %
’-1_500_8’	76,0 %	’0_800_12’	80,3 %
’0-10_f00_t00_f36_t36_64’	75,5 %	’-1_400_16’	79,3 %
’-1_-1_4’	75,3 %	’0-05_f00_t00_f36_t36_12’	79,3 %
’-1_700_12’	75,0 %	’-1_400_32’	79,0 %

* Robustness represents the average accuracy across 200 training and testing repetitions.

Table 16. Percentual differences in average robustness scores for different cartilages between results collected with selected features versus all features

Cartilage	**L10**	**L50**	**L90**	**SUM**
Femur	-3,22 %*	-2,03 %	3,11 %	-0,12 %
Tibia	3,14 %	5,08 %	2,52 %	4,66 %

* Green color indicates a positive impact of the selected feature set. Red color indicates a negative impact.

Overall best performers were collected from all predictions, femur-based predictions and tibia-based predictions according to the number of times each 3D Texture Analysis output appeared amongst the ten best performers for each machine learning output. Table 17 shows only the 20 best and worst performers overall. The highest amount of appearances was counted for -1_500_4. Second and third place was then occupied by the 0_-1 grey level quantization scheme. Out of the 20 best performers, 13 were based on dynamic grey level assignment and only two (no. 16 and 19) were utilizing the special average-based grey level quantization scheme. The median bin number for the twenty best performers was 8; The median bin number for the twenty worst performers was 32. The complete bin distribution across the sorted best performers is depicted in figure 17.

Table 17. Best overall performers and worst overall performers with offset combined

	Best of all ML* results	**Best of FEMUR-based ML results**	**Best of TIBIA-based ML results**		**Worst of all ML results**
1	’-1_500_4’**	’0_-1_12’	’-1_500_4’	**113**	’0-02_f_t_16’
2	’0_-1_12’	’0_700_4’	’0_-1_4’	**114**	’0-05_f_t_12’
3	’0_-1_64’	’0_-1_32’	’-1_500_8’	**115**	’0-02_f00_t00_f36_t36_16’
4	’0_700_4’	’0_-1_64’	’0_500_16’	**116**	’0-02_f00_t00_f36_t36_32’
5	’0_-1_32’	’0_700_8’	’-1_700_12’	**117**	’0-05_all_different_64’
6	’0_700_8’	’0_-1_8’	’-1_600_12’	**118**	’0_500_64’
7	’0_-1_4’	’-1_600_4’	’-1_700_8’	**119**	’0-05_all_different_12’
8	’0_-1_8’	’0_800_4’	’-1_400_8’	**120**	’0-05_all_different_32’
9	’-1_-1_4’	’-1_-1_4’	’-1_400_4’	**121**	’0-02_all_different_16’
10	’-1_-1_8’	’0_-1_16’	’0_600_16’	**122**	’-1_500_64’
11	’-1_600_4’	’-1_-1_8’	’0_400_12’	**123**	’-1_700_64’
12	’0_800_4’	’0-10_f_t_4’	’-1_400_16’	**124**	’0-02_f_t_64’
13	’0_-1_16’	’-1_500_4’	’-1_700_16’	**125**	’-1_400_64’
14	’0_600_4’	’0_400_4’	’0_600_4’	**126**	’0_300_16’
15	’-1_500_8’	’0_800_8’	’0_700_12’	**127**	’0_600_32’
16	’0-10_f_t_4’	’0-10_f00_t00_f36_t36_4’	’-1_400_12’	**128**	’0-02_all_different_12’
17	’-1_700_12’	’0-02_f00_t00_f36_t36_4’	’-1_600_16’	**129**	’0_800_64’
18	’0_500_16’	’-1_-1_12’	’0_700_16’	**130**	’0_700_32’
19	’0-10_f00_t00_f36_t36_4’	’0-05_f00_t00_f36_t36_4’	’0_800_12’	**131**	’-1_300_32’
20	’-1_-1_12’	’0-10_all_different_16’	’0_500_12’	**132**	’-1_300_64’

* Machine Learning

** -1 indicates dynamic grey level assignment. 0_02, 0_05, 0_10 indicate cutoff percentages used to shorten the cartilage vector and assign the grey levels. f_t, f00_t00_f36_t36 and all_different utilize averaged pixel intensities as minimum and maximum grey level.

Average robustness scores were calculated for each offset and compared against the average robustness scores from offset 1. Table 18 demonstrates the percentual impact of utilizing higher offset settings compared to offset 1. Offset 3 improved the L10 robustness scores by approximately $3,41\%$, while offset 4 positively impacted the average robustness scores of L50 by $0,93\%$, L90 by $3,15\%$ and SUM by $1,79\%$.

The average robustness scores achieved by a given classifier per each layer, separate for femur and tibia, are shown in Table 19. The highest average robustness was achieved with Naive Bayes using tibial layer L50. In terms of femur, using MLP yielded the best average robustness scores for femoral L10 and L50, $62{,}8\%$ and $63{,}2\%$ respectively.

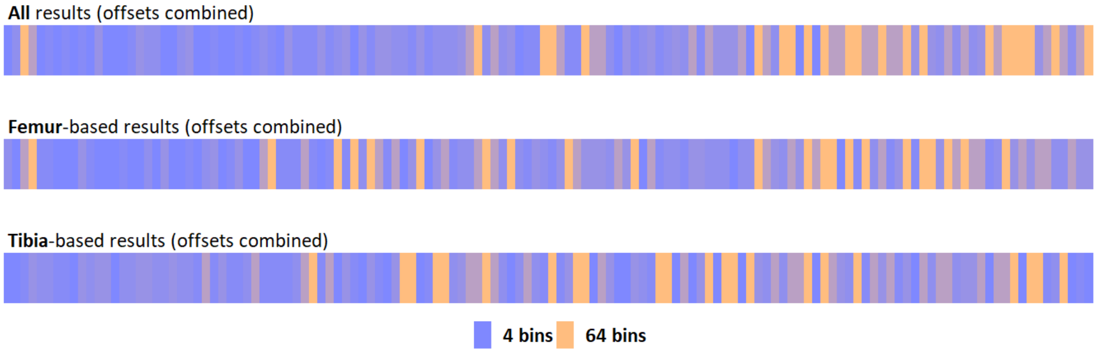


Figure 17. Bin distribution across the best performers overall performers with offset combined. The scales range from the most robust performer (left) to the least robust performer (right). Purple color corresponds to the 4-bin quantization scheme and beige color corresponds to the 64-bin quantization scheme.

Table 18. Average improvements in robustness scores for offsets 2,3 and 4 compared to offset 1 for each cartilage layer

Offset setting	L10	L50	L90	SUM
Offset 1	-	-	-	-
Offset 2	1,83 %*	0,61 %	0,86 %	-0,73 %
Offset 3	3,41 %	0,66 %	0,66 %	0,69 %
Offset 4	2,51 %	0,93 %	3,15 %	1,79 %

* Green color marks the highest improvement.
Red color marks the highest deterioration.

Table 19. Average robustness scores for each cartilage and their corresponding layers generated by the utilized classifiers

	Femur				Tibia			
Classifier*	**L10**	**L50**	**L90**	**SUM**	**L10**	**L50**	**L90**	**SUM**
MLP	62,8 %**	63,2 %	53,5 %	58,6 %	62,7 %	70,3 %	64,6 %	66,6 %
NB	53,7 %	53,8 %	48,5 %	48,1 %	67,2 %	77,5 %	71,1 %	77,4 %
SVM	28,7 %	29,5 %	26,1 %	29,4 %	36,6 %	45,4 %	38,3 %	43,0 %

* MLP - Multilayer perceptron; NB - Naive Bayes; SVM - Support Vector Machines
** The cell color saturation is directly proportional to the cell value.

5. DISCUSSION

The initial cartilage histogram analysis provided valuable information about the pixel distributions within the cartilage. Both femur and tibia seemed to follow a pattern in terms of their histograms and can be clearly differentiated from each other by mere visual inspection. The major finding, however, is the increase in the maximum pixel intensity over time for both femur and tibia. Unfortunately, it is only a speculation if such longitudinal change is associated with physiological changes or if it is simply a result of the reduced reproducibility of the measurements over the years. The fact that there is a significant increase for both CTRL and PRGS subjects indicates that it does not reflect the progression of the disease. All the cartilages were averaged into a single cartilage and a cumulative histogram was created. The 100% cumulation was achieved before the 800 pixel intensity mark and therefore, a static-static grey level range from 0 to 800 would provide a coverage of all the pixel intensities present within the studied dataset. The large difference in intensity between the 99% (399) cumulation and 100% cumulation is almost 400 intensity values. This observation suggests a presence of outliers within the pixel distribution. Both femur and tibia showed a difference larger than 100 intensity values between their original maximum intensity value and the maximum intensity value after a 2% cutoff. The differences between 2%, 5%, and 10% cutoffs are less than 20 intensity values. This result suggests that the 2% cutoff should be enough to remove the extremities and therefore was chosen as the cutoff percentage for all the dynamic grey level quantization schemes.

Before any outputs were collected, the algorithm of the 3D Texture Analysis was tweaked in order to accommodate for multiple grey level and bin quantizations to be calculated at a time. The algorithm update was possible due to the fact that the grey level boundaries and the bin number are independent from the interpolation and extrapolation tasks. Therefore, the GLCMs can be calculated based on the same interpolation data. Such update resulted in approximately $6\times$ faster computation time, which means that around 83% of the computation time was reduced. Additional improvements that could be implemented are adjusting the measurements of the computation time and making it resilient to interruptions.

The statistical analysis provided not only a glimpse into the importance of the input parameters for the 3D Texture Analysis of 3D DESS images to differentiate between the controls and the progressive cohort, but also outlined the abilities of individual texture features to differentiate between CTRL and PRGS subjects. The approach was based on calculating the amounts of the effect sizes found per the studied layers (L10, L50, L90 and SUM), cartilage types and timepoints (00m Femur, 00m Tibia, 36m Femur, 36m Tibia), grey level quantization schemes, bin quantization schemes and offsets. Additionally, charts indicating the total amounts of effect sizes found per bin quantization scheme and offset setting for each texture feature were analyzed. The approach of calculating the amount of measured effect sizes is in itself a questionable method. Although finding a large amount of effect sizes with certain combination of parameters might indicate superior results, a large quantity might not be strictly associated with better quality. The contrast between quality and quantity was pointed out with the differences between the total amounts of large effect sizes found per texture feature. Even though the differences between the texture features might be visible from the small effect size chart, the large effect sizes clearly accentuate those differences and make them much more interpretable. Also, the limited number of samples might have a negative influence on the effect size calculation and the estimated numbers might not be as precise and reliable as they might be with a larger sample size. [73]

The study of quantized cartilage layers showed that L50 and L10 might provide the most insight into the OA differences. This result might be due to the fact, that cartilage tends to be denser and contain less fluid towards its base and therefore deeper layers could provide more detail into the intra-cartilaginous changes. Further research is recommended to establish new methodologies how to effectively combine the information from all the layers and achieve maximum predictive potential.

In terms of the cartilage data from different timepoints, results from tibia at baseline showed the highest amount of medium and large effect sizes. This is a surprising result, because by logic, the most amount of differences would be expected at the 36 month follow-up time point, where the progressive subjects had developed their OA significantly while the control subjects stayed approximately the same and therefore the differences should theoretically get accentuated. Although not following the initial expectation, this result supports the primary application of the 3D Texture Analysis software in this thesis and that is to differentiate between CTRL and PRGS subject at baseline. However, the amount of effect sizes found using femurs from baseline is around twice as lower than the amount found in femurs from 36 month screenings. Nonetheless, more than 70% of all the effect sizes detected at baseline were collected by utilizing the tibial cartilage.

The statistical results for the studied **grey level** quantization schemes indicated superiority of the dynamic grey level assignment. Dynamic grey level quantization assigns the grey levels according to the pixel intensity distribution of the analyzed cartilage, which seemed to benefit the calculations of δ effect sizes.

The statistical evaluation of **bins** suggested that smaller number of bins might be beneficial. 4 bin quantizations yielded the most amount of small and medium effect sizes. However, large effect sizes benefited mostly from 8 up to 16 bin quantization schemes. The benefit of smaller number of bins follows the preconceived expectations and is caused likely due to the reduced amount of noise within such schemes. The higher the number of bins, the higher the noise effect to the GLCM. This might subsequently result in worsening the predictive power. However, the bin number has an inconsistent effect on the individual texture features. For example, features *Information Measure of Correlation 1* seemed to benefit from smaller number of bins in terms of large δ effect sizes. On the other hand, *Sum of Squares (Variance)* found the most large effect sizes from 64-bin quantized outputs. Overall, the range between 4 and 16 bins is probably more beneficial in general applications. Lower number of bins does not allow much noise to affect the GLCMs, which might be the reason for their better performance.

The highest amount of effect sizes was observed with **offset** 4. However, the offset study provided a way to extract features for a selected feature subset to be utilized for the machine learning. Only features that had a total of identified large effect sizes higher than the average total from all features were selected.

At the point of writing this thesis, only linear bin quantization was supported by the 3D Texture Analysis. Adjusting the bin size according to the histogram density might also have an impact on the 3D Texture Analysis output features. In 2017, Di et. al. [74] implemented non-linear GLCM quantization in order to improve seismic texture analysis and their results show a potential for the non-linear approach. Therefore, further research might attempt to study and implement non-linear quantization methods and their impact on the knee OA prediction.

The machine learning analysis provided a more practical evaluation of the input parameters and shows how they might influence the early prediction of knee OA based on the cartilage data. Only subject data from baseline was used to train and test the classifiers. The entire ML pipeline was constructed so that various types of machine learning approaches would be utilized and, as a results, the various 3D TA outputs would show their robustness. Due to the small amount of subject data, the ML algorithms seemed to be prone to overfitting. Therefore, the training and testing phase for each 3D TA output was repeated $200\times$ with randomly selected training and testing subjects and random initial weights. This was crucial for the study, because the ultimate goal of the machine learning analysis was not to find the algorithm which performs the best, but rather to see, which 3D TA outputs perform the best and showed overall robust results and kept their classification power regardless the chosen ML method.

The inclusion of multiple classifiers and the entire machine learning pipeline provided a rigorous testing site for the 3D Texture Analysis outputs. Multilayer perceptron was the most time-consuming, however provided the best possible results from Femur (on average 63,2% from L50). The 2 hidden layers each containing 11 neurons created a powerful neural network, which might have pushed the model towards overfitting and therefore diminishing the predictive potential. Naive Bayes, on the other hand, showed the best results overall (on average 77,5% from L50, 77,4% from SUM). Naive Bayes exceled with its simplicity and therefore low time consumption. SVM showed the worst performance, however that is most likely due to the fact that no kernel function was utilized for the training. This, on one side, significantly sped up the collection of the results, but on the other, cost the possible predictive power. Probably the most relevant possibility for a future research is the optimization of machine learning algorithms for the knee osteoarthritis prediction using the 3D Texture Analysis outputs. The study could utilize some of the findings from this thesis and expand the knowledge by maximizing the predictive power of the various machine learning methods.

The feature selection for the machine learning might benefit from further testing. On average tibia-based predictions showed improvements by using only the selected feature subset, however femur-based predictions suffered. In the future, some more sophisticated methods for feature selection might be implemented and analyzed, for example Principal Component Anlaysis (PCA). The predictions based on tibial features significantly outperformed the predictions based on femur. Similarly to the cartilage layers, a possible future study focusing on combining the femoral and tibial features to maximize the predictive potential would provide the necessary understanding and yet again push the osteoarthritis prediction forward.

The machine learning results indicated that dynamic grey level quantization schemes yield improved predictions. This finding supports the results from the statistical analysis. The dynamic grey level assignment is slightly more difficult to implement, however the results support its effectiveness. *0 - dynamic* was placed amongst the 20 best overall performers with all 6 bin quantization schemes. It is the only grey level quantization approach which placed itself amongst the performers with bin quantizations 32 and 64. *Dynamic - static* showed above average results in the statistical analysis. The table of best overall performers also includes *static - static* quantizations 0_700 and 0_800, which support the recommendations by Brynolfsson et. al. [11] to use a static grey level quantization which encapsulates all the pixel intensities present within the ROIs. *Static - static* grey level quantization is the simplest to use, however the *0 - dynamic* approach seems to provide the same results plus seems to be more resilient and capable of the same robustness with higher bin quantization schemes. The special approach was included to see, if a customized grey range for different cartilages, timepoints and study groups yield superior differentiation and prediction. Although their results were competitive, the dynamic approach seemed to provide better results. Moreover, in real-life clinical setting, the special approach is not optimal. The idea is to find a simple grey level quantization scheme, which can independently extract the most amount of information from a single subject at a single time point.

The machine learning analysis further supports the smaller bin quantization schemes. As seen from the overall best performers, despite the *0-dynamic 64-bin* second best performer, a lower number of bins seems to correlate with better prediction of OA subjects. Additionally, low bin quantization schemes are faster to calculate due to smaller GLCMs. This is another reason to opt for a lower number of bins, especially on large datasets.

The results for the **offset** indicate that using higher offset setting might be beneficial. Although higher offset seems to correlate with better machine learning output, it is not very clear which offset number might yield the best result. Based on the results, offset settings 2 and higher can be recommended. The comparison of offsets in terms of robustness improvement also showed that offsets 3 and 4 might be better, however that evaluation is based on heavy averaging. Although the indication is supported by the statistical results, further research is necessary to generalize the offset setting. Follow-up studies might be interested in collecting the 3D TA outputs from various offset settings and averaging those outputs into one. Such protocol increases the calculation time, since each offset requires its own TA run. However, provided there is a good computational power, an average output across various offset might provide valuable insights.

The importance of the input parameters was demonstrated by measuring the difference in robustness between the best performing and worst performing 3D Texture Analysis outputs in each collected ML pipeline output. The average measured difference was approximately 25% and the maximum found difference was over 60%. In other words, there was an extreme case of more than 60% difference in robustness between the 3D Texture Analysis outputs applied with the same machine learning parameters. The average difference of 25% basically says, that by optimizing the 3D TA input parameters, the predictive performance can increase on average by 25%.

This thesis was a subject to a number of limitations. The subject dataset in this thesis consisted of only 20 subjects screened at two timepoints (baseline and 36 month follow-up). The small sample size might be arguably one of the biggest limitations of this study due to its probable impact on the classifiers. However, a limited amount of subjects was necessary in order to collect all the desired 3D Texture Analysis outputs in the given time. Moreover, the available computer hardware might struggle severely with a larger dataset. Future studies with larger sample sizes should definitely be considered. The number of simultaneously calculated grey level and bin quantizations was also limited by the computer hardware. Unfortunately, the 3D TA is exclusively CPU operation and cannot be accelerated by a graphics card. Therefore, more CPU cores and higher clock speeds should result in a reduction of the computation time and larger RAM should be able to accommodate for more quantization schemes to be calculated simultaneously.

6. CONCLUSION

This thesis attempted to optimize the input parameters of a novel GLCM-based 3D Texture Analysis software in order to maximize its predictive capabilities of knee osteoarthritis. The studied parameters were assumed to have a significant impact on the output texture features and, therefore, have a potential to influence the early prediction of knee OA. The software was applied onto a study dataset of 10 control subjects and 10 progressive subjects containing cartilage data from the baseline and the 36-month follow up screening. The outputs were compared in terms of their ability to identify significant δ effects between the subject cohorts. The machine learning phase attempted to predict the knee OA purely from the baseline data, where both control and progressive subjects showed no signs of the disease.

Adjusting the grey level quantization scheme according to the cartilage pixel intensity distribution had a positive impact on the predictive power of the calculated texture features. Bin quantization schemes utilizing 4 to 12 bins were found to not only require less amount of computation time but also yield the most robust predictions. Although offsets > 1 were associated with improved results, the optimal offset setting remains undetermined and further investigation is recommended.

The results indicated that the 3D Texture Analysis input parameters can have a significant impact on the output features and consequently on their predictive power. The presented results also prove that the 3D Texture Analysis software holds a solid potential for accurate early prediction of knee osteoarthritis from the baseline data. The findings of this thesis might provide some guidance for the possible future research activities utilizing the 3D Texture Analysis software. The goals of this thesis have therefore been achieved. Further studies, especially focusing on optimization of the machine learning methods, should improve and stabilize the predictive capabilities of the software.

7. REFERENCES

[1] Mora J.C., Przkora R. & Cruz-Almeida Y. (2018) Knee osteoarthritis: pathophysiology and current treatment modalities. Journal of pain research, 11 p. DOI: `https://doi.org/10.2147/JPR.S154002`.

[2] Ayhan E., Kesmezacar H. & Akgun I. (2014) Intraarticular injections (corticosteroid, hyaluronic acid, platelet rich plasma) for the knee osteoarthritis. World J Orthop. DOI: `https://doi.org/10.5312/wjo.v5.i3.351`.

[3] Dulay S.G., Cooper C. & Dennison E.M. (2015) Knee pain, knee injury, knee osteoarthritis & work. Best Practice & Research Clinical Rheumatology. DOI: `https://doi.org/10.1016/j.berh.2015.05.005`.

[4] Lespasio M.J., Piuzzi N.S., Husni M.E., Muschler G.F., Guarino A. & Mont M.A. (2017) Knee Osteoarthritis: A Primer. The Permanente journal, 21, 16–183. DOI: `https://doi.org/10.7812/TPP/16-183`.

[5] McRobbie D., Moore E., Graves M. & Prince M. (2006) MRI From Picture to Proton. Cambridge University Press.

[6] Eckstein F., Cicuttini F., Raynauld J., Waterton J. & Peterfy C. (2006) Magnetic resonance imaging (MRI) of articular cartilage in knee osteoarthritis (OA): morphological assessment. Osteoarthritis and Cartilage, Volume 14, Supplement 1. DOI: `https://doi.org/10.1016/j.joca.2006.02.026`.

[7] Haralick R.M., Shanmugam K. & Dinstein I. (1973) Textural Features for Image Classification. IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-3, no. 6, pp. 610-621. DOI: `https://doi.org/10.1109/TSMC.1973.4309314`.

[8] Nailon W.H. (2010) Texture Analysis Methods for Medical Image Characterisation. Intech Publishing, Biomedical Imaging. DOI: `https://www.researchgate.net/publication/221907976_Texture_Analysis_Methods_for_Medical_Image_Characterisation`.

[9] Väärälä A. (2018) 3D Texture Analysis of Knee Articular Cartilage Using Isotropic MR Images. University of Oulu, 44 p.

[10] Schooler J., Kumar D., Nardo L., McCulloch C., Li X., Link T. & Majumdar S. (2014) Longitudinal evaluation of T1rho and T2 spatial distribution in osteoarthritic and healthy medial knee cartilage. Osteoarthritis and Cartilage, Volume 22, Issue 1, Pages 51-62. Https://doi.org/10.1016/j.joca.2013.10.014.

[11] Brynolfsson P., Nilsson D., Torheim T., Asklund T., Thellenberg-Karlsson C., Trygg J., Nyholm T. & Garpebring A. (2017) Haralick texture features from apparent diffusion coefficient (ADC) MRI images depend on imaging and pre-processing parameters. Sci Rep 7, 4041. DOI: `https://doi.org/10.1038/s41598-017-04151-4`.

[12] Peuna A., Hekkala J., Haapea M., Podlipská J., Guermazi A., Saarakkala S., Nieminen M. & Lammentausta E. (2018) Variable angle gray level co-occurrence matrix analysis of T2 relaxation time maps reveals degenerative changes of cartilage in knee osteoarthritis: Oulu knee osteoarthritis study. Imaging, 47: 1316-1327. DOI: `https://doi.org/10.1002/jmri.25881`.

[13] Gomez W., Pereira W.C.A. & Infantosi A.F.C. (2012) Analysis of Co-Occurrence Texture Statistics as a Function of Gray-Level Quantization for Classifying Breast Ultrasound. IEEE Transactions on Medical Imaging, vol. 31, no. 10, pp. 1889-1899. DOI: `https://doi.org/10.1109/TMI.2012.2206398`.

[14] Osteoarthritis initiative (oai) study protocol. URL: `https://nda.nih.gov/oai/study-details`.

[15] Osteoarthritis initiative. URL: `https://www.niams.nih.gov/grants-funding/funded-research/osteoarthritis-initiative`. Last updated: July 2020.

[16] Affatato S. (2015) Biomechanics of the knee. Surgical Techniques in Total Knee Arthroplasty and Alternative Procedures 2015, Pages 17-35. DOI: `https://doi.org/10.1533/9781782420385.1.17`.

[17] A. H. (2017) Clinical Anatomy - Knee. YouTube. URL: `https://www.youtube.com/watch?v=0R95KGJc0GY`.

[18] Wang M.L. & Peng Z.X. (2015) Wear in human knees. Biosurface and Biotribology, Volume 1, Issue 2, June 2015, Pages 98-112. DOI: `https://doi.org/10.1016/j.bsbt.2015.06.003`.

[19] Sophia Fox A.J., Bedi A. & Rodeo S.A. (2009) The basic science of articular cartilage: structure, composition, and function. Sports health, 1(6), 461–468. DOI: `https://doi.org/10.1177/1941738109350438`.

[20] Vincent K.R., Conrad B.P., Fregly B.J. & Vincent H.K. (2012) The pathophysiology of osteoarthritis: a mechanical perspective on the knee joint. PM & R : the journal of injury, function, and rehabilitation, 4(5 Suppl), S3–S9. DOI: `https://doi.org/10.1016/j.pmrj.2012.01.020`.

[21] Hsu H. & Siwiec R. (2020) Knee Osteoarthritis. StatPearls. URL: `https://www.ncbi.nlm.nih.gov/books/NBK507884/`.

[22] uncredited A. (2020) Guide to Severe Knee Osteoarthritis. Spring Loaded. URL: `https://springloadedtechnology.com/guide-to-severe-knee-osteoarthritis/`.

[23] Kellgren J. & Lawrence J. (1957) Radiological assessment of osteo-arthrosis. Annals of the rheumatic diseases. DOI: `https://doi.org/10.1136/ard.16.4.494`.

[24] Kohn M.D., Sassoon A.A. & Fernando N.D. (2016) Classifications in Brief: Kellgren-Lawrence Classification of Osteoarthritis. Clinical Orthopaedics and Related Research. DOI: https://doi.org/10.1007/s11999-016-4732-4.

[25] Ding C., Cicuttini F. & Jones G. (2007) Tibial subchondral bone size and knee cartilage defects: relevance to knee osteoarthritis. Osteoarthritis and Cartilage, Volume 15, Issue 5, May 2007, Pages 479-486. DOI: https://doi.org/10.1016/j.joca.2007.01.003.

[26] Thakkar R.S., Flammang A.J., Chhabra A., Padua A. & Carrino J.A. (2011) 3T MR Imaging of Cartilage using 3D Dual Echo Steady State (DESS). MAGNETOM Flash. URL: https://mriquestions.com/uploads/3/4/5/7/34572113/3t_mr_imaging_of_cartilage_using_3d_dual_echo_steady_state-00011808.pdf.

[27] Castelvechi D. (2020) Just a moment. Nature Physics. Link: https://www.nature.com/articles/s41567-020-1022-6.

[28] Kaunitz J.D. (2018) Magnetic Resonance Imaging: The Nuclear Option. Digestive diseases and sciences vol. 63,5. DOI: https://doi.org/10.1007/s10620-018-4992-9.

[29] Rabi I.I., Zacharis J.R., Millman S. & Kusch P. (1938) A New Method of Measuring Nuclear Magnetic Moment. American Physical Society. DOI: https://doi.org/10.1103/PhysRev.53.318.

[30] Lindley D. (2006) Landmarks: NMR–Grandmother of MRI. Physical Review Focus 18, 18. Link: https://physics.aps.org/story/v18/st18.

[31] Purcell E.M., Torrey H.C. & Pound R.V. (1946) Resonance Absorption by Nuclear Magnetic Moments in a Solid. Physical review. DOI: https://doi.org/10.1103/PhysRev.69.37.

[32] Bloch F., Hansen W.W. & Packard M. (1946) The Nuclear Induction Experiment. Physical review. Link: https://mri-q.com/uploads/3/4/5/7/34572113/bloch._nuclear_induction_experiment_1946.pdf.

[33] Lauterbur P.C. (1973) Image Formation by Induced Local Interactions: Examples Employing Nuclear Magnetic Resonance. Nature 242. Link: https://www.nature.com/articles/242190a0.

[34] Chodos A., Ouellette J. & Tretkoff E. (2006) July: This Month in Physics History. APS Physics. Link: https://www.aps.org/publications/apsnews/200607/history.cfm#:~:text=The%20first%20images%20were%20produced,soft%20tissues%20like%20the%20brain.

[35] Edelman R.D. (2014) The History of MR Imaging as Seen through the Pages of Radiology. RSNA. DOI: https://doi.org/10.1148/radiol.14140706.

[36] Grover V.P.B., Tognarelli J.M., Crossey M.M.E., Cox I.J., Taylor-Robinson S.D. & McPhail M.J.W. (2015) Magnetic Resonance Imaging: Principles and Techniques: Lessons for Clinicians. Journal of Clinical and Experimental Hepatology. DOI: https://doi.org/10.1016/j.jceh.2015.08.001.

[37] Bitar R., Leung G., Perng R., Tadros S., Moody A.R., Sarrazin J., McGregor C., Christakis M., Symons S., Nelson A. & Roberts T.P. (2006) MR Pulse Sequences: What Every Radiologist Wants to Know but Is Afraid to Ask. RadioGraphics, Vol. 26, No. 2. DOI: https://doi.org/10.1148/rg.262055063.

[38] Hammer M. (2014), Mri physics: K-space trajectories. URL: http://xrayphysics.com/traject.html.

[39] Bruder H., Fischer H., Graumann R. & Deimling M. (1988) A new steady-state imaging sequence for simultaneous acquisition of two MR images with clearly different contrasts. Magn Reson Med. 1988 May;7(1):35-42. DOI: https://doi.org/10.1002/mrm.1910070105.

[40] Abraham C.L., Bangerter N.K., McGavin L.S., Peters C.L., Drew A.J., Hanrahan C.J. & Anderson A.E. (2015) Accuracy of 3D dual echo steady state (DESS) MR arthrography to quantify acetabular cartilage thickness. Journal of magnetic resonance imaging : JMRI, 42(5), 1329–1338. DOI: https://doi.org/10.1002/jmri.24902.

[41] Kimpe T. & Tuytschaever T. (2006) Increasing the Number of Gray Shades in Medical Display Systems—How Much is Enough? Journal of Digital Imaging. DOI: https://doi.org/10.1007/s10278-006-1052-3.

[42] Castellano G., Bonilha L., Li L.M. & Cendes F. (2004) Texture analysis of medical images. Clinical Radiology, Volume 59, Issue 12. DOI: https://doi.org/10.1016/j.crad.2004.07.008.

[43] Vallières M., Freeman C.R., Skamene S.R. & El Naqa1 I. (2015) A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. Phys. Med. Biol. 60 (2015) 5471–5496. DOI: https://doi.org/10.1088/0031-9155/60/14/5471.

[44] Leijenaar R.T.H., Nalbantov G., Carvalho S., Van Elmpt W.J.C., Troost E.G.C., Boellaard R., Aerts H.J.W.L., Gillies R.J. & Lambin P. (2015) The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis. Sci Rep. 2015 Aug 5;5:11075. DOI: https://doi.org/10.1038/srep11075.

[45] Carballido-Gamio J., Stahl R., Blumenkrantz G., Romero A., Majumdar S. & Link T.M. (Spatial analysis of magnetic resonance T1rho and T2 relaxation times improves classification between subjects with and without osteoarthritis.) 2009. Med Phys. 2009 Sep;36(9):4059-67. DOI: https://doi.org/10.1118/1.3187228.

[46] Williams A., Winalski C.S. & Chu C.R. (2017) Early articular cartilage MRI T2 changes after anterior cruciate ligament reconstruction correlate with later changes in T2 and cartilage thickness. J Orthop Res. 2017 Mar;35(3):699-706. DOI: https://doi.org/10.1002/jor.23358.

[47] Joseph G., Baum T., Carballido-Gamio J., Nardo L., Virayavanich W., Alizai H., Lynch J.A., McCulloch C.E., Majumdar S. & Link T.M. (2011) Texture analysis of cartilage T2 maps: individuals with risk factors for OA have higher and more heterogeneous knee cartilage MR T2 compared to normal controls - data from the osteoarthritis initiative. Arthritis Res Ther 13, R153. DOI: https://doi.org/10.1186/ar3469.

[48] Zayed N. & Elnemr H.A. (2015) Statistical Analysis of Haralick Texture Features to Discriminate Lung Abnormalities. International Journal of Biomedical Imaging, 1687-4188. DOI: https://doi.org/10.1155/2015/267807.

[49] Blumenkrantz G., Stahl R., Carballido-Gamio J., Zhao S., Lu Y., Munoz T., Graverand-Gastineau M.P.H.L., Jain S.K., Link T.M. & Majumdar S. (2008) The feasibility of characterizing the spatial distribution of cartilage T(2) using texture analysis. Osteoarthritis and cartilage, 16(5), 584–590. DOI: https://doi.org/10.1016/j.joca.2007.10.019.

[50] Li X., Pai A., Blumenkrantz G., Carballido-Gamio J., Link T., Ma B., Ries M. & Majumdar S. (2009) Spatial distribution and relationship of T1rho and T2 relaxation times in knee cartilage with osteoarthritis. Magn Reson Med. 2009 Jun;61(6):1310-8. DOI: https://doi.org/10.1002/mrm.21877.

[51] Materka A. & Strzelecki M. (2015) On The Effect Of Image Brightness And Contrast Nonuniformity On Statistical Texture Parameters. Foundations of Computing and Decision Sciences, Volume 40, Issue 3. DOI: https://doi.org/10.1515/fcds-2015-0011.

[52] Kokkotis C., Moustakidis S., Papageorgiou E., Giakas G. & Tsaopoulos D.E. (2020) Machine learning in knee osteoarthritis: A review. Osteoarthritis and Cartilage Open, Volume 2, Issue 3. DOI: https://doi.org/10.1016/j.ocarto.2020.100069.

[53] Guyon I., Gunn S., Nikravesh M. & Zadeh L.A. (2008) Feature extraction: foundations and applications (Vol. 207). Springer. URL for book preview: https://books.google.fi/books?hl=en&lr=&id=FOTzBwAAQBAJ&oi=fnd&pg=PA1&dq=feature+extraction&ots=5Ug9N9ari_&sig=GNFwo2sJMnHOEp8FyC5j8edqwhc&redir_esc=y#v=onepage&q=feature%20extraction&f=false.

[54] Brownlee J. (2016) Master Machine Learning Algorithms. Machine Learning Mastery. URL: https://books.google.fi/books?id=n--oDwAAQBAJ&printsec=copyright&redir_esc=y#v=onepage&q&f=false.

[55] DeNero J., Klein D., Abbeel P. & et. al. (2017) The Pac-Man projects. UC Berkeley CS188 Intro to AI – Course Materials. URL: http://ai.berkeley.edu/project_overview.html#:~:text=Overview,techniques%20to%20playing%20Pac%2DMan.&text=Instead%2C%20they%20teach%20foundational%20AI,probabilistic%20inference%2C%20and%20reinforcement%20learning.

[56] Brownlee J. (2018) A Gentle Introduction to k-fold Cross-Validation. Statistics, Machine Learning Mastery. URL: https://machinelearningmastery.com/k-fold-cross-validation/.

[57] Brownlee J. (2016) Overfitting and Underfitting With Machine Learning Algorithms. Machine Learning Algorithms. URL: https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/.

[58] Alpaydin E. (2004) Introduction to machine learning. Cambridge, Mass.: MIT Press. URL: https://books.google.fi/books?id=1k0_-WroiqEC&lpg=PP1&dq=machine%20learning%20books&pg=PA234#v=onepage&q&f=false.

[59] Brownlee J. (2016) How to Code a Neural Network with Backpropagation In Python (from scratch). Code Algorithms From Scratch. URL: https://machinelearningmastery.com/implement-backpropagation-algorithm-scratch-python/.

[60] Du Y., Almajalid R., Shan J. & Zhang M. (2018) A Novel Method to Predict Knee Osteoarthritis Progression on MRI Using Machine Learning Methods. IEEE Transactions on NanoBioscience, vol. 17, no. 3, pp. 228-23. DOI: https://doi.org/10.1109/TNB.2018.2840082.

[61] Deokar D.D. & Patil C.G. (2015) Effective Feature Extraction Based Automatic Knee Osteoarthritis Detection and Classification using Neural Network. International Journal of Engineering and Techniques - Volume 1 Issue 3, May - June 2015. URL: https://www.semanticscholar.org/paper/Effective-Feature-Extraction-Based-Automatic-Knee-Deokar-Patil/46e7236ccf2a8d40cf726bd730e87152f6a29f09.

[62] Panfilov E., Tiulpin A., Klein S., Nieminen M.T. & Saarakkala S. (2019) Improving robustness of deep learning based knee mri segmentation: Mixup and adversarial domain adaptation. 2019 International Conference on Computer Vision Workshop, ICCVW 2019. DOI: https://doi.org/10.1109/ICCVW.2019.0005.

[63] Clausi D.A. (2002) An analysis of co-occurrence texture statistics as a function of grey level quantization. Canadian Journal of Remote Sensing 28, 45–62. DOI: https://doi.org/10.5589/m02-004.

[64] Soh L.K. & Tsatsoulis C. (1999) Texture Analysis of SAR Sea Ice Imagery Using Gray Level Co-Occurence Matrices. IEEE Transactions on Geoscience and

Remote Sensing, Volume 37, Issue 2. DOI: https://doi.org/10.1109/36.752194.

[65] Romano J., Kromrey J.D., Coraggio J., Skowronek J. & Devine L. (2006) Exploring methods for evaluating group differences on the NSSE and other surveys: Are the t-test and Cohen's d indices the most appropriate choices? Southern Association for Institutional Research. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.595.6157&rep=rep1&type=pdf.

[66] Sullivan L. (2017) Nonparametric Tests. Boston Univesity School of Public Health. Link: https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Nonparametric/index.html.

[67] Cliff N. (1993) Dominance statistics: Ordinal analyses to answer ordinal questions. Psychological Bulletin. DOI: https://doi.org/10.1037/0033-2909.114.3.494.

[68] Hentschke H. & Stüttgen M.C. (2018) Matlab Toolbox 'Measures of Effect Size'. GitHub. Mathworks File Exchange: https://se.mathworks.com/matlabcentral/fileexchange/32398-hhentschke-measures-of-effect-size-toolbox.

[69] Cohen J. (1988) Statistical Power Analysis for the Behavioral Sciences. Lawrence Erlbaum Associates. Link: http://www.utstat.toronto.edu/~brunner/oldclass/378f16/readings/CohenPower.pdf.

[70] Chen D., Chen X., Li H., Xie J. & Mu Y. (2019) DeepCPDP: Deep Learning Based Cross-Project Defect Prediction. IEEE Access. DOI: https://doi.org/10.1109/ACCESS.2019.2961129.

[71] Eraqi H. (2016) MLP Neural Network with Backpropagation. MATLAB Central File Exchange. URL: https://www.mathworks.com/matlabcentral/fileexchange/54076-mlp-neural-network-with-backpropagation.

[72] Brownlee J. (2017) Estimate the Number of Experiment Repeats for Stochastic Machine Learning Algorithms. Statistics, Machine Learning Mastery. URL: https://machinelearningmastery.com/estimate-number-experiment-repeats-stochastic-machine-learning-algorithms/.

[73] Whitley E. & Ball J. (2002) Statistics review 6: Nonparametric methods. Critical care (London, England), 6(6), 509–513. DOI: https://doi.org/10.1186/cc1820.

[74] Di H. & Gao D. (2017) Non-linear GLCM texture analysis for improved seismic facies interpretation. Interpretation 5(3):1-34. DOI: https://doi.org/10.1190/int-2016-0214.1.