The Modelling and Analysis of Comet Assay Data

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy
by

Selina Dhinsey

Department of Physics
University of Liverpool
March 2022

Abstract

Approximately half the UK population will be diagnosed with cancer at some point in their lives, therefore effective treatment is crucial. Using ionising radiation to damage the DNA of cancer cells to such an extent that they are unable to reproduce, and subsequently die, is one favoured treatment. Cells are able to repair some types of damage, so it is important that cancer treatment is based on a good understanding of cell repair mechanisms, particularly for protons and heavy ions where there is less clinical experience. Therefore, measurement methods that allow assessment of the degree and type of radiation damage are needed. One such method is the Comet Assay.

The Comet Assay involves fixing cells in agarose, chemical treatment to free the DNA, electrophoresis, staining, followed by imaging. Image analysis allows the degree of DNA damage to be determined for individual cells. Here, a range of classical and neural net-based methods have been tested, with regard to their comet identification and measurement capabilities.

A complete and automated analysis chain has been established, allowing the damage measured to be output in a spread sheet. Care has been taken to maximise the number of comets measured in each image, whilst ensuring these measurements are as accurate as possible. Due to the high variability of comet structures, an accompanying output plot is also provided to the user, to enable cross referencing of measurements.

A Monte Carlo model was developed, producing simulated Comet Assay data to aid the training and testing of neural network models, and also to provide data with which the performance of all analysis methods could be tested. A model was designed with parameterisation of DNA mobility in agarose. The simulation was successful in producing Assay images that mimic cells with varying levels of DNA damage, however it was not accurate enough to perform the desired tests. This indicates that further study is required to produce a more complex model, incorporating the motion of DNA through agarose in a more realistic manner.

It was found that there is a high degree of variability in Comet Assay data, originating from both the experimental and imaging procedures, affecting the DNA damage measures acquired. Image exposure, contrast and contamination affect DNA damage measurements and this was found to present differently across the analysis methods explored. The neural network architecture coped well with image contamination, but not with large changes in image contrast and exposure. Classical analysis methods were found to be more robust in their measurements. Recommendations are made for producing a standardised Comet Assay protocol and encouraging researchers to outline both their experimental and imaging procedures, to more accurately convey results.

Acknowledgements

There have been a number of people that have helped shape the last four and a half years, making the tough times better, and the good times great. I am so thankful for everyone I have connected with through my PhD research and all that I have learned. It's incredibly difficult to articulate how much you all mean to me, but I'll attempt to.

I'd like to start by thanking my supervisors; Prof Tim Greenshaw, Dr Jason Parsons and Prof Carsten Welsch for their support and guidance. Tim particularly, you have been exactly what I've needed in a supervisor. I truly appreciate every time you've accommodated me and made the space to discuss and share ideas, listen to my concerns and support me as an all-round mentor. Thank you for your patience and calming nature; you've been integral in me making it through to the end of this journey.

Furthermore, my gratitude extends to Jason and his lab team, for answering my cellular biology related questions and conducting the Comet Assay experiments that supplied the data analysed in this work.

Thank you to STFC for funding this work, as part of the LIV.DAT CDT.

My peers, LIV.DAT friends in particular, have also been integral in my PhD experience. Thank you for always making things more bearable and walking this journey alongside me.

I extend my thanks to Naomi Mwasabili and Megan Charles for providing me with the opportunity to take part in such great work at Chanua. You have both been such sources of strength and kindness, especially during lockdown.

Thank you, Milena. You continue to inspire and push me to be the best version of myself.

My family are the biggest definers of my character, and have helped me build the resilience necessary to make it through this project, and are the only people to have seen me in all of my various states.

Thank you to my mother-in-law, Susan, for all your love, support and always being there to listen.

Tom, you are my anchor and stability, and have often been the sole source of keeping some form of balance in my life, especially through the hardest days. Thank you for never failing to make me laugh and for all the small moments. You truly mean the world.

Parji, thank you for the belief you have in me, and for always having my back.

To my grandparents, thank you for all your sacrifice. None of this would be possible without your determination and willpower to always succeed. I am the culmination of a dream to lift up the succeeding generation.

And finally, but most importantly, to my mother, Culdip. None of this exists without you. Thank you for your unwavering and fierce belief in me, especially when I haven't had it for myself. Thank you for all the sacrifice that has made all of this possible, and for continually being my source of strength. No words could ever describe the value of our relationship, and this thesis is dedicated to you.

ਮੇਰੀ ਮਾਂ ਨੂੰ

Contents

List of Figures

List of Tables

Chapter 1

Introduction

As a society we face many pressing issues. Growing populations and climate change coupled with an ageing population globally present real challenges in terms of healthcare. Cancer is one such healthcare concern to which tremendous research efforts have been dedicated for over 200 years. One of the most difficult aspects of cancer treatment is the individualised and specific nature of each case; the way in which a tumour develops and affects each person can be very different. This requires specialised treatment that has been curated with an individual in mind, rather than for a particular cancer type.

In 1775, initial research by Percivall Pott [1] investigated the causes of cancer, identifying relationships between chimney soot exposure and the incidence of squamous cell carcinoma of the scrotum in chimney sweeps. Developments were then made in understanding how cancer presents in the body through inflammation by Rudolf Virchow in 1863 [2], followed by the first breast cancer treatment via radical mastectomy by William Halsted in 1894 [3]. Preceding the turn of the century came the discovery of X-rays by Wilhelm Röntgen in 1895 [4], followed by Marie and Pierre Curie discovering the radioactive elements in polonium and radium [5], which were then used in cancer treatments. The treatment of cancer really took off in the 20th century after the mechanism through which tumours arise was proposed by Theodor Boveri in 1902 [6], outlining the processes of chromosome damage leading to uncontrollable cell division. This vital work has led to much research in the field of DNA, chromosome abnormalities and individual gene studies as we push to understand the specific causes and triggers of cancer and whether they can be predicted prior to tumour development.

Not only are we striving to understand these precise mechanisms that lead to cancer development, but it is also vital to create targeted treatment therapies, allowing for the best chance of total cancer eradication. It is therefore also of utmost importance to understand the cellular effect of different treatments, how effective they are in fatally damaging DNA and how this is expressed biologically and clinically.

1.1 Cancer is Caused by Structural Changes in DNA

Cancer is a disease resulting in the frantic and uncontrollable division of cells in the body. It is caused by structural changes to DNA through which a cell's normal processes are overridden, resulting in a tumour. It can affect bodily functions in a range of ways and its prevalence amongst the population is on the rise. During the period 2016-2018 there were 375 400 new cases of cancer diagnosed per year in the UK, with a predicted total incidence to increase by 2% between 2014 and 2035 [7]. It is therefore of critical importance that the mechanisms through which cancer arise and develop are thoroughly understood to create preventative treatments.

Changes in DNA can be caused by a number of factors that result in a genetic mutation. This is when a change in the DNA sequence occurs, which can cause abnormal processes in the body. Mutations can happen during incorrect cell division, discussed in Section 1.2.1, which lead to future daughter cells containing the mutated gene. External factors such as radiation can physically damage the bonds that hold DNA together, and even knock out some bases. This type of damage can be lethal and completely kill the cell if it is unable to sufficiently repair the DNA strands, causing it to be the aim of cancer treatment, and further discussed in Section 1.3.
To fully comprehend how mutations affect DNA, we first need to understand what DNA is, how it is structured and the role it plays.

1.2 DNA

Deoxyribonucleic acid (DNA) is the blueprint of all eukaryotic cells and dictates their function, structure, enzyme production and even their lifespan. It was first identified as 'nuclein' in 1869 by Swiss physician Friedrich Miescher [8], whilst the landmark discovery of the internal structure of DNA's double helix was made in 1953 by Watson and Crick [9]. This breakthrough was only possible through utilising the X-ray diffraction work conducted by Rosalind Franklin [10], [11], [12], shown in Figure 1.1. Today, this type of experiment is such a standard procedure that it is regularly performed in undergraduate Physics laboratory experiments [13].

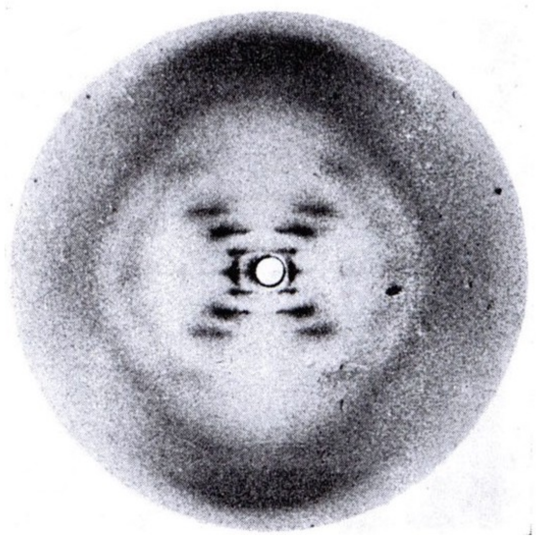


Figure 1.1: *Franklin's X-ray diagram of the B form of sodium thymonucleate (DNA) fibres showing the characteristic features of helical structures. Adapted from [10].*

DNA, depicted in Figure 1.2 is a molecule consisting of two long polynucleotide strands. Each double stranded unit of DNA is known as a chromosome. Somatic cells (all human cells of the body except sperm, egg and red blood cells) are diploid, meaning their nuclei contain 23 pairs of chromosomes (one inherited from each parent), providing a total of 46. DNA is only visible as chromosomes when it condenses, as cells enter mitosis - the process of cell division. Otherwise it is not observable and is "diffuse and dispersed within the nucleus" [14].

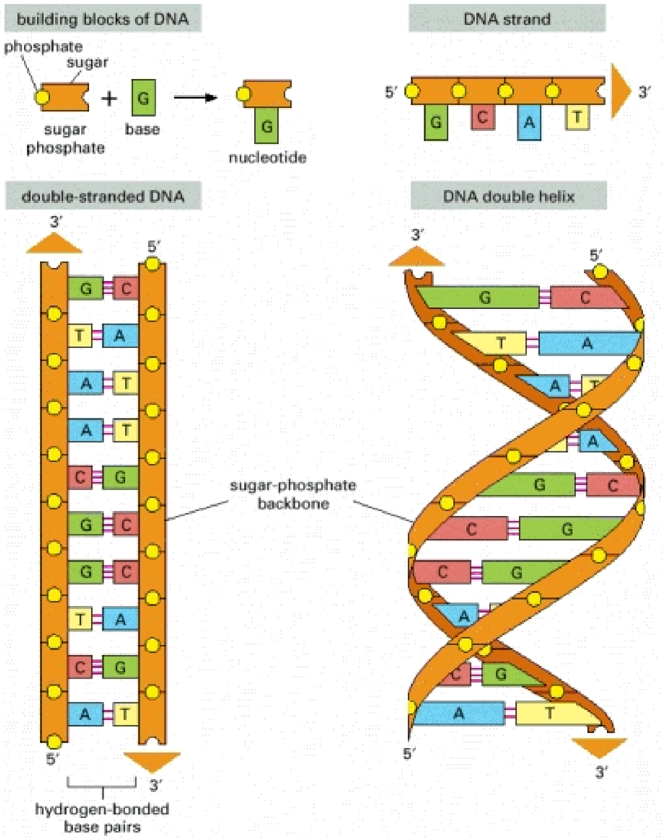


Figure 1.2: *The structure of DNA and how the sugar phosphate backbone and base pairings are arranged to form a double stranded polynucleotide in a double helical structure, [15].*

Each strand of DNA is formed of many sub-units existing in four different types. The sub-units, known as nucleotides consist of a phosphate sugar molecule (forming the DNA backbone) and one of four nitrogenous base molecules; adenine (A), thymine (T), cytosine (C) and guanine (G) [15]. Uracil is the 5th nitrogenous base but it is only found in RNA structures, replacing thymine. The bases fall into two types; adenine and guanine are purines whereas thymine and cytosine are pyrimidines. These pair up in a specific and complementary way so that adenine will always pair with thymine whilst cytosine will bond with guanine, producing purine-pyrimidine couples. This is due to the chemical structure of the bases, shown in Figure 1.3. Whilst thymine and adenine are held together by two intermolecular hydrogen bonds, cytosine and guanine are joined by three of these bonds. This specific bonding of the base pairs is instrumental in the DNA replication phase of the cell cycle and for cellular repair mechanisms, discussed further in Section 1.2.2. It is also the cause of DNA's overall structure; hydrogen bonded bases inside whilst the outside is composed of a sugar-phosphate backbone.

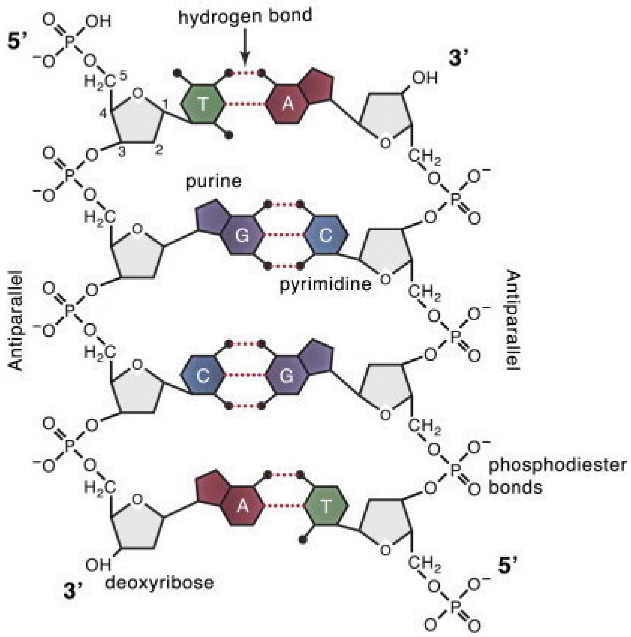


Figure 1.3: *A view of DNA's chemical structure in terms of the nucleotides, comprising of phosphate sugar and base units and joined together with phosphodiester bonds. Hydrogen bonds between complementary bases are shown with dotted red lines. Adapted from [16].*

The phosphate sugar molecule is the supplement of each nucleotide unit. Nucleotides are joined through phosphodiester bonding between a phosphate ion and deoxyribose molecule (lending its name to the D in DNA). As shown in the top left deoxyribose molecule of Figure 1.3, the carbon atoms are numbered from 1'-5' in an anti-clockwise fashion. The 3' and 5' carbons are of most interest as they are the positions at which phosphodiester bonds form. This sequential 5'3' bonding throughout a DNA strand gives it polarity, also know as directionality, and dictates that each DNA strand has a free 3' end and a free 5' end [16]. Due to this, an individual DNA strand can be characterised with either a 5'-3' or 3'-5' directionality and because they are complements of one another, they run in anti-parallel directions. Figure 1.3 also shows that the phosphate ions are negative due to free, unbonded electrons on the O^- ions. This has the effect of causing the overall charge of the backbone and thus the DNA molecule to be negative, an essential characteristic for electrophoresis

during the Comet Assay procedure, detailed in Section 1.4.
The specific bonding of the bases is what gives DNA its helical structure, which is right-handed and of B-form in "normal" eukaryotic cells. Each turn of DNA contains about 10-10.5 nucleotides and is ∼3.4 nm in length. The distance between individual nucleotides is ∼0.34 nm [16].

1.2.1 The Cell Cycle

DNA is stored within the cell's nucleus, where it replicates and then divides into newly formed cells during the final stage of the cell cycle. A cell's life cycle is broken down into two main stages. The first is where DNA replication occurs and is known as the S phase. To prepare for replication, the strands are separated, an easy process as they are only held together by weak hydrogen bonds between bases. This is the most energetic and time consuming part for the cell as it makes an identical copy of its 46 chromosomes, lasting 10-12 hours and taking up around half of the cell cycle in mammalian cells [17].
Replication is performed by utilising the complementary nature of DNA strands. Only one strand is required to create the other, simply by using it as a template and assigning the associated base in the correct place in the sequence. For example, if one single strand has base sequence CCTGACTGG at a certain position, the partner strand will have GGACTGACC at the complementary position. This allows an exact and precise copy to be made, requiring only the original DNA and the base/sugar phosphate building blocks to construct the new strands. The duplication via template methodology is depicted in Figure 1.4.

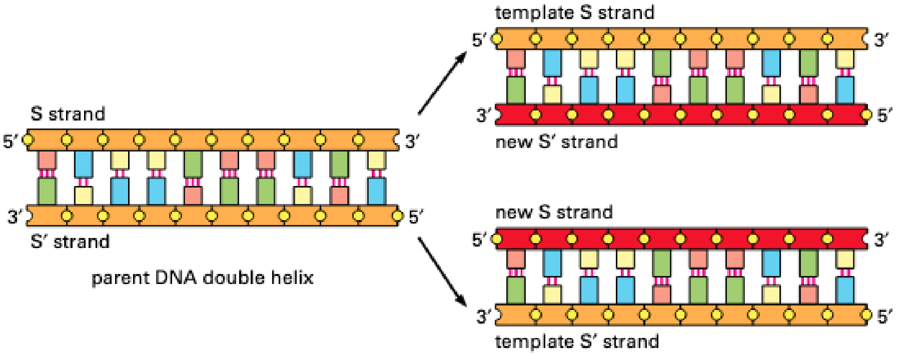


Figure 1.4: *A representation of how DNA replication is performed by using the parent strands as templates to build new, complementary strands to produce exact copies of the parent DNA [18].*

Following the replication period, or S phase, cell division occurs comparatively quickly taking around only an hour during the M phase, Figure 1.5. This phase gets its name from mitosis, the process of nuclear division. Cytokinesis follows mitosis within the M phase, in which a daughter cell is produced by the splitting of the cell's cytoplasm. At the beginning of mitosis, DNA and its replicate condense to form extremely compact chromosomes. They then attach to the mitotic spindle via microtubules and the nuclear envelope breaks down in preparation for the division. During the metaphase, the chromosomes line up along the equator of the mitotic spindle to ensure an even split containing one each of the sister chromatids in the two divided cells. The anaphase is when the replicated sister chromatids are pulled apart to opposite poles of the spindle, decondense and reform separate nuclei. The final stage of mitosis is the telophase and runs into cytokinesis. It is the process in which the mitotic spindle disassembles as the cell is separated to form two cells and the nuclear membrane of each new nucleus reforms.

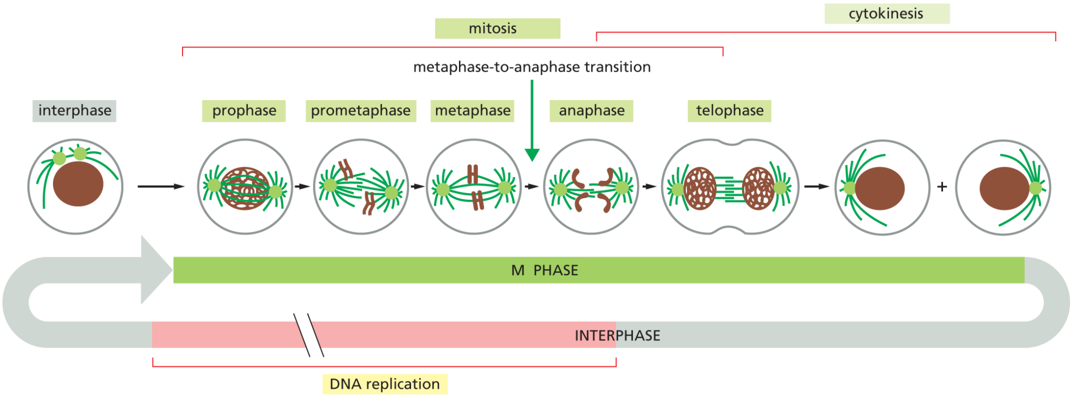


Figure 1.5: *The M phase of the cell cycle, showing a breakdown of the processes that occur during mitosis and cytokinesis, resulting in nuclear and cellular division [17].*

The S and M phases are separated by two growth periods known as the G_1 and G_2 phases. Not only do they serve as periods for cell growth but they also allow for checking the internal and external environments, to ensure conditions are appropriate for cell division to proceed. An overview of the cell cycle is shown in Figure 1.6, depicting the four main stages; G_1 phase, S phase, G_2 phase and M phase.

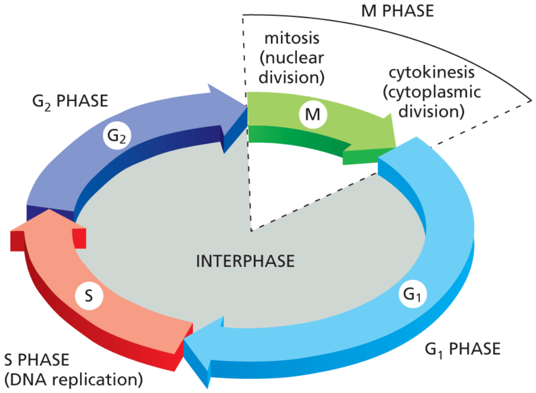

M PHASE
mitosis (nuclear division)
cytokinesis (cytoplasmic division)
G_2 PHASE
G_2
M
INTERPHASE
S
G_1
S PHASE (DNA replication)
G_1 PHASE

Figure 1.6: *The cell cycle, consisting of G_1 phase, S phase where DNA replication occurs, G_2 phase and M phase in which nuclear and cell division take place [17].*

1.2.2 Damage and Repair

The DNA molecule is a complex structure that undergoes continuous day-to-day damage due to internal and external factors such as hydroxyl ions and ionising radiation [19]. Damage can result in structural changes to the DNA which can alter how it functions and in turn, affect how the cell operates e.g. carcinogenesis [20]. Some damage types can cause permanent changes to the nucleotide base sequence, and are known as mutations. They are an integral part of species diversity and evolution, but can also pose a threat to overall functionality of organisms. There are four different types of mutation. Point mutations are the smallest alteration and occur when a single nucleotide base is substituted for another, added or deleted from the DNA sequence [21]. Deletion is where a section of chromosome is deleted and translocation occurs when a segment fractures and attaches to another chromosome. Hereditary mutations occur in germline cells (sperm and egg) that are passed from parent to offspring and are permanently part of the genetic code. This is a spontaneous process caused by errors in DNA replication. Other mutations occur through damage to the DNA via metabolic and environmental sources [21].

Consider the case of transcription. This is a regular cell process in which a segment of DNA is copied into RNA for the synthesis of proteins in order to carry out cellular functions [22]. If the transcribed RNA is “different” due to a mutation in the DNA, different proteins may be produced or, in the most extreme case, no proteins are produced at all. Here, the change in DNA could be inherited and in the worst case scenario of not producing certain proteins, can present in the body as a disease from birth. It could also be the case that this mutation was not inherited and was due to some other factor, that causes a disease later in life.

Cancer is a genetic disease resulting from mutations in somatic cells [23]. It is caused by mutation of specific genes, that can build up over a lifetime, which is why cancer often has a higher incidence in the elderly population [24]. The BRCA1 and BRCA2 gene mutations are heavily linked with the development of breast cancer, and are an example in which a particular gene mutation is linked to a specific cancer [25]. Mutations can be considered a type of DNA damage that cause lasting alterations, but other damage types exist that cells are readily able to reverse, as they are equipped with repair enzymes [15]. Some other damage types present increased challenge, requiring more intricate repair pathways. For example, strand lesions occur in two forms; single strand breaks (SSBs) and double strand breaks (DSBs), with the latter being more difficult to repair.

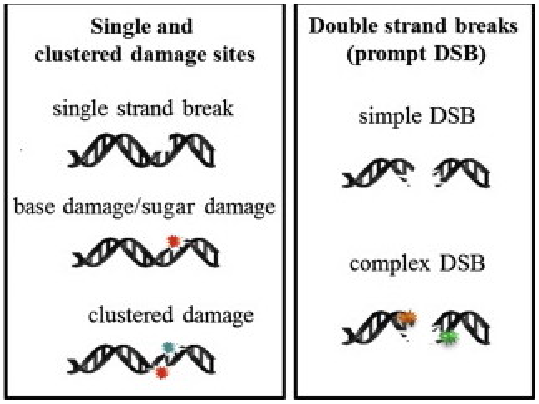


Figure 1.7: *A physical representation of SSBs and DSBs and some complex damage, of both types caused by a clustering of multiple lesions within a small area [26]. The coloured regions depict where complex damage has occurred.*

SSBs occur when only one of the two strands of DNA has a lesion, and DSBs are when both strands experience breaks, Figure 1.7. SSBs are the most common lesion that mammalian cells experience, undergoing tens of thousands per cell, per day [27], translating into one SSB per $\sim 10^5$ bp, or one SSB per $\sim 10^4$ helical turns [28]. Some of these are indirect, for example during the base excision repair (BER) pathway [29], which repairs damage by creating further SSBs in order to reach the lesion site. Attack of the DNA by free radicals, such as reactive oxygen species, is one way SSBs are caused directly. Regardless of the source, a cell's ability to repair SSBs is integral to its survival, thus it has evolved efficient and rapid repair processes. The intact strand can be utilised as a template to reconstruct the base pattern of the damaged nucleotide chain, via mechanisms such as base excision repair and nucleotide excision repair [30][31], which are actively employed during the cell cycle to maintain genetic stability [32].

DSBs arise naturally during cell development, but can also be physically induced through external ionising radiation and chemotherapeutic agents [33]. Their incidence scales linearly with the dose administered, producing $\sim$20-40 DSBs per Gy of dose absorbed. For DSB repair, there is no template for the base order to be reconstructed from, so the cell has to undergo alternate and more complex repair pathways such as non-homologous end-joining and homologous recombination [34]. DSBs can also be created during the DNA replication phase of the cell cycle, due to unrepaired SSBs. This can cause an increase of DSBs with time, as base damage or SSBs are magnified as a consequence of unsuccessful strand transcription. Research suggests that $\sim$1% of the 5×10^3 SSBs that occur during a normal cell cycle are converted into $\sim$50 endogenous DSBs per cell per cycle [35].

Figure 1.7 also depicts some damage known as complex (clustered) damage. This type of damage can refer to SSBs, DSBs and even base damage. Clustered damage is the result of multiple strand breaks forming in a small region (typically a few nm). It is known to be the most damaging to DNA, as its repair process is slow or often even impossible [36]. For example, DNA ligase is an enzyme that joins together DNA lesions [37]. It however requires the break to be "clean", meaning it consists of one 3' hydroxyl end and one 5' phosphate end [38]. This is more unlikely in clustered damage, meaning other more complex repair pathways must be utilised or the lesions will remain. To maximise the efficacy of cancer treatment, it is important that research efforts focus on treating patients with methods that cause the most complex damage to cancerous cells.

1.3 Cancer Treatment

The treatment of cancer has two main focal points; the effectiveness of killing tumours and the unwanted side effects on patients. The overall aim is to maximise the first and minimise the second, creating a compromise between various treatment parameters. For example, when treating cancer using radiation, administering a higher dose provides a better chance of killing the tumour via fatally damaging nuclear DNA, but the impact this dose would have on the patient's non-cancerous tissue would severely compromise their overall health. As a result, the various methods used to treat cancer are continually being modified and refined to increase tumour death and reduce damage to healthy tissue. The main types of cancer treatment are surgery, chemotherapy and radiotherapy. The first is an invasive procedure in which the tumour (or as much as possible) is cut out of a patient, with the aim of removing as many cancerous cells as possible. This presents additional risks and complications, as most surgeries do, whilst also still needing to consider the cancerous cells that may be left behind and continue to grow subsequent tumours. Surgery is often coupled with another treatment in order to reduce the number of cells left behind.

Chemotherapy is a treatment method delivered intravenously or orally, with the aim of stopping cancer cells from growing and spreading within the body [39]. It has many uses, ranging from curative treatment to relieving symptoms through palliative care. As chemotherapy is delivered as a whole body treatment through the blood, with cancerous cells as the target, it also has a lot of unfavourable side effects such as hair loss, fatigue, sore skin and higher risk of infection.
Similarly radiotherapy treatment, using energetic particles/electromagnetic radiation to damage the tumour's DNA and destroy cancerous cells presents its own challenges [40]. Many studies today are looking into the exact biological response caused by radiotherapy, using either X-rays as the damaging agent or ionising radiation such as protons, in particular focusing on the type and amount of damage administered to cellular DNA.

1.3.1 X-ray Radiotherapy

X-rays are a form of high energy electromagnetic radiation that can be utilised in cancer treatment. They were first used clinically at the end of the 19th century [41], shortly after their discovery by Röntgen, published in 1898 [42]. They are effective in damaging DNA by releasing electrons from molecules through indirect ionisations. This is due to their characteristic low LET (linear energy transfer), stemming from charge neutrality. LET, the energy deposited per unit length, is low for X-rays in comparison to high LET radiation such as alpha particles and protons, which are charged. This is due to X-rays interacting indirectly with matter, causing damage by breaking covalent bonds within DNA, producing strand lesions, DSBs in particular.

Starting with initial intensity I_0, X-rays traversing a material with linear attenuation coefficient, μ lose some of their energy and become attenuated to have intensity I, at a depth d, where

$$I = I_0 e^{-(\mu d)}. \tag{1.1}$$

Due to this relationship, when X-rays are used therapeutically in radiotherapy, their initial energy deposition is high, and gradually decreases as they travel through the patient. The result is that X-rays initially deposit a higher amount of energy when in first contact with the body, which can be one of the causes of skin damage such as dryness, soreness and increased sensitivity in cancer patients [43].

The amount of energy deposited by a material, per unit mass is defined as dose, measured in Gray (Gy). Doses of radiation for cancer treatment are often divided into fractions, to limit the damage caused to healthy tissue. Three different fractionation schemes are implemented for treatment; accelerated fractionation [44], hyperfractionation [45] and hypofractionation [46]. Each of these tailors the frequency, amount of dose and time frame treatment is delivered in and is specifically adapted for each patient case.

There are three main interactions in which X-rays can cause DNA damage. The first is the photoelectric effect, first discovered in 1887 by Hertz [47]. This was further built on in 1905 by Einstein [48], in which he outlined how an X-ray photon transferred all its energy to an electron, ejecting it from the atomic shell and out of the atom, proposing the particle-like behaviour of light. Another interaction of X-rays was discovered by Henry Compton in 1923 and thus named Compton scattering [49]. Here, X-rays are inelastically scattered by interactions with outer-shell atomic electrons. Some of the X-ray energy is absorbed by an electron, producing a new scattered photon of lower energy and new direction. If the electron absorbs enough energy to overcome its binding energy, it will be ejected from the atom. The third interaction is pair production, discovered by Blackett and Occhialini in 1933 [50]. The X-rays are completely absorbed, producing electron positron pairs from interactions with atomic nuclei. This requires the X-ray photon to have a minimum energy of 1.022 MeV (the rest mass energy of an electron and positron). The electrons produced via these mechanisms generate free radicals within cells, which are able to react with DNA molecules and cause molecular strand lesions. As eukaryotic cells are 70% water, indirect ionisations account for the majority of radiation-induced damage [51].

The physical geometry in which a tumour's irradiation will occur is carefully considered through the creation of a treatment plan [52] This is when the radiation delivery is simulated in the volume to be irradiated. It is used to calculate the dose that will be absorbed and optimised to deliver enough dose to the cancerous region whilst minimising the dose to healthy tissue. Although treatments plans are optimised to cover the entire tumour volume, a significant dose can still be deposited in healthy tissue encountered along the X-ray path, due to the way they attenuate, an issue that is significantly reduced in proton beam therapy.

1.3.2 Proton Beam Therapy

Proton therapy is a newer area of radiotherapy treatment and its capabilities in treating cancer continue to advance. It was first proposed that protons could be utilised in the clinical environment by Wilson in 1946, where he outlined the physical properties of high energy protons that make them advantageous [53]. He particularly focused on the penetration depth that high energy (up to 200 MeV) protons could reach and the specific, direct ionisation of the matter they interact with. It is well accepted and understood today that protons ionise particles along the path of their track. This ionisation is maximal near the end of their path, just before they stop, as this is the site of maximal energy deposition. Zirkle first highlighted the important relationship between the density of ionisations and the resulting biological effect in 1935 [54]. He discussed how the concentration of ionisations along the path of charged particles is inversely proportional to the energy of the particles, a concept now widely known as LET. Charged particles of higher energy have a lower LET due to their higher velocity producing fewer ionisations along their track, resulting in fewer interactions with matter.

The Physics community is more familiar with the quantity -$\frac{dE}{dx}$ for energy deposition per unit length, rather than LET, which is derived via the Bethe-Bloch equation, Eq. 1.2.

$$-\frac{dE}{dx} = Kz^2\frac{Z}{A}\frac{1}{\beta^2}\frac{1}{2}ln\frac{2m_\mathrm{e}c^2\beta^2\gamma^2T_\mathrm{max}}{I^2} - \beta^2 - \frac{\delta}{2}, \tag{1.2}$$

where

$$K = \frac{4\pi e^4}{c^2 m_\mathrm{e}} N_\mathrm{A}. \tag{1.3}$$

Here, z is the relative charge of the particle, Z is the atomic number of the absorber atoms, A is the atomic mass of the absorber material and β is the relativistic variable defined as $\frac{\text{velocity}(v)}{\text{speed of light}(c)}$. The electron rest mass is m_e and γ is the relativistic variable defined as $\frac{1}{\sqrt{1-\beta^2}}$. T_max is the maximum transferable energy from the charged particle to an electron in the medium it is traversing and I is the characteristic ionisation constant of the material. δ is the density effect correction term, dependent on the particle's velocity, and N_A is Avogadro's constant.

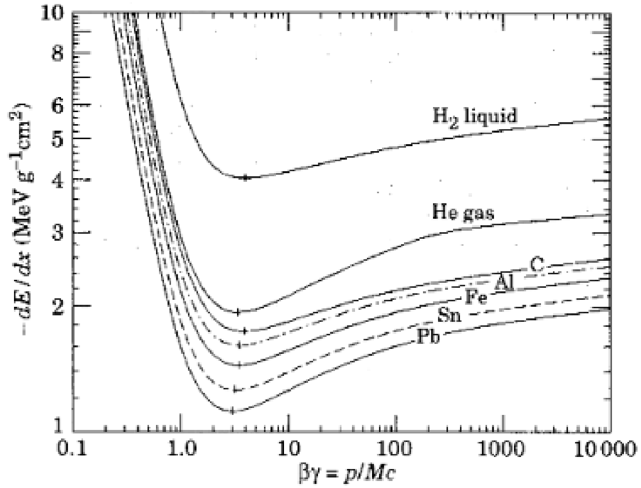


Figure 1.8: *Plot showing the energy loss of protons as a function of momentum (GeV/c) in various materials [55].*

Figure 1.8 depicts the relationship between energy loss per unit length and proton momentum in various materials. A general trend in this relationship can be seen for all materials investigated in which the energy loss decreases rapidly for protons of momentum of up to around 3 GeV/c and then gradually increases with momentum. The spatial range that radiation is able to traverse is of vital importance for clinicians as it determines the amount and possibly the type of tissue that is subjected to radiation during treatment. By integrating the Bethe-Bloch equation we can acquire the radiation range [55], where E' is the initial kinetic energy of the ion.

$$\int_0^E \frac{dE'}{dx}^{-1} dE' \approx \sum_0^E (\frac{dE'}{dx})^{-1} \Delta E' \tag{1.4}$$

The value obtained from Eq. 1.4 is the particle's path length, acting as a good approximation of range for clinical use [55]. In 1905, Bragg and Kleeman realised that the logarithmic range and logarithmic energy had a linear relationship [56]:

$$R(E) = \alpha E^p \tag{1.5}$$

Here, α is a constant of the material in question, E is the proton beam's initial energy and the power p is dimensionless; its value can be found by comparison to solutions of Eq. 1.2. The range equation is based upon the continuous slowing down approximation (CSDA) [57], in which radiation is assumed to continuously deposit energy along its track. By plotting range against dose, we can obtain the familiar Bragg peak graph, Figure 1.9. This plot shows the dose deposition as a function of penetration depth for various types of radiation, with charged particles (here carbon ions and protons) exhibiting particularly interesting characteristics.

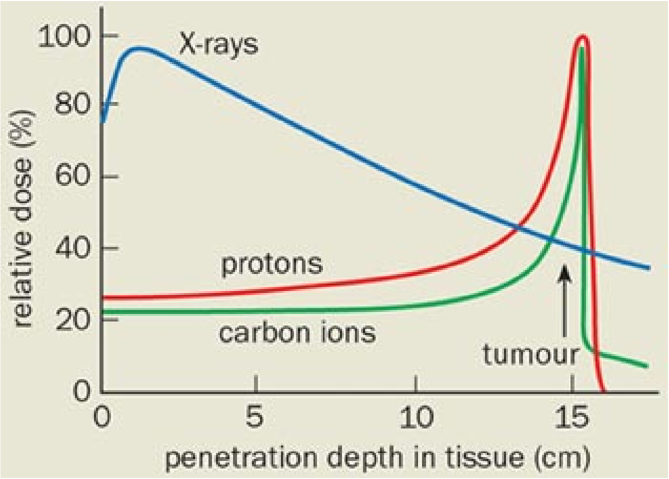


Figure 1.9: *Plot showing the different energy depositions along the paths of protons, X-rays and carbon ions, [58].*

Unlike X-rays, charged ions have a low dose deposition at the beginning of their track with minimal increase until just before they stop. At this point, there is a strong increase in deposited dose, peaking just before the particle comes to rest. This property of charged particles results in a much lower entry dose to patients, reducing skin and healthy tissue damage and maximising the dose deposited in the tumour.

Proton energy directly affects LET because energy is proportional to particle velocity. High energy (58 MeV), fast protons do not have enough time to interact with the matter they pass through, causing fewer ionisations resulting in a similar damage spectrum to that of X-rays and γ-irradiation [59]. The same particles at a lower energy (11 MeV) travel at a lower speed and are able to cause many more ionisations as they interact. The density of ionisations is especially important with regard to the biological effects of the radiation on cells. Studies show that clustered damage is most prevalent when treatments of high LET protons are used, as they result in high density ionisations [59]. Today we recognise that there is a direct relationship between the density of ionisations causing damage, in the form of strand breaks, and the resultant failure of the cell to repair the damage. This is an active area of research with the aim of identifying and exploiting the damage that cells find most difficult to repair following irradiation.

1.4 Measuring DNA Damage

1.4.1 Immunofluorescence

Immunofluorescence (IF) is a method which can be utilised to measure the number of DSBs present in cells. A biological marker, in the form of antibodies, is chemically marked with fluorescent dyes (known as fluorophores) which bind to DSB sites [60]. Direct IF is implemented when a single antibody is used and is conjugated with fluorophores. Indirect, also known as secondary IF, is when an additional antibody is used, to which fluorophores are attached. This second antibody binds to the first, fluorescently tagging a site of DSBs indirectly. There are various advantages and disadvantages of the two IF methods which need to be evaluated before determining which is best suited for the task at hand. For example, although utilising direct IF is less time intensive as it involves fewer steps, conjugated primary antibodies are usually more expensive than their unconjugated counterparts [61]. Figure 1.10 shows an image produced from indirect IF using two antibodies. Mouse antibodies, $\gamma H2AX$ (also known as Alexa Fluor 488), were used as the primary antibody which bound to DSB sites in the cell line used (HeLa cells in this case, discussed further in Section 1.5). A secondary antibody from rabbits, 53BP1 (also known as Alexa Fluor 594), is then applied to the cells and binds to $\gamma H2AX$. 53BP1 is a conjugated antibody, thus has been tagged with a fluorescent marker which can be observed during fluorescence imaging of the cells. It would also be possible to view these antibodies separately as $\gamma H2AX$ emits red light (emission peak at 520 nm)and 53BP1 emits green (emission peak at 617nm), and they each have excitation wavelength peaks of 499 nm and 590 nm, respectively.

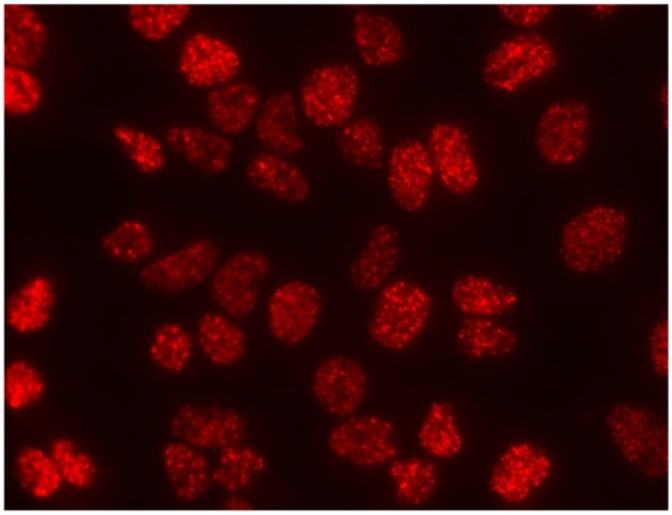


Figure 1.10: *An image taken after applying the immunofluorescence process to HeLa cells 1 hour after irradiation with 58 MeV protons.*

A single IF experiment is defined as at a specific dose damaging agent delivered at a specific time point. For repair kinetics studies, as in [62], the experimental procedure is as follows. For each experiment, two microscope slides containing prepared cells are analysed. A fluorescence microscope, at 40x magnification, is used to take five images (more if images are not clear) in the dark of each coverslip. Images are taken in areas of high cell population but with no overlapping cells. An automated program reads the image files and counts the number of foci (focus points of fluorescence) per cell. How the program counts the cells can vary between cell lines. This is due to different size/shape of the cells, brightness of the images captured etc. so comparisons between cell lines should not be made. The mean number of foci per cell is determined by averaging across images and is indicative of the DSB damage present.

Literature states that 20-40 DSB are expected per 1 Gy dose of X-rays [33]. However, this is not always seen experimentally and there is great variability. Factors such as radiation source and energy can effect the number of DSBs measured, and the cell line used affects results also. Unlike Comet Assay images, which are discussed later on, IF images show that cell repair mechanisms are still active four hours after irradiation.

Although IF studies are a direct method of visualising double strand breaks, IF is not a sensitive enough technique to resolve individual strand breaks and instead gives an overall impression of the damage present, This is useful in comparative studies but not detailed enough to give accurate, quantifiable measures. The Comet Assay is a more widely used technique that measures a number of other DNA damage metrics.

1.4.2 The Comet Assay

The Comet Assay, also known as Single Cell Gel Electrophoresis (SCGE) is a technique that can be utilised to measure DNA strand breaks within eukaryotic cells in vitro. Both DSBs and SSBs can be measured by adjusting the conditions in which the assay is carried out in and it has incredible sensitivity, allowing for individual cells to be observed and analysed. The main principle of the assay is to expose any cellular DNA strand breaks by applying a small electric field to stretch the DNA out along the field direction axis [63]. The Comet Assay can be broken down into 5 main parts as follows:

1. Coating slides with agarose gel
2. Embedding cells into the agarose gel
3. Lysis of cells to break down the cellular and nuclear membranes
4. Electrophoresis in which an electric field is applied across the cells
5. Staining and imaging of cells

Agarose is a gel-like substance that has an open cell matrix structure composed of pores. In the Comet Assay a low melting point molten agarose is utilised, into which a small number of cells are embedded as it cools. This allows DNA fragments, ranging from 100 base pairs (bp) to 25 kbp [64], to migrate through the pores at a later stage during electrophoresis. The pore size of agarose is inversely proportional to its concentration, and ranges from 100nm to over 500 nm in diameter for a 1% concentration gel [65]. It is important that agarose is prepared shortly before use in the comet assay and not left for extended periods of time. This is due to a process called syneresis, in which a separation of low-molecular-weight liquid occurs on the gel's surface [66].

Once cells are embedded in agarose, they are lysed to break down their cellular and nuclear membranes, by addition of a salt-detergent solution. This frees the DNA from the nucleus, allowing it to populate the hole in the agarose made by the cell. Lysis is also required to relax the super-coiled structure of DNA so that strand break damage can be realised through DNA migration during electrophoresis.

In order to reveal any strand breaks in damaged DNA, cells are exposed to an electric field in order to stretch out the DNA. This is conducted by adding an electrophoresis buffer and then placing the slides into an electrophoresis tank, which consists of an anode and a cathode. The buffer facilitates movement of charged ions when a potential is applied across the cells. During electrophoresis, the agarose gel acts as a molecular sieve, filtering DNA strand lengths. This results in shorter strands migrating further away from the original cell body whilst longer strands do not travel as far, due to their reduced mobility in agarose pores. The length of migration is an important metric in determining level of DNA damage, discussed in more detail in Section 2.3. For use in the Comet Assay, electric fields of around 1Vcm^{-1} are applied. Once DNA has been stretched out sufficiently, it is neutralised by rinsing with a suitable buffer and is then ready to be stained and imaged, producing an image such as that shown in Figure 2.2.

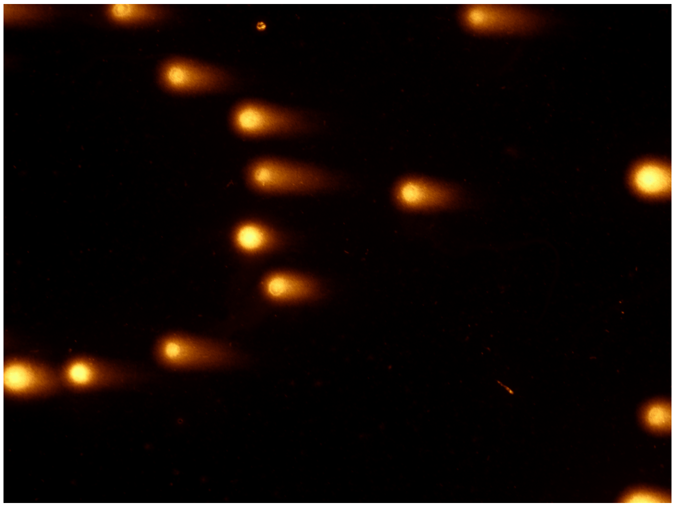


Figure 1.11: *An example image of HeLa cells following a neutral comet assay, irradiated with protons at a 4Gy dose.*

Staining is commonly carried out with SYBR-green, DAPI or $10\mu g/ml$ propidium iodide [67]. It is worth mentioning that staining power can vary both with the medium used to stain, but also through how the stain is applied. This is a source of error that is important to consider and is further discussed in Section 2.4. The final step is to image the stained cells in order to analyse them and quantify the apparent DNA damage. An overview of the Comet Assay process is depicted in Figure 1.12. For detecting only DSBs, cells undergo a neutral lysis and are incubated in neutral electrophoresis buffer [68]. An example Comet Assay protocol can be found in Appendix A.1.

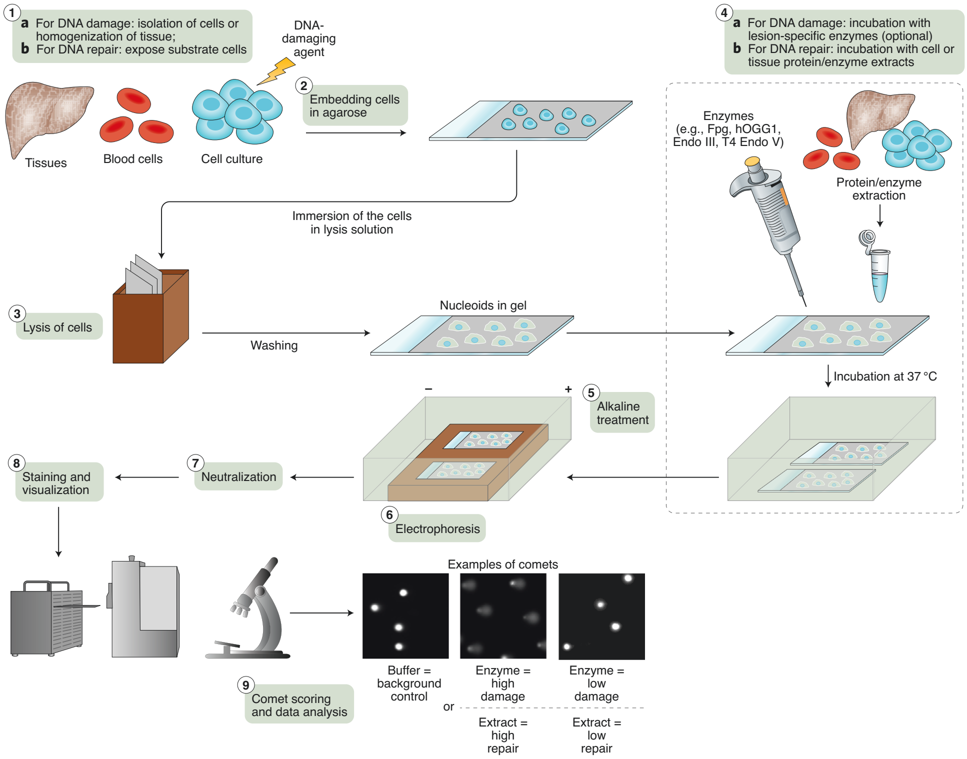

1
a For DNA damage: isolation of cells or homogenization of tissue;
b For DNA repair: expose substrate cells
DNA-damaging agent
Tissues
Blood cells
Cell culture
2
Embedding cells in agarose
4
a For DNA damage: incubation with lesion-specific enzymes (optional)
b For DNA repair: incubation with cell or tissue protein/enzyme extracts
Enzymes (e.g., Fpg, hOGG1, Endo III, T4 Endo V)
Protein/enzyme extraction
Immersion of the cells in lysis solution
3
Lysis of cells
Nucleoids in gel
Washing
Incubation at 37 °C
−
+
5
Alkaline treatment
8
Staining and visualization
7
Neutralization
6
Electrophoresis
Examples of comets
9
Comet scoring and data analysis
Buffer = background control
or
Enzyme = high damage
Enzyme = low damage
Extract = high repair
Extract = low repair

Figure 1.12: *An overview of the Comet Assay procedure, adapted from [69].*

The Comet Assay technique was first implemented in 1984 by Ostling and Johanson and was known, due to the small scale of cells and the use of an electric field, as a “microelectrophoretic study” [70]. In this study, they irradiated mammalian cells with ^{60}Co gamma rays at a dose rate of 1.8 Gy/min. A neutral detergent solution was used in lysis with the aim of measuring the DSBs present in the DNA. Since that time, the technique has been modified many times for specific uses. For example, Singh went on to study the damage of lymphocyte cells in 1988, using alkaline conditions after exposure to either X-rays at a dose of 200 rad/min (2 Gy/min) or H_2O_2 [71]. The use of an alkaline electrophoresis solution creates an environment in which the weak hydrogen bonds holding double stranded DNA together are broken. The individual strands can then be further processed through the stages of electrophoresis, staining and imaging as usual in order to observe both SSBs and DSBs.

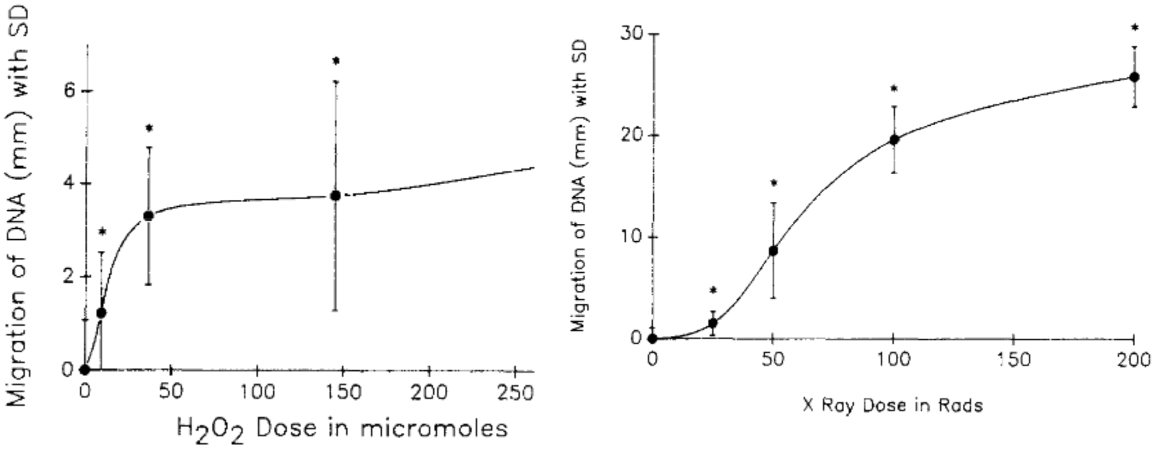


Figure 1.13: ***(a)*** *DNA migration length as a function of H_2O_2.* ***(b)*** *DNA migration length as a function of X-ray dose, [71]. The units of dose used (micormoles and Rads) are equivalent.*

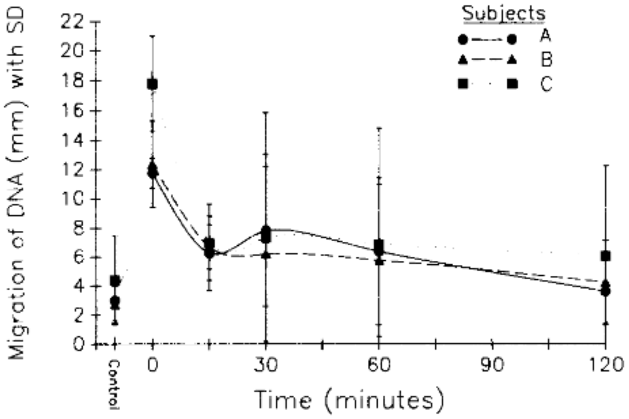


Figure 1.14: *Migration of DNA (towards the anode under electrophoresis) as a function of time, [71].*

Singh's study demonstrated the relationship between dose and relative damage to DNA, with Singh's findings of increased X-ray dose resulting in a larger migration distance of the DNA. This proved to be true for both types of agent used but the effect was significantly lower for the hydrogen peroxide solutions, Figure 1.13, giving evidence that the X-rays applied dose was more damaging to the DNA than H_2O_2 in the concentration used.
The repair kinetics of cells were monitored and it was found that the majority of repair occurred in the first 15 minutes following irradiation, Figure 1.14. More recently, studies have been conducted using the Comet Assay in order to measure the repair response resulting from different radiation sources, such as UV [72] and exposure of the DNA to other chemicals [73]. The Comet Assay presents some limitations, especially for making comparisons between independent data, as discussed in Section 2.4.

1.5 This Work

The Comet Assay technique is explored in this work, in particular the acquisition and analysis of assay images, which are fundamental in determining the efficacy of DNA damage following some DNA damaging treatment. The majority of data has been obtained following a neutral Comet Assay of either HeLa, UMSCC6, or UMSCC74A cells. The latter two cell lines are oropharyngeal squamous cell carcinomas; head and neck cancer cells. The HeLa cell line gets its name from the person it originates from; Henrietta Lacks.

A sample of Lacks' cancer cells were taken by Johns Hopkins researcher Dr. Geogre Gey in 1950, to develop surgical cell tissue for cervical cancer studies [74] [75]. Of all the samples collected from many patients, the HeLa cells were the only to survive and continue to grow. It was soon found that this cell line was immortal; it continues to proliferate indefinitely offering several advantages such as low cost, unlimited supply of materials and ease of use [76]. The HeLa cell line has been involved in a plethora of vital work such as studying the effects of zero gravity on cells in outer space, developing the polio and COVID-19 vaccines [74] and cancer treatment studies, as seen in this work.

Cells have been damaged by either X-ray or proton beam irradiation. For the case of X-ray irradiation, cells were exposed to low-LET X rays (100 kV) using a CellRad irradiator (Faxitron Bioptics, Tucson, AZ; dose rate of 3 Gy/min) [59]. Proton irradiation took place at Clatterbridge Cancer Centre [77], using a horizontal, passive-scattered beam line of 60 MeV maximal energy from the Douglas Cyclotron at Clatterbridge [78]. Cell line dose calibration curves were utilised in order to deliver equivalent doses to differing cell lines, as cell sensitivity is unique to the cell line used. Analysis of Comet Assay images is performed in Jupyter Notebooks using Python 3.6.

The goals of this work are to explore how variations in experimental protocol and image acquisition processes affect comet measurements, alongside developing a coherent image analysis pipeline that improves and optimises the image analysis process. The aims of the analysis pipeline are to:

- Minimise human interaction as much as possible
- Maximise the number of comets that are measured in each image
- Maximise the removal of false positive measurements
- Allow the user to review measurements, and their source in a meaningful and easy to understand manner

Chapter 2 introduces image data as a whole, and examines the process of capturing Comet Assay data, detailing how images are created. It is important to understand how image file formats work, as varying formats handle and store data differently. This can have a considerable impact on the image analysis process, and is an important factor when considering techniques to implement and their outcomes. Chapter 2 also explores the high variability in the experimental, imaging and analysis processes, outlining how slight adjustments can have a large effect on the images obtained.

In Chapter 3 some image segmentation methods are introduced and their advantages and disadvantages are explored, before applying the methods to Comet Assay data in Chapter 4. Results are compared in terms of their effectiveness in separating objects of interest from background.

The actual measurement of comets is then explored in Chapter 5. Widely used measures are analysed and a new measurement metric is proposed, which aims to reduce the variation in how damage levels are calculated and simplify the process.

Chapter 6 presents a Monte Carlo model that has been developed to produce simulated Comet Assay images to aid in analysis.

Recommendations are offered on how analysis can be standardised across the field to allow reproducibility and direct comparison between research labs. This would also help flag any missed assumptions whilst allowing a more complete picture of data acquisition to be presented.

Chapter 2

Comet Assay Data

2.1 An Introduction to Image Data

Hundreds upon thousands upon millions of individual data points in the form of pixels coalesce to produce the vast majority of images we see everyday. Individually they offer very little information, but when combined their power is vast. Humans view a single image as one entity. We may see different regions, objects and colours within an image but generally we process an image as the sum of all its parts, rather than distinguishing each individual pixel (especially in the case of high resolution images containing a huge number of pixels). Computers are much better equipped to break down an image into smaller components or even individual pixels, and analyse them in ways the human brain is incapable of. Conversely, things that are obvious to the human brain, such as edges and specific objects, can easily be overlooked by a computer. Digital image files are created to organize, store and display image data on electronic devices such as computers and smart phones. This data is the outcome of charge deposition from exposure to electromagnetic radiation, such as visible light, X-rays, infrared radiation etc. Image files come in many different formats, each with their advantages and disadvantages, but they intrinsically have common features.

Pixels are comprised of channels and bits. The bit depth determines the number of values an individual pixel can take on, producing its colour or greyscale range. A 1 bit image is purely black and white (each pixel can either be of value 1 or 0), otherwise known as a binary image. An 8 bit image has pixels that can take on $2^8 = 256$ possible values. Similarly, a 16 bit image indicates each pixel can take on values in the range 0-65563.

Channels can be viewed as layers that combine to produce colour and opacity in images but can also provide information from different types of light. An image formed of only 1 channel can be either binary (black/white and therefore only 1 bit) or greyscale. Colour images are comprised of red, green and blue channels, and are known as RGB images. RGBA images are also possible comprising of the three RGB channels mentioned, plus an additional A (alpha) channel. This determines the pixel's opacity, used in layered images etc.

2.1.1 Image Compression

A higher number of pixels provides a better image resolution, allowing more detail to be captured and observed, whilst additional channels increase the range and type of data collated. This is preferred of course but comes at the cost of producing substantial amounts of data, which needs to be stored, transferred and read. In order to reduce the amount of data required to represent the information in an image, compression techniques have been developed, which aim to reduce the total amount of space or bandwidth required to store or transfer information [79]. Data compression is a process in which the same amount of information is conveyed using less data and comes in lossless or lossy forms; where the original information is able to be completely recovered or suffers some information loss respectively. Image compression utilises a concept called relative data redundancy. A vast number of pixels, or bits in an image are often irrelevant in conveying useful information, making them redundant to the overall purpose of the image. The relative data redundancy, R, of an image is a calculation which determines how much of it is useless in conveying information, [80].

$$R = 1 - \frac{1}{C} \tag{2.1}$$

C in Eq. 2.1 is the compression ratio which is defined as the proportion of bits, b' required to represent the same information that a larger number of original bits, b provide.

$$C = \frac{b}{b'} \tag{2.2}$$

There are three types of redundancy that can be exploited for image compression; coding redundancy, psychovisual redundancy and interpixel (spatial) redundancy [81]. Coding redundancy utilises a system of symbols (letter, numbers, bits etc.) known as a code to represent each piece of information in an image by assigning it a code symbol sequence, called a code word. The length of these codewords is optimised accoridng to the intensity range present, producing the smallest codewords possible in representing the image's information.

If we take the adapted example from [82] of a M x N image that has pixels of intensity value r_k, where $r_k = 87$, 128, 186 or 255 exclusively, we find that the probability of r_k occurring is $p_r(r_k)$ using Eq. 2.3.

$$p_r(r_k) = \frac{n_k}{MN} \qquad k = 0, 1, 2, 3, ..., L-1 \tag{2.3}$$

L is the number of intensity values and the frequency of a given intensity k is defined by n_k. When defining the number of bits representing each value r_k as $l(r_k)$, the average number of bits necessary to represent each pixel is found by summing for all pixel intensity values k, the product of the bit number required to represent that intensity and the probability of that intensity's occurrence, Eq. 2.4.

$$L_{avg} = \sum_{k=0}^{L-1} l(r_k) p_r(r_k). \tag{2.4}$$

From this average number of bits, the total number of bits required to represent an $M \times N$ image can be found from MNL_{avg}. Table 2.1 shows an overview of this example, providing pixel intensities, their probability of occurrence and 2 sample codes used to represent the different pixel intensities.

r_k	$p_r(r_k)$	Code 1	$l_1(r_k)$	Code 2	$l_2(r_k)$
$r_{87} = 87$	0.25	01010111	8	01	2
$r_{128} = 128$	0.47	10000000	8	1	1
$r_{186} = 186$	0.25	11000100	8	000	3
$r_{255} = 255$	0.03	11111111	8	001	3
r_k for k $\neq$ 87, 128, 186, 255	0	-	8	-	0

Table 2.1: *An example of variable-length coding for an image containing 4 different intensities. Adapted from [82].*

For the case of Code 1 where $L_{avg} = 8$, all intensities have the same code length,

$l_1(r_k) = 8$. When using a simplified coding, as done for Code 2, we obtain a much lower bits per pixel average of 1.81, calculated as follows

$$L_{avg} = (0.25 \times 2) + (0.47 \times 1) + (0.25 \times 3) + (0.03 \times 3) = 1.81. \tag{2.5}$$

Using the original bit average of 8 and the smaller calculated bit average of 1.81 we can calculate the compression ratio from Eq. 2.2 as

$$C = \frac{8}{1.81} \sim 4.42 \tag{2.6}$$

and using this in Eq. 2.1 we obtain a relative data redundancy, R of

$$R = 1 - \frac{1}{4.42} = 0.774. \tag{2.7}$$

This tells us that 77.4% of the data in the original 8-bit image are redundant, and it is possible to utilise only the relevant data by using fewer bits per pixel to represent the same information.

Psychovisual redundancy exists due to human perception. The brain is a very complex system that has several factors affecting image perception, including brightness uniformity and spatial/contrast resolution [83]. This results in unequal sensitivity to incoming visual information from which we infer a varying importance to different parts of an image. For example, very small differences in colour are not registered by the brain, but it has a much higher sensitivity for dark intensity changes than for light ones [84]. The consequence is that if the brain cannot perceive pixels that are slightly different to those surrounding it, the data is essentially irrelevant. Some compression techniques take advantage of this psychovisual redundancy by removing small fluctuations, simplifying how the data can be stored using much less data, resulting in a lower quality image and not fully representative of the original.

Inter pixel redundancy is overcome by exploiting pixel correlation in an image, as many images have large regions in which pixels have the same or similar intensity value. In this way, the individual value of a pixel can be predicted from its neighbours for some cases of low intensity variation. An image's intensity array is transformed to some non-visual representation called a mapping, which can be reversible if the original image is able to be reconstructed from it exactly, or irreversible if not. Such mappings include an array of differences between adjacent pixels or run length encoding (RLE), expressed using run length pairs. The latter is a record of consecutive pixel value occurrence in a certain direction (rows, columns or even diagonals across an image can be used), known as the run length. For example, it could be encoded that all 1020 pixels in the top row have value 0, rather than recording all 1020 pixels. A run length pair details the start of a new pixel value and the number of consecutive pixels with this value. The result of this new run-length delineation is compressing an example (256 × 256 × 8) image by a factor of (256 + 256 × 8), stemming from each 256 row of the original image becoming a single 8 bit intensity value with length 256 in the RLE, producing a 128:1 compression.

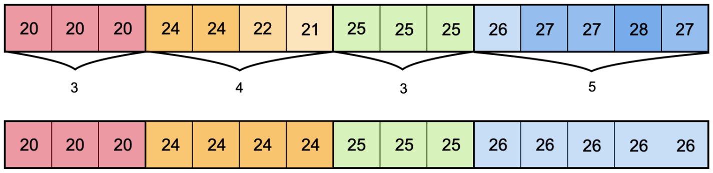


Figure 2.1: *An example of applying lossy RLE to a set of pixels (top) with a minimum run length of 3. The output is the pixels on the bottom that have been grouped according to the number labels shown.*

There are occurrences in which RLE can produce negative compression, where the data used to describe the image is actually larger than the original image data. This arises when an image is very dynamic in terms of intensities and contains few sections of high intensity repetition. These types of images can still be compressed in a positive manner, thus reducing the data used to represent it, but the resulting compression is now of the lossy type. This means the original image cannot be reconstructed from the compressed data, only some representation close to it. To employ lossy RLE, some threshold for run length is set. For an example run length threshold of 3, any consecutive intensity occurrences with frequency less than 3 will be grouped with the preceding or proceeding run and take on their intensity value, as shown in Figure 2.1. Here the pixels with run length < 3 are grouped with the preceding run pixel intensity and take on this value themselves. A method of optimisation is usually applied in choosing the value short run length pixels taken on to reduce the variation from the original image. The minimum run length value is also optimised dependant on the input data composition. Of course the consequence of using lossy RLE is a loss of small areas of intensity variation that cannot be retrieved during image reconstruction. Sometimes this loss of data is negligible and has little effect when displaying the image, but for cases where this loss of data would impact the image's use, for example in the analysis of comet assay data, it is important to choose a suitable image format.

2.1.2 Image File Types

Many different image file types exist, each with its own methods for encoding and compressing images in either a lossy or lossless manner. Some commonly used examples include Joint Photographic Experts Group (JPEG), Portable Network Graphics (PNG), Tagged Image File Format (TIFF) and Bitmap Image Files (BMP). These are all raster images, also known as bit-mapped images that utilise pixels in a row × column matrix formation to store image data. Vector images also exist which make use of mathematical vector elements rather than pixels to define an image. This makes them completely lossless and infinitely scalable by definition as the resolution is unaffected, which is not always the case for compressed raster images. Examples of vector image file formats include Scalable Vector Graphics (SVG), Encapsulated PostScript (EPS) and Portable Document Format (PDF), commonly used in graphic design. Raster image files will only be explored further here, as they are the most common format for storing photographic images produced by a camera.

The JPEG format is used for images comprised of 24 bits (3 colour channels of 8 bits each). It was created in 1992 by the Joint Photographic Experts Group, which developed a new generation of image compression specifically for photographic images [85]. The compression technique utilised in JPEG images is the Discrete Cosine Transformation (DCT), developed by Nasir Ahmed in 1974 [86] [87]. It is a lossy technique; data is sacrificed in order to reduce image size. The DCT works on the basis of applying a transform matrix to 8 × 8 blocks of pixels, starting in the upper right hand corner and working across and down the image. This produces a matrix of coefficients for each block which can be compressed during quantization, achieved by multiplying the coefficient matrix by a quantization matrix. there are 100 predetermined quantization matrices, offering a varitey of quality levels. A higher value quantization matrix produces a higher quality image in which less data is lost.

The PNG format, created in 1995 [85], is an image file that utilises lossless image compression. It supports up to 32 bit images as it allows for a fourth alpha channel, determining a pixel's transparency. The result of such a high bit depth is that a 32 bit image can be displayed in over 16 million different colour shades, with the added benefit of changing the transparency of each pixel, due to an additional 8 bit alpha channel. The data stored in PNG images is compressed using deflate compression.

The native bitmap image format for the Microsoft's Windows operating system is the BMP. Bit depths of 1, 4, 8, 16, 24 and 32 per pixel are supported, providing a large colour/intensity spectrum, although files utilising 16 and 32 bits are very rare [88]. RLE can be utilised in BMP files, but is only of use in images with large blocks of identical colour. This often means that they go uncompressed, retaining all data and information that was originally captured in the image, but at the cost of producing large file sizes.

The Tagged Image File Format (TIFF) was created in 1986 by Aldus and is capable of storing greyscale, palette colour and full colour image data in a range of colour spaces [89]. Its name stems from its ability to store "tags" of supplementary information about the image, in the form of metadata. It can be compressed in a lossy or lossless form, but is almost exclusively used as a lossless image format with virtually no compression [90]. This implies that TIFF files are often very large and are therefore not used in everyday applications, such as on web browsers. TIFF files are usually saved with 24 or 48 bits (8 or 16 bits per colour channel), allowing for a very dynamic range of colour intensities to be stored.

Image files can be converted to other formats, to make use of other compression methods that produce smaller files, making it easier to store and transfer lots of data. However, this transform is often at the cost of information loss, as the bit depth, resolution and resulting intensity range are usually decreased. It is therefore important to consider whether the conversion of image data is useful, considering the information that may be lost in the process.

2.2 Producing Comet Assay Images

Images are captured following a Comet Assay to document the elongation of DNA in the presence of strand break damage. This is possible by using a flourescent marker that binds to DNA and produces light, captured by microscopy. Comet Assay images are generally produced with a minimum of 8-bits in order to capture sufficient intensity range for analysis. A camera attached to a microscope captures the light and is represented as a digital image through some imaging program. The microscope is necessary to magnify the image to an observable size. The most appropriate image bit value is determined by the dynamic range of light produced, dependent on the choice of fluorescent marker and the configuration of other camera settings such as exposure time. An example imaging program used to capture the data used in this work is MicroManager [91]. Manual operation of the imaging system is required and this creates a bottle neck in the image acquisition process, limiting the number of images that can be produced in a given time frame. This slow process can mean that a minimum benchmark number of images are produced per slide across an experiment, which may not be representative of the cells across a whole slide. It will remain necessary, at least for the foreseeable future, that the imaging process requires human operation to ensure a good standard of data is produced. This step will therefore continue to be time consuming. However, there are other areas proceeding image acquisition that can be shortened significantly, such as image analysis, discussed in Section 2.4.3 .

2.3 Scoring Comet Assay Data

The imaging of Comet Assay slides permits measurement of the apparent DNA damage, which is widely known in the field as scoring. The methods of measurement can vary considerably, impacting the results and interpretation which is discussed further in Section 2.4.3. Common metrics for measurement are tail DNA, tail length and DNA distribution profile also known as tail moment [92]. These require identifying regions of the comet known as the “head” and “tail”, depicted in Figure 2.2. The comet head specifies the region where DNA was located prior to electrophoresis, and where the majority of DNA remains except for cases of high damage. The tail is composed of DNA that has migrated through the agarose gel towards the anode during electrophoresis. Tail length is simply a pixel measure of the tail along the electric field axis, converted into microns [92]. A longer tail is indicative of smaller fragment formation (formed from a larger number of breaks) as only small fragments are able to migrate through many agarose pores to traverse longer distances away from the comet head. The tail length measure is disregarded in publications by many researchers as the maximal tail length can be achieved at relatively low doses, beyond which further damage is indicated by an increased fluorescence in the tail [93].

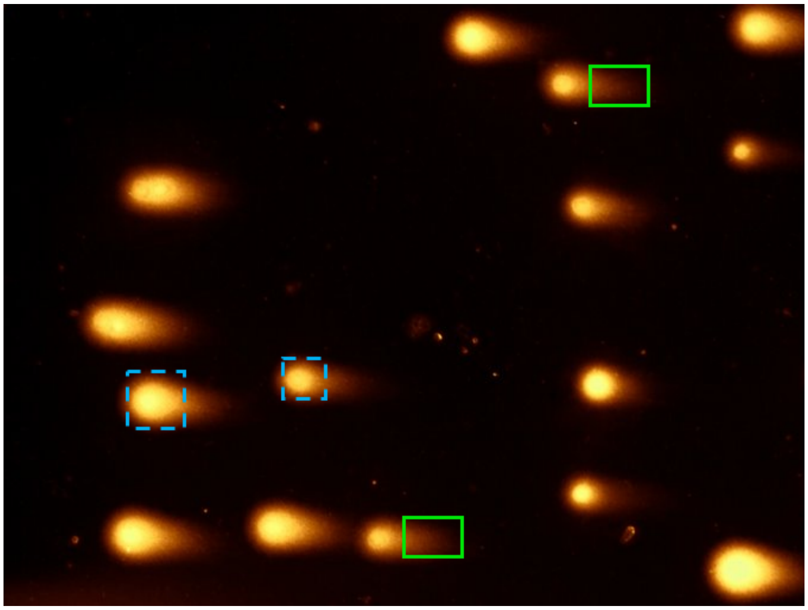


Figure 2.2: *An example image of irradiated cells following a neutral comet assay. The comet tail is highlighted in green (solid rectangle) and the head in cyan (dashed rectangle).*

Tail DNA is a measure of the fluorescence in the tail as a proportion of the whole comet's fluorescence and is often expressed as a percentage, Eq. 2.8, where n represents an individual pixel belonging to the tail, which contains a total of N pixels, and m represents those that belong to the head, comprised of M pixels.

$$TailDNA(\%) = \left(\frac{\sum_0^{N-1} Intensity_n}{\sum_0^{N-1} Intensity_n + \sum_0^{M-1} Intensity_m} \right) \times 100 \tag{2.8}$$

The tail distribution (or moment) measure incorporates the product of both tail DNA and tail length. This is often important as individually these measures can be misleading. For example, a longer tail length can be formed of fewer DNA fragments, resulting in a lower intensity tail. From the tail length measure alone one could assume it depicted a higher level of damage than a shorter tail, which may in fact be composed of higher intensity pixels due to more strand breaks. It is therefore essential that a combination of these measurements is made and reported in order to provide a more complete picture of the DNA damage.

2.4 High Variability in Comet Assay Data

There are a number of factors that contribute to the high variability found in Comet Assay data. These come from the processing of cells and slides prior to the Comet Assay, the Comet Assay procedure itself, image capture and image analysis processes. General experimental protocols and procedures exist to mitigate these factors but they are just guidelines and are often not described and examined in published work. The necessity to calibrate and report parameters is vital in producing comparable and reproducible results. This type of standardisation allows for inter- and intra- laboratory experiments to be compared, pushing the field further forward as new damaging agents and repair inhibitors are found. An example can be found in the human biomonitoring tool (hCOMET), in which a collaboration of researchers are combining Comet Assay results to create a database recording the effects of various factors on DNA damage [94].

Cells have a radiosensitivity limit which defines the level of damage they can withstand before entering apoptosis; programmed cell death [95]. It then follows that highly damaged cells within a single image can present somewhere on a spectrum between comets with long, bright tails and disappearing completely as they have undergone apoptosis. In the case where a considerable number of cells have died, the damage measures would not only be skewed due to the lower number of cells available, but also biased towards the remaining cells that have a lower radiosensitivity and have survived the high dose. Therefore experiments where a high dose is delivered to cells needs to carefully consider these factors and report results accordingly. One possible recommendation is to present cell survival data in the case of very high damage experiments, as suggested by Bankoglu *et al.* [96].

2.4.1 Experimental processing of slides

There are many experimental steps that form a Comet Assay protocol. Due to its variable use in different applications where DNA damage is studied, there is a lot of variability in the chemical quantity/forms used and time elapsed during/between steps, impacting the equivalence of results produced [97]. Firstly, slides are prepared, with the goal of obtaining a uniform agarose gel layer that is stable enough to survive the whole assay procedure. Different methods for this exist, utilising one, two or three agarose layers [98]. This will affect the imaging procedure as the transmission of light can vary over a number of layers, if they have differing refractive indexes, and should be a consideration during analysis. The multi-layer options also allow more potential contamination to be introduced into the gel. This should be minimised at all stages as impurities can interfere with image analysis.

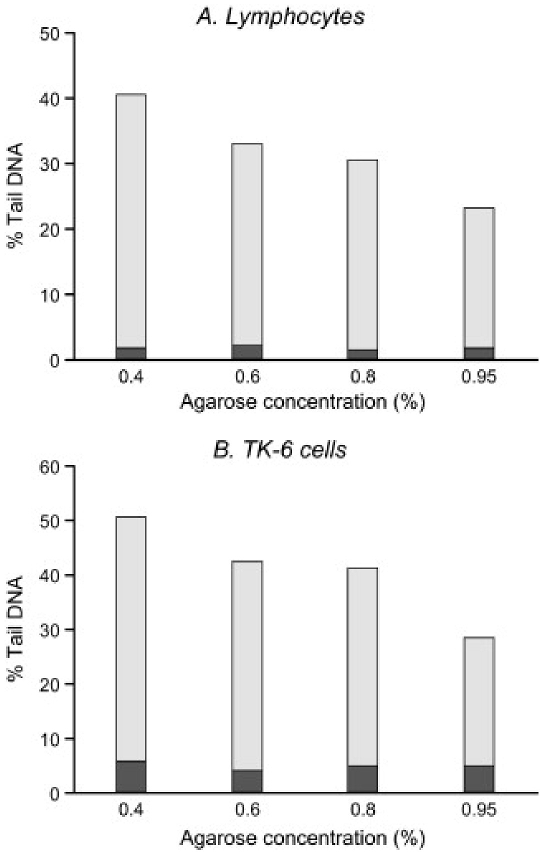


Figure 2.3: *Graphs showing how agarose concentration has a direct relationship on the migration of DNA in treated lymphocyte and TK-6 cell lines, [99]. Irradiated data is shown in light grey whereas control data is in dark grey.*

The concentration of agarose used is also an important factor that affects DNA migration and in turn damage analysis. As previously mentioned, agarose concentration affects the gel's pore size. High concentration results in small pore sizes which restrict and impair the migration of longer DNA fragments. Typically concentrations of ∼0.5-1% are reported, but the actual final concentration can remain unknown due to the indeterminate volumes cells are suspended in prior to introduction into the agarose [99]. The affect agarose concentration has on DNA migration is depicted in Figure 2.3, showing that variability from 0.4 to 0.95% agarose concentration can reduce the tail DNA measured by more than 10%

in lymphocytes and $\sim 20\%$ in TK-6 cells.

Another important factor is the alkaline treatment time during lysis, where the cellular and nuclear membranes are broken down, freeing the nucleoids of DNA. The time cells are left in this treatment affects the amount of unwinding that occurs, impacting DNA migration [100]. A time of 20-40 minutes is usually reported [101], and it is deemed that 20 minutes is sufficient for this process [98]. However, an increase in migration is observed for a higher treatment time. This is understood to be due to radiation induced alkali labile sites [102], areas in DNA that become highly sensitive to alkaline solutions and contribute to SSB production, as the increase was not observed in control cells [103]. The conversion rate of alkali labile sites into SSBs can be affected by temperature changes [104]. Some studies have also shown that an overnight lysis time of the alkaline comet assay significantly increases the sensitivity for detection of DNA damage. This is attributed to an increase in efficiency of damage detection, caused by a longer time frame for DNA winding and diffusion prior to electrophoresis [105].

Other lysis factors such as temperature also affect DNA repair pathways and in turn, the DNA migration observed [106]. It has been shown that DNA repair can be inhibited by a small heat increase (a temperature of 45 °C compared to 42°C) for DSBs [107]. It is therefore important to first choose an appropriate temperature when planning experiments and then control for it during lysis, but perhaps is more crucial to report these exact conditions to the scientific community.

As for lysis, electrophoresis also allows for a high degree of variability in regard to the length of time it is conducted for, the electric field strength applied and the amount of buffer used. Electrophoresis is usually carried out over a time period of 20-30 minutes [99]. This is due to the expectation that an increase of electrophoresis time would not have a proportional affect on DNA migration beyond this point, stemming from a belief that tail DNA is formed of DNA loops still attached to the nuclear matrix [108]. However it has been found that an increase of time from 20 to 30 minutes can produce a significant extension in tail DNA measured [103], [99], as shown in Figure 2.4.

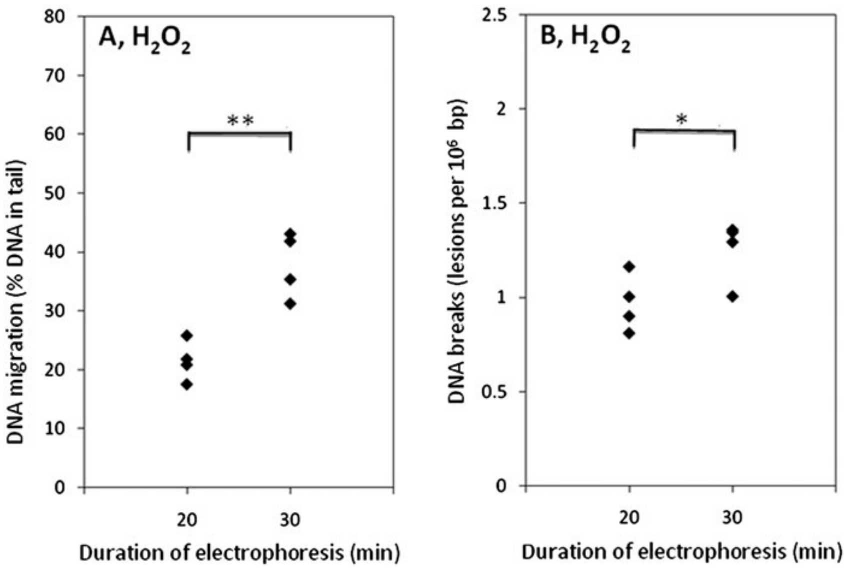


Figure 2.4: *Graphs showing the effect of conducting electrophoresis for 20 and 30 minutes on cells treated with H_2O_2 . The left plot shows the tail DNA measured for each set of cells and the right shows the corresponding number of DNA breaks the tail DNA represents, transformed using calibration curves. Adapted from [103].*

The electric field strength contributes to the potential DNA migration that can occur. As previously discussed in Section 1.2, electrophoresis causes migration of DNA due to its negatively charged backbone structure, which is attracted to the anode within an electrophoresis tank. The mobility of DNA is partially attributed to its charge, and in part to the movement of positively charged ions in the electrophoresis buffer, known as electroosmotic flow (EOF) [109]. Both DNA mobility and EOF are dependent on the agarose pore size, dictated by its concentration, which is another crucial reason it is accurately known and reported.

The electric potential applied across the cells, measured in Vcm^{-1}, is another variable that affects these measures, instructing the strength of attraction the charged ions have to the anode and cathode of the electrophoresis tank. Typical voltages for an electrophoresis chamber are roughly determined by multiplying the distance between electrodes (in cm) by a factor of 0.6 [110]. This of course introduces huge error and is not a robust measure to report or use in calculations. It also overlooks voltage drops at each electrode and can have serious effect on the potential cells experience in small tanks. As the electric potential is dependent on a solution's concentration gradient, the electrophoretic buffer used can be circulated using an external pump to keep the electric potential stable in concentration across cells, but also to ensure temperature stability [111]. It is also crucial cells have an even covering of electrophoresis buffer to ensure a consistent EOF contribution to DNA migration.

Staining of slides to prepare them for imaging is the final experimental aspect of the Comet Assay that introduces variability in DNA migration measurements. A first consideration is the staining agent used, as different chemicals have differing binding ability. Commonly used stains include ethidium bromide, propidium iodide and SYBR Green, which are chosen depending on an individual investigator's needs [112]. It is important that little time is elapsed between staining and imaging, as some stains are known to fade. For example, SYBR Green is often the DNA marker of choice due to its bright fluorescence compared to other stains, but it known to fade during scoring [92]. It has also been observed to have varying binding affinities between single- and double- stranded DNA by a factor of ∼11 [113]. It is therefore useful for the scientific community to be informed of not only the stain used, but the time taken to image, post-stain. Some fluorescent dyes allow for anti-fade to be used, offering a great reduction in the rate of signal quenching [114].

It is recommended that as much detail is supplied in publications where the Comet Assay has been utilised to measure DNA damage in the form of strand breaks, as stated in [69]. This will standardize the parameters that are reported and allow for a more comprehensive understanding of the research conducted, while affording inter-experiment and inter-laboratory reproducibility.

2.4.2 Imaging Procedure

The correct imaging of cells requires various parameters and settings to be defined on the microscope and the computer-controlled camera. A first consideration is the camera settings. Saturation, contrast and the grouping of pixels (binning) are all features that can be controlled for image acquisition, and directly affect the image produced. In fact, variations of tail DNA as high as 40% can be attributed to changes in exposure and microscope focus [115]. Image exposure is a measure of the incident light multiplied by the exposure time, measured in lux seconds. Optimum exposure is dependent on the sensitivity of the

camera's sensor and the subject's luminance [116]. Under/overexposure of an image occurs when the light captured is not represented well, and does not make full use of or exceeds the available dynamic range of intensity. Histograms are routinely used in imaging, and are often built into digital cameras. They offer a visual representation of an image's intensity spectrum, allowing the user to see the range captured for a set exposure. These histograms can sometimes be used to identify poor exposure by a clipping of intensities recorded at the bottom and/or top of the intensity range respectively, but this is entirely dependent on the subject material and its luminescence.
Similar to exposure, a camera's contrast setting also affects the captured images. Contrast is defined by the range of intensities that an image covers. A low contrast image has a narrow range of intensities and a high contrast image has a broad range. For comet assay imaging, the intensity measured is vital in describing the apparent DNA damage, therefore it is crucial that camera settings are optimised to best represent the fluorescence.

A further feature that affects image acquisition and in turn the analysis of Comet Assay images is camera binning. This is a process in which adjacent pixel intensities (for example a 4×4 region) are combined to increase the overall signal in individual pixels. This boosts the signal to noise ratio in an image, better separating objects in an image from the background. However, binning a large number of pixels has the disadvantage of reducing the image's resolution by outputting a lower number of pixels. It is therefore important to find a suitable compromise between binning value and resolution, so as to not lose too much information.

Another consideration is the microscope focus, which is required to clearly see cells. This can be controlled manually and the operator is the sole judge of when the observed image is in focus. As imaging progresses and the microscope is moved across the slide, it often becomes out of focus, requiring focusing once again. As this is individually judged, there is no way of knowing if the new focus matches the original. Even a slight difference of focus can impact the image acquired, which can affect analysis outcomes. Many digital microscopes have a built in auto-focus feature, designed to automatically adjust the focus when the microscope is used. Although this provides a much more consistent method for ensuring focus, it is important to calibrate this to ensure images are acquired in the sharpest possible focus, reducing subsequent analysis errors.

2.4.3 Image Analysis

As previously mentioned, the measurements obtained from Comet Assay analysis are linked to the manner in which comets are scored. It therefore follows that the same images analysed by different groups using different methods can lead to very different conclusions. This can be misleading at best or dangerous in the worst cases, where research feeds into clinical decisions. In order to perform the measurements of tail DNA, tail length etc. a methodology is required to determine which pixels belong to the comet head, tail and which are background pixels.
One step back from this is the most basic damage assessment method, which discretely categorises apparent damage. A user manually observes Comet Assay images and assigns them to a bracket of damage i.e. a number between 0-4 [117]. This method is very quick and gives a rough answer regarding the level of damage exhibited by the electrophoresed cells. It does however require an expert with a high level of experience to perform, and is a very objective and potentially misleading measure with a unquantifiable error.

A simple way of dealing with the human variation manual scoring encompasses is to hand the work over to a computer. Many software options exist which are designed to read in Comet Assay images and perform analysis of them [118]. They come in semi and fully automated varieties with varying requirement of user input. An example of a widely utilised semi-automated software package is Komet [119]. Using software allows for the same algorithms to be applied to every image, creating consistency and reproducibility in the analysis. They also do not require the user to be highly experienced in handling Comet Assay data. There is still an issue present in semi-automated software due to the requirement of the user to select images and define comet regions during the image analysis. A solution is present in the use of automated image analysis systems that only require a user to select the files to be analysed and the rest is dealt with by the program, which presents the results in an ordered manner, such as a spreadsheet [120]. Automated software is a current popular method of performing Comet Assay image analysis but it often still requires parameters and environments to be explicitly stated for reproducibility and comparison between research. It is also not often clear exactly how measurements on comets are performed, and exactly how comets are identified from the image background.

Using computer software does however have its own limitations. It can often be expensive to purchase and use, excluding smaller research groups or individuals. The time taken to analyse large datasets is also a factor, acting as an additional bottleneck following the imaging procedure. A further point to note is the difference in comet formation for the alkaline and neutral assay procedures. Neutral comets tend to have a more homogenous staining, whilst those from an alkaline treatment appear more granular [117].

The use of deep neural networks for object detection is a fast growing field of research, discussed further in Section 3.5, and is utilised for Comet Assay data analysis [121, 122, 123]. Rosati *et al.* have explored automatic comet identification and classification using the Faster R-CNN network [124]. This allows a complete automation of DNA damage analysis, with a high performance compared to other methods, but not in a quantifiable way. Instead, comets are classified as representing high, medium and low DNA damage following testing on images from two datasets. The limited range of data sources mean it is unclear how this model would perform on other data.

Chapter 3

Image Segmentation Methods

Measuring the parameters of imaged cells allows us to determine the level of damage they have undergone in terms of strand breaks. In order to perform measurements such as tail DNA and tail length, the comet instances need to be identified in the microscope image. This process is termed segmentation. There are many segmentation methods that can be used to identify objects in an image and some examples are analysed and compared in Chapter 4.

A simple way to separate the comet data from the background is to apply a global threshold. This creates a binary image in which all pixels above the threshold value are set to 1 and all other pixels are set to 0. Automatic thresholding algorithms have been developed to remove the manual selection of a threshold value. Such methods include the Otsu [125], Li [126, 127], triangle [128] and adaptive thresholds, which are all described here and their use evaluated. There are also more sophisticated ways, in which variations within an image are considered and thresholds are applied locally. These methods consider the relationship between pixels in a much smaller region and a threshold value is calculated for each pixel individually.

3.1 Global Thresholding

Global thresholding is a simple method that performs a cut at an assigned value of intensity, T. All image pixels are then tested against this cut; those below that value are assigned as 0 and those above to 1, Eq. 3.1.

$$g(x,y) = \begin{cases} 1 & \text{if } f(x,y) \geq T \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

This produces a binary image, where the aim is to have an image where all the relevant and interesting components are represented (as 1's) and the remaining image is assigned to background and discarded. There are two main issues with this method. Firstly, how is the value of T chosen? As global thresholding only has one parameter to tune, its value directly impacts the output. A range of threshold values are often explored to find the optimum one for the desired outcome. But this is not a robust method and further problems can arise when applying the global threshold to new images. Therefore, optimisation for each image is required to find the best T value. The second issue is the consequences of choosing a specific value of T.

Consider three options; a lower value of T, a perfect, middle value of T and a higher value of T. The lower value will likely include the majority if not all of the pixels belonging to the image objects we wish to segment. However, for objects in close proximity to one another, we run the risk of identifying them as one whole object rather than two or more individual ones. For the case of a higher T value, we are more likely to avoid the issue of depicting two objects as one. But this is often at the expense of losing outer pixels and not fully defining the object. The middle ground is a value that balances these two conflicting issues and defines a T value somewhere in the middle; our Goldilocks threshold parameter. The issue we run into here is that the threshold has been optimised specifically for that image (or set of images) and may then be inadequate for any subsequent image set it is used upon, thus requiring a tedious and cumbersome optimisation for every image to be thresholded.

An alternative to the manual selection of a global threshold value is to use algorithms that compute T automatically. This removes the need for complex and manual parameterisation for an image set, and a new threshold is calculated based on the current image, for each new image. Many of these methods work by manipulating an image's histogram, an example of which is depicted in Figure 3.1. Image histograms, known simply as histograms from here on, display the pixel intensity spectrum of an image, with intensity on the horizontal axis and the frequency on the vertical. They allow a visualisation of the tonal range and are useful in observing the contrast and potential under/overexposure of an image.

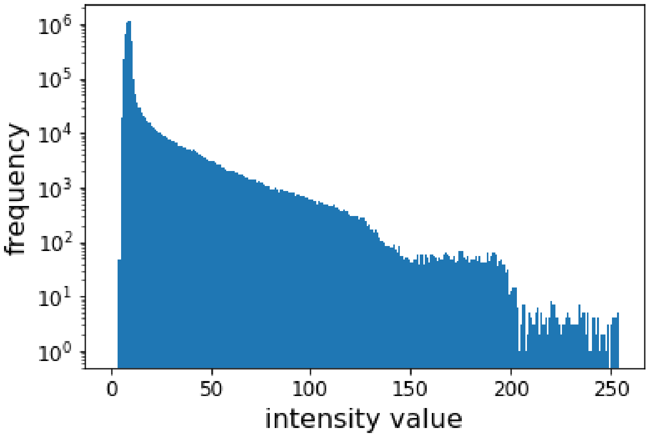


Figure 3.1: *A histogram of an image with a minimum intensity of 4 and a maximum of 255. The pixel intensity frequency is shown on a logarithmic scale.*

The most simple of the automated thresholding methods include the mean threshold, which as the name suggests, is set as the average pixel intensity value of the entire image. Another example is the minimum threshold. This is implemented by a continual histogram smoothing until there are two distinct peaks. The minimum intensity value between these peaks is then chosen as the threshold. Many more complex global thresholding algorithms exist, including the Otsu, Li, IsoData, Yen and triangle thresholds. These algorithms all aim to provide an optimum intensity cut and their suitability is dependent on the histogram's composition.

3.1.1 Otsu Method

The Otsu method was developed in 1979 by Nobuyuki Otsu as a non-supervised process for automatically determining a threshold value [125]. It works by maximising the inter-class variance of an image, or minimising the intra class variance (either of these produce the same result). For a given image comprising N pixels made up of L total intensity values, there are n_i pixels each with intensity value i. The probability of a specific intensity occurring is $p_i = \frac{n_i}{N}$. Some arbitrary threshold, T divides the intensities into two classes, C_0 and C_1, for which the cumulative probability, mean intensity value and class variances are given by $P_0(T)$, $P_1(T)$, $\mu_0(T)$, $\mu_1(T)$, $\sigma_0^2(T)$ and $\sigma_1^2(T)$, respectively. The mean intensity value of the overall image, μ can be calculated as shown in Eq. 3.2 and used to calculate the between-class (inter-class) and within-class (intra-class) variances, as shown in Eq. 3.3 and Eq. 3.4.

$$\mu = \sum_{i=1}^{L} ip_i \tag{3.2}$$

$$\sigma_b^2(T) = P_0(T)(\mu_0(T) - \mu)^2 + P_1(T)(\mu_1(T) - \mu)^2 \tag{3.3}$$

$$\sigma_w^2(T) = P_0(T)\mu_0(T) + P_1(T)\sigma 1^2(T) \tag{3.4}$$

The optimum threshold, $T*$ is then found by maximising the inter-class variance, Eq. 3.5 or minimising the intra-class variance, Eq. 3.6.

$$T* = \underset{1 \leq T < L}{\arg} \max \{\sigma_b^2(T)\} \tag{3.5}$$

$$T* = \underset{1 \leq T < L}{\arg} \min \{\sigma_w^2(T)\} \tag{3.6}$$

3.1.2 Li Threshold

Li thresholding is a method that defines an image's "optimum" threshold by minimising the cross entropy (information content) between the original image's probability distribution, P, formed of values p_i, and its segmented version's distribution Q, made up of q_i values [126][127], defined in Eq. 3.7.

$$D(P, Q) = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} \tag{3.7}$$

This method begins with three unknown values, the optimum threshold, t, the range of gray values, i to be assigned to the background, $[1, t-1]$, denoted as class a, and the range to be assigned to the foreground, $[t, L]$, denoted by class b. L is the image's maximum gray value. By using the image histogram distribution's zeroth and first moments (area and mean) as defined in Eq. 3.8 and Eq. 3.9, we can obtain the portions' means, $\mu_a(t)$ and $\mu_b(t)$ as in Eq. 3.10.

$$m_{0a}(t) = \sum_{i=1}^{t-1} h(i) \qquad m_{0b}(t) = \sum_{i=t}^{L} h(i) \tag{3.8}$$

$$m_{1a}(t) = \sum_{i=1}^{t-1} ih(i) \qquad m_{1b}(t) = \sum_{i=t}^{L} ih(i) \tag{3.9}$$

$$\mu_a(t) = \frac{m_{1a}(t)}{m_{0a}(t)} \qquad \mu_b(t) = \frac{m_{1b}(t)}{m_{0b}(t)} \tag{3.10}$$

The function to be minimised is found, as in [129], to be

$$\eta(t) = -m_{1a}(t)\log(\mu_a(t)) - m_{1b}(t)\log(\mu_b(t)) \tag{3.11}$$

As the minimum of Eq. 3.11 is found at the point where its derivative is equal to 0, the optimum value of t, is finally defined as

$$t = \frac{\mu_b(t) - \mu_a(t)}{\log(\mu_{b(t)}) - \log(\mu_{a(t)})} \tag{3.12}$$

By applying the one-point iterative method to Eq. 3.12 an iterative formula shown in Eq. 3.13 can be found for calculating the optimal threshold [129].

$$t_{n+1} = \text{round}\left(\frac{\mu_b(t_n) - \mu_a(t)}{\log(\mu_b(t_n)) - \log(\mu_a(t_n))}\right) \tag{3.13}$$

The use of Li thresholding is of course dependent on the composition of the image to be segmented. Factors such as intensity difference between the background and foreground classes and the size of the foreground objects can impact the results obtained, as discussed by Yazid *et al.* [130].

3.1.3 ISODATA Method

The ISODATA method, also known as the Ridler-Calvard method utilises mean class values [131]. Here, the process for thresholding into two classes is given, but this method can be extended for as many classes as required. Firstly some random values, μ_{a0}, μ_{b0} are assigned as the initial mean values of two pixel classes, a and b. The threshold value separating these classes is then found by taking the average of the mean class values, as shown in Eq. 3.14. All pixels are then assigned to the class whose mean their value is closest to.

$$T_0 = \frac{\mu_{a0} + \mu_{b0}}{2} \tag{3.14}$$

Updated mean class values are then calculated along with the corresponding new threshold value, T_1. If $T_0 \neq T_1$, T_1 replaces T_1 and the process repeats. When $T_0 = T_1$, the method has converged on the final threshold value to be used in the image segmentation.

3.1.4 Yen

The Yen threshold was developed in 1995 and is based on a maximum correlation criterion [132]. It was created as a more computationally efficient alternative to algorithms utilising entropy measures and is based on two quantities; the discrepancy between the thresholded and original image, and the number of bits required to represent the thresholded image. The optimum threshold values are found by defining a cost function that takes both of these into account and minimising both it and the classification number of grey levels. This algorithm was created for multi-level thresholding but by setting the number of grey classification values to two it can be applied for a single threshold.

3.1.5 Triangle

The triangle threshold method was first implemented by Zack *et al.* in 1977 when analysing sister chromatids [128]. It works by first constructing a line between the histogram's peak and farthest end. The end is set as the first (or last) non-zero bin. The longest perpendicular line connecting the top of a bin to the first line is then found. An offset can be applied to the obtained intensity, as in Figure 3.2 to correct the threshold value, but was not done in the implementation of the algorithm presented here.

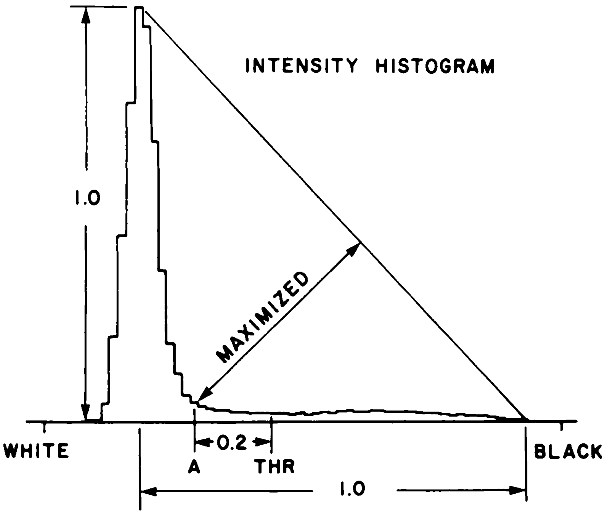


Figure 3.2: *A visualisation of how the triangle method is applied to automate a threshold value [128].*

3.2 Adaptive Thresholding

The global thresholding method processes an image as a whole, and disregards any relationships that may be present in subsections of the image. Local thresholding (also known as adaptive or dynamic thresholding) attempts to identify and exploit these relationships, potentially leading to more accurate segmentation. Examples where local thresholding can be highly effective include cases where an image has illumination changes e.g. shadowed areas. It would make sense in this case for threshold values to be determined in local regions rather than taking the image as a whole. The threshold chosen for each pixel is dependent on its neighbours, the number of which and the methods used can be optimised for a particular use case. An example would be using a 9 × 9 block around the pixel to be thresholded. This local region would then be analysed using some threshold finding method to determine the threshold to be applied to the central pixel. Varying the block size affects the threshold calculated, with large block sizes being more computationally demanding. The best method to implement is dependent on the input image and its histogram. Some studies have found that entropy methods can give the best results, but it is important to investigate the results of applying a variety of thresholds for a specific image set [133].

3.3 DBSCAN Algorithm

In some cases, it is necessary to define individual objects following segmentation. This allows them to be manipulated individually for measurements. Clustering algorithms are unsupervised methods for regions with specific characteristics. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is an algorithm that is successful in finding sample regions of high density and expanding clusters from them [134].
It has the advantage of not requiring the user to pre-define the number of clusters to be found, unlike the K-Means clustering algorithm [135] [136]. It also is a lot more user friendly, only requiring two parameters to be set and optimised for its use; *epsilon* and *minPoints*. For a two dimensional example, *epsilon* is the radius of the circle defined around each point within which the density is defined, and *minPoints* is the minimum number of points required within the circle for it to be classified as a core point.

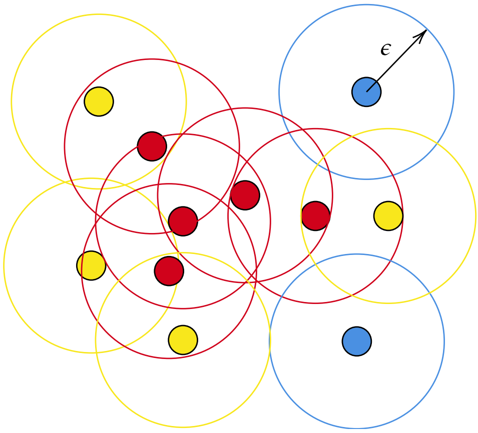


Figure 3.3: *A depiction of the DBSCAN algorithm. All data points have a circle of radius epsilon around them used to determine how many points are in their local area. Each point is then categorised as core (red), border (yellow) or noise (blue) dependent on the local points density.*

Border points are defined as those that have less than *minPoints* but at least one point in their circle. Noise is defined as solitary points with no others in their circle. Figure 3.3 shows an example of how points are classified for a *minPoints* value of 3.
DBSCAN uses Euclidean distance to locate points in space and has the added advantage over other algorithms of only needing to scan the data once. While there are only two parameters to optimise, the algorithm is very sensitive to the precise values of *epsilon* and *minPoints*. It is therefore important to explore and optimise these values according to

the dataset. The original paper suggests setting $minPoints = 4$ for two dimensional data whilst *epsilon* is found through inspection of a graph which plots the average distance between each point and its k-nearest neighbours [134]. There will be a turning point in this graph which separates the noise and cluster components of the image, as shown in Figure 3.4. This is widely known as the graph's elbow and is used to determine *epsilon*.

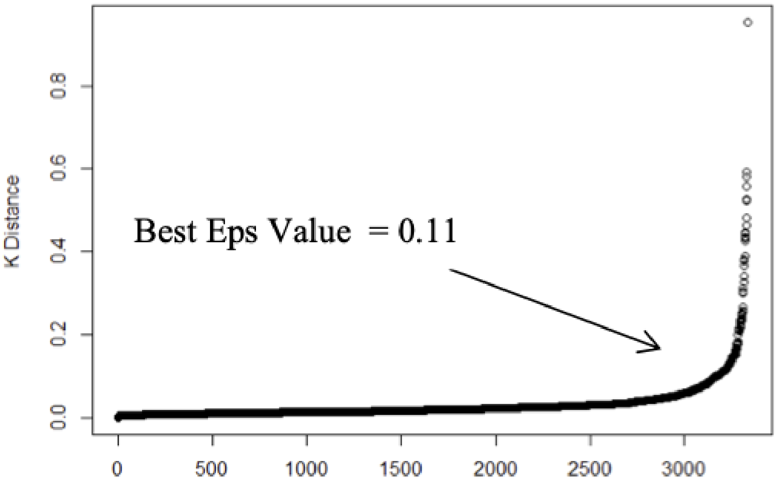


Figure 3.4: *An example plot of the distance between between each point and its k-nearest neighbours (3rd nearest in this example). The turning point of the graph is highlighted, from which the best epsilon value is set for use in the DBSCAN algorithm [137].*

3.4 Watershed Algorithm

The watershed algorithm is widely used to perform segmentation of images in which individual regions need to be identified and labelled. The concept of this algorithm stems from considering the pixel intensities as representative of the elevation above sea level. To implement the algorithm, markers are required which can be user defined or found using some automatic procedure implemented for the specific image type [138]. Examples of marker choice include the centre position of objects, calculated from the maximum distance from the image background and local minima calculated from the image gradient. The process can be visualised as water filling up from these minima in catchment basins. Watershed boundaries are defined as regions where catchment basins meet and are what segments adjacent objects from each other, as shown in Figure 3.5.

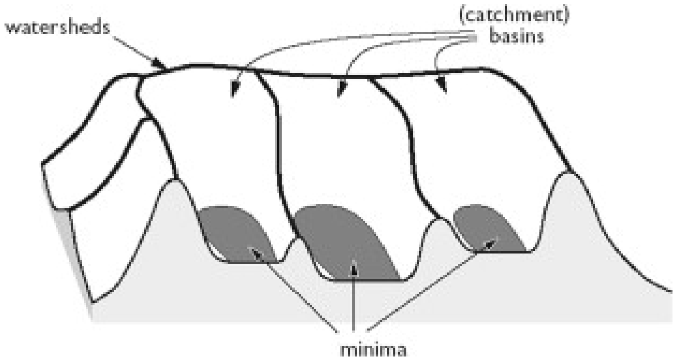


Figure 3.5: *A visual representation of the watershed algorithm and how catchment basins are filled with water from points of minima until adjacent regions meet to define watershed boundaries [139].*

This algorithm is not suited to noisy images as "over-segmentation" occurs, where small basins caused by background variations are found. Such images require some pre-processing to remove or reduce noise [139]. Variations of the watershed algorithm have been proposed for use in medical image segmentation to tackle the issue of over-segmentation [140].

3.5 Neural Networks

Neural Networks were first theorised in McCulloch and Pitts' paper of 1943, in which they outlined the logical calculus of nervous activity within the brain and its various applications [141]. Since then, the field has gone through many periods of rapid advancement. The present day understanding of the perceptron was introduced in 1957 by psychologist Frank Rosenblatt, as a means of conveying the brain and its receptor functions [142]. An overview of the perceptron model is shown in Figure 3.6.

This simple initial model shows how various inputs from the retina are sent to "Associate units" within the brain and undergo some transformation according to weights vector **y** to become the **a** vector. A summation is then performed on all elements of **a** before a final classification takes place in the "Response unit", dependent on the summed value. Simply put, the perceptron is the most basic form of a neural network, consisting of a single layer and acting as a binary classifier.

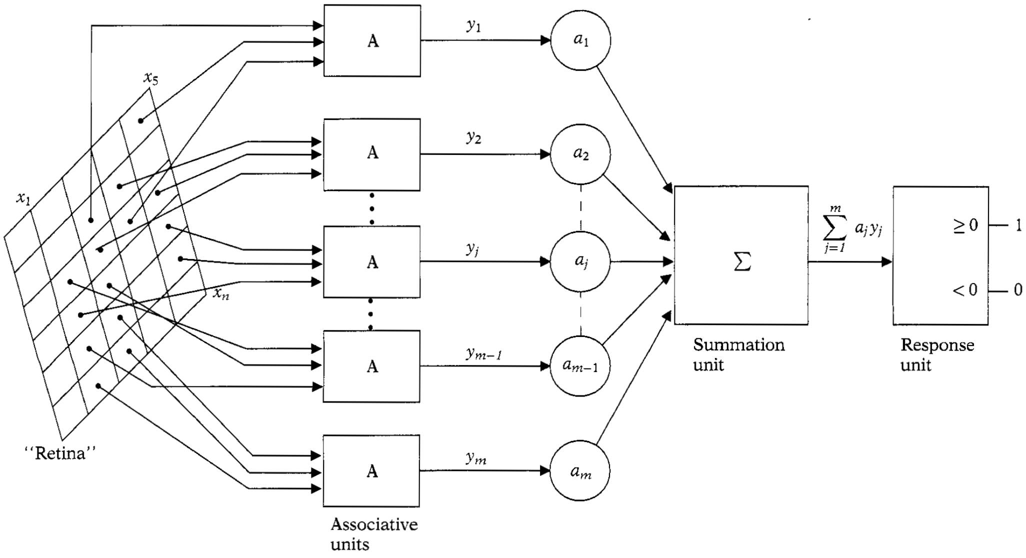


Figure 3.6: *The Mark 1 perceptron; a model of the neurons in the brain, adapted from [143].*

For more complex problems, for which there are more than two possible outputs, more complex neural networks are required; the multi-layer perceptron. This type of network, as the name suggests, has additional layers, known as hidden layers between input and output layers, as depicted in Figure 3.7.

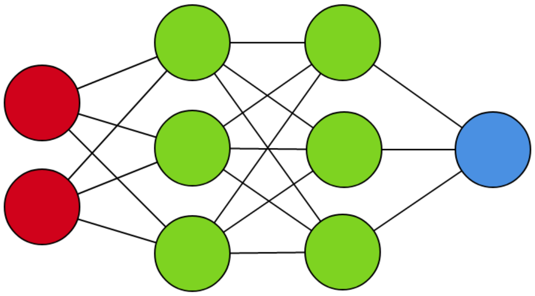
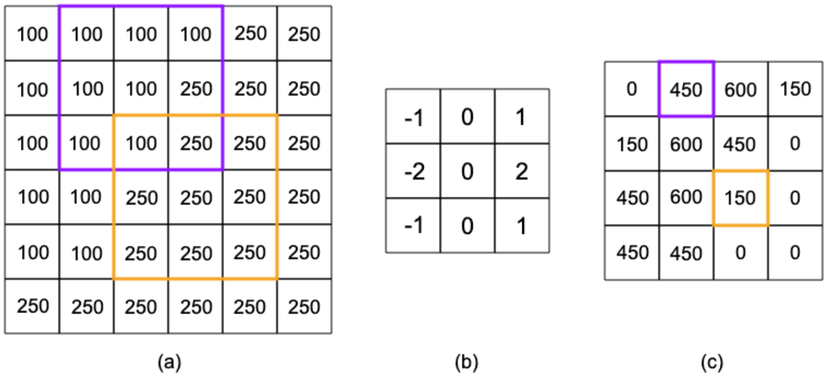


Figure 3.7: *An example multi-layer perceptron that has 2 hidden, fully connected layers (green), between input (red) and output (blue) layers.*

The extra layers allow a deeper type of learning to occur, during which the network can realise a whole range of relationships and features of the input data, and is why they are commonly known as deep neural networks. The hidden layers also can be of varying types. One example is known as a fully connected layer, in which each node of a layer is connected to all nodes of the previous layers, so that it receives all the information from that previous layer.

Another commonly used layer is the convolutional layer. This is much more selective in how adjacent layers connects to one another, so not all nodes receive the same information. This provides a higher level of flexibility in learning, which can be useful when searching for specific features of data, such as the edges within an image. Here, the data from neighbouring pixels is aggregated to find relationships between them dependant on the convolution filter that is applied. For example, in Figure 3.8 a set of pixels are convolved with a filter designed to detect vertical edges, called the Sobel operator (also known as the Sobel filter or mask) [80]. This filter slides across the image, performing calculations on each 3 × 3 local region and outputs a convolved feature of smaller dimensions than the input image.



100	100	100	100	250	250
100	100	100	250	250	250
100	100	100	250	250	250
100	100	250	250	250	250
100	100	250	250	250	250
250	250	250	250	250	250

(a)

-1	0	1
-2	0	2
-1	0	1

(b)

0	450	600	150
150	600	450	0
450	600	150	0
450	450	0	0

(c)

Figure 3.8: *An example of applying the vertical Sobel filter, (b) to a set of pixels (a). The output is shown in (c) where the convolved values for two sets of pixels are highlighted in orange and purple.*

In Figure 3.8, the region indicated by the purple square in (a), vector A, is convolved by the vertical edge filter shown in (b), vector B, to produce the value of 450 in (c). This is executed by calculating the dot product of vectors A and B, the output of which is 450.

The benefits of convolutional layers over fully connected ones is the, often large, computation decrease. If the information from all nodes is passed to all nodes of the following layer, a lot of information is passed on all the way through the network before some transformation and summations are performed prior to the output layer. This is both very computationally and time intensive, requiring a lot of resources, whereas convolutional layers by definition transform data into a simpler, more compact form which requires less time and computing power. However, the optimum architecture required is dependant on the problem it is applied to, and therefore should be considered and optimised depending on the specific use needs.

3.5.1 Mask R-CNN Instance Segmentation

Machine Learning provides options for implementing automated image segmentation tools in a more robust way to traditional methods. Mask Region-Based Convolutional Neural Network (R-CNN) is an architecture that is able to segment targeted objects within an image and create a pixel-wise mask of the pixels it identifies as belonging to that object [144]. It is based on the Faster R-CNN network [145], adding on the additional mask feature that provides its name, as show in Figure 3.9. The aim of Faster R-CNN is to take input images and produce three outputs: potential bounding boxes, a label of the identified object in that bounding box and a probability for each label.

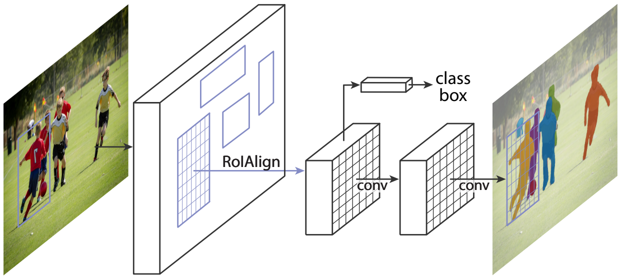


Figure 3.9: *The Mask R-CNN framework showing an input image passing through the network. Regions of interest are aggregated in the RoI Align layer and the following convolutional layers perform classifications to produce the output bounding box and and segmentation masks [144].*

3.5.2 Faster R-CNN

Faster R-CNN is comprised of three main sections; the backbone Convolutional Neural Network (CNN), a Region Proposal Network (RPN) and Region of Interest (ROI) Pooling, as shown in Figure 3.10. The CNN is a deep learning network that has already been trained, such as ResNet [146] or VGG [147]. The exact architecture of this initial network determines the images it can read in. For example, the VGG network has fully-connected layers in its final block instead of convolutional layers. These require a fixed size input which dictates that the original input to the network must be 224 × 224 × 3 (square 224 × 224 RGB images). ResNet has convolutional layers throughout so does not have this requirement. It takes the input image and represents them as $Height \times Width \times Depth$ tensors, according to the image dimensions and reduces the dimensionality down considerably in order to extract features. Convolutional layers are able to do this by aggregating neighbouring pixels to find relationships between them dependant on the convolution filter that is applied. The features obtained from the CNN are stored in a feature map and passed to the RPN.

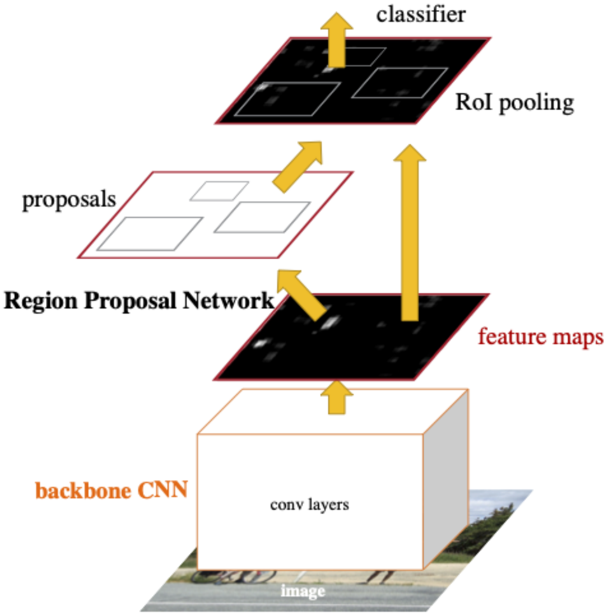


Figure 3.10: *The Faster R-CNN network, comprising of a backbone convolutional neural network, a region proposal network that produces feature maps and region proposals. A RoI pooling layer collates all the regions of interest before classification. Adapted from [145].*

The RPN's aim is to find potential bounding boxes (bboxes) for the target object using the extracted features. "Anchor" boxes are utilised to act as reference points of known co-ordinates within an image. These are tied to specific locations and are of varying scale and aspect ratio, so that 9 separate anchors cover each pixel. Checks are applied to each anchor to firstly see if it contains the object to be detected and then to determine how best the anchor could be resized to better fit the object. The region found from this process then becomes a ROI, and a calculation is performed on each identified ROI to determine its intersect over union (IoU) value. This determines the amount of overlap between the proposed ROI and the ground truth bbox for that object, as defined in Eq. 3.15.

$$IoU = \frac{pix_{ROI} \cap pix_{truth}}{pix_{ROI} \cup pix_{truth}} \tag{3.15}$$

An IoU value of ≥ 0.7 is commonly used for determining which anchors are providing a positive class label during training of the RPN, although this can be adjusted to another value or use the highest IoU ROIs [145]. Anchors with IoU ≤ 0.3 are deemed non-positive and the network learns that the information within these regions is not the object class and also contribute to training. All other regions with values outside the defined thresholds ($0.3 <$ IoU < 0.7) do not contribute to the training effort. A RoI Pool layer then extracts features from each proposed region to perform classification and bounding-box regression. A disadvantage of RoI Pool is that it uses quantization to reshape bounding boxes to the correct size for the subsequent fully connected (FC) layer. This results in a loss of region information due to the use of floor division, and lowers the models precision. Mask RCNN overcomes this issue by implementing an updated RoI Align layer instead, which uses the whole RoI's area to pool from, losing no data. The FC layers that follow perform classification where the region is defined with an object or background label, and the final bounding box co-ordinates are produced.

Instance Segmentation

This is where Faster RCNN's work ends and the "mask" part of Mask RCNN comes into play. In order to produce a mask of an identified object, a region's pixels need to individually classified as belonging to the object or not; the region needs to be segmented. There are two main types of segmentation implemented under the Faster RCNN architecture, semantic and instance segmentation. As shown in Figure 3.11 the semantic version groups all identified object pixels into an overall identification. It recognises all balloons in the image but cannot distinguish between each one. Only one mask would be produced when using semantic segmentation, and it would contain all pixels identified as balloons. Instance segmentation on the other hand separates each individual balloon and defines them each with a distinct mask. This is achieved by utilising the bounding box identification method with pixel-wise classification. This of course has a great advantage as individual masks can be used to further analyse each object separately, lending to a whole range of applications such as medical health and intelligent driving [148].

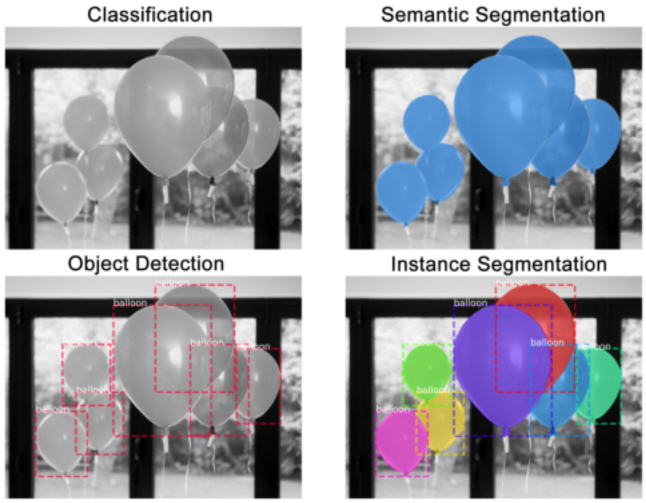


Figure 3.11: *An example of balloon identification using different methods; classification, object detection, semantic segmentation and instance segmentation [149].*

This part of the Mask R-CNN's architecture is what dictates the different training input compared to the Faster R-CNN network. In order to learn pixel wise classification during training Mask R-CNN requires segmentation masks alongside train images to learn from. These are given in the form of x/y co-ordinate points of polygons that enclose the target object, rather than the traditional bounding boxes. These co-ordinate files are read in as JSON files and can be created using a variety of programs such as VGG [150] [151] and, LabelMe [152].

Transfer Learning

It is very rare that an individual has the data available to train a model using the Mask R-CNN framework from scratch. It requires hundreds of thousands of images in order to learn the necessary relationship and features to accurately detect objects in new images. And even if one did have the necessary data to train from, it is a very time intensive process to annotate this training set and produce the required co-ordinate file to learn from. Transfer learning is a machine learning method that removes the need for an extensive dataset. It allows a user to take a trained model that has learned a lot of object detecting features and stored this knowledge, and apply it to their new, smaller dataset [153]. This is possible by only retraining the old model's final few layers, where classification takes place but now on the new data.

There are a number of extensive models that have been trained on enormous datasets including Microsoft's Common Objects in Context (COCO) model [154]. It includes over 90 object classes of which 82 have over 5 000 labelled instances [154]. The total dataset contains 2 500 000 labelled instances within 328 000 images. This dataset has been trained using the Mask RCNN framework in TensorFlow to create a model that identifies and segments objects within an image. By using the existing model, only the head layers of the model require training, allowing not only a minimal training time but also a considerably smaller dataset to train from.

Chapter 4

Identifying Comets in an Image

4.1 Identifying comets

The first step prior to comet measurement is to identify each comet present within an image. This is carried out by segmenting the foreground (comets) from the image background. Humans are easily able to perform this segmentation as comets appear in relatively regular shapes as bright objects on a dark background. This is more difficult for a computer, especially with the added complexity of contaminants that are often present. A variety of methods are explored here to perform this segmentation, investigating the results from applying a range of global and local thresholds alongside the watershed and DBSCAN algorithms. They are compared against manual segmentations in which images have been annotated by hand to test their accuracy and robustness.

4.1.1 Thresholding Analysis

Thresholding creates a binary class of pixels that are equal to or above an assigned intensity value, or below it, as discussed in Chapter 3. An initial test was performed using scikit-image's *try_all_threshold* method [155]. This applies the isoData, Li, mean, minimum, Otsu, triangle and Yen thresholds in a global fashion to an image and outputs the thresholded versions, as shown in Figure 4.1. The purpose of this is to observe the different thresholds and their effect on comet data.

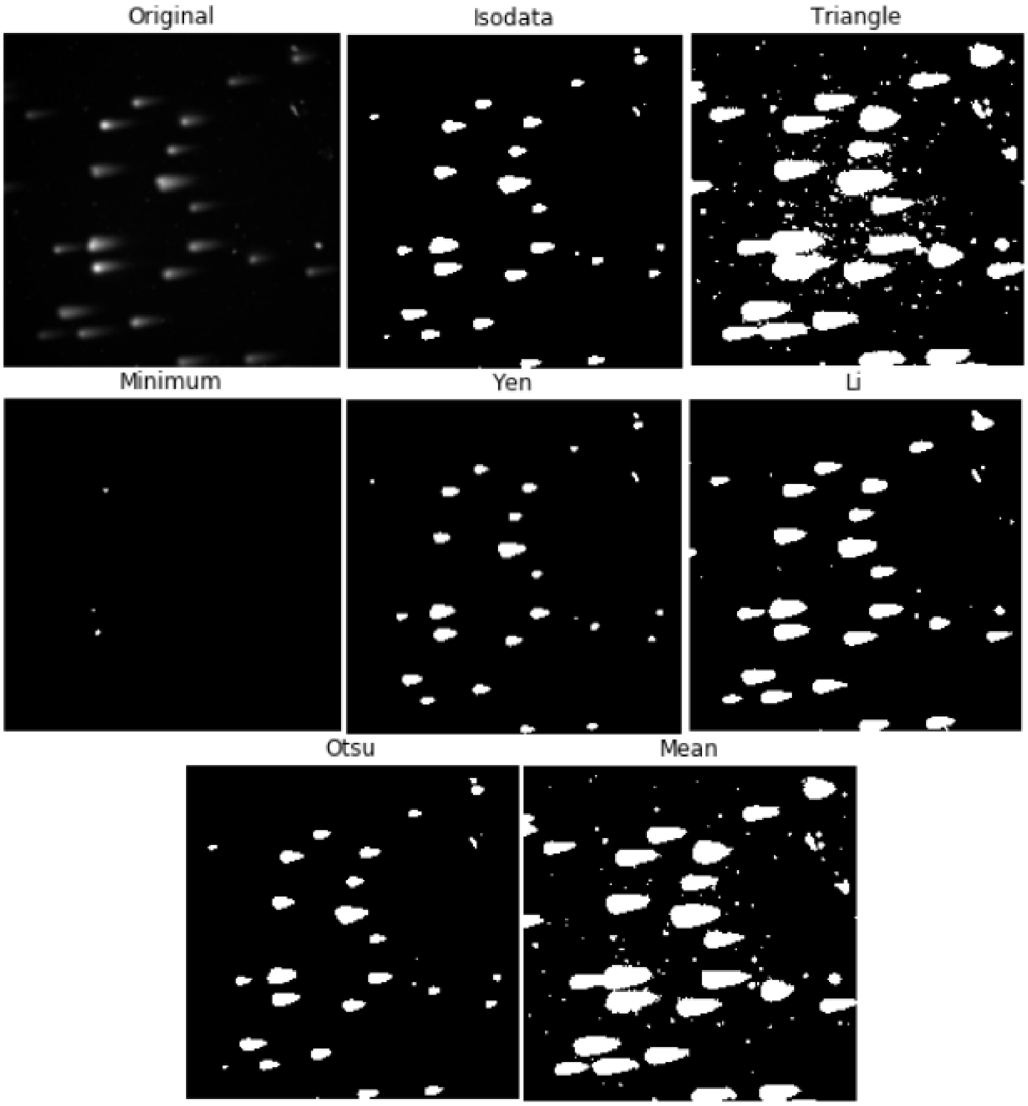


Figure 4.1: *Applying scikit-image's try_all_threshold method to a comet assay image to produce various thresholded images.*

A variety of 20 comet images were tested in this manner and it was found that the minimum and Yen thresholds consistently performed the poorest of the evaluated methods in segmenting the comets. This is expected from the minimum method, as comet images do not have two distinct peaks in their intensity distribution. Excessive smoothing is required to produce a second peak to then find the subsequent minimum value between peaks. Considering the spectrum of a typical image, shown in Figure 4.2, the second peak is likely to be found towards the right end of the intensity range, producing a higher threshold than needed for this image composition.

The Yen threshold consistently produced too high thresholds, causing a loss of comet data comapred to the other methods (Otsu, Isodata and Li in particular) and was therefore not explored further. The automated results using the Mean and Triangle thresholds segmented the comets, but the calculated thresholds were not high enough to achieve the separation required.

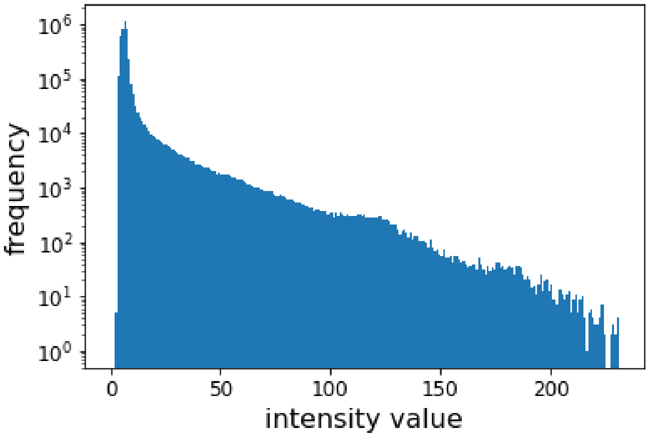


Figure 4.2: *The histogram of a raw comet assay image, showing its intensity distribution. Frequency has been plotted on a logarithmic scale.*

An aim of automating comet measurements is to maximise the number of comets detected per image to better represent the available data. Local thresholding was performed through scikit-image's local threshold routine to determine if this would improve the comet segmentation by taking into account the local pixel neighbourhood relationships. This requires parameterisation of the block size, determining the pixel neighbourhood size to analyse, and the thresholding method to implement. Block size must be a positive odd value as it includes the central pixel plus the surrounding neighbours. For example, the smallest block size possible is the 3 × 3 neighbourhood around a pixel for which the threshold is determined. Any other neighbourhoods must also have an odd value for x in an $x \times x$ environment, in order to keep symmetry about the central pixel in both the x and y directions. The built-in threshold methods implemented by this routine are the Gaussian, mean and median rank filters. The Gaussian method weights pixels closer to the central pixel with a higher contribution to the threshold value calculation and those further away with less. It uses a σ value (representing the standard deviation of the Gaussian) provided by the user or automatically calculated using the defined block size, which will cover $>$ 90% of the distribution, as in Eq. 4.1.

$$\sigma = \frac{blocksize - 1}{6} \tag{4.1}$$

The mean method, as the name suggests, calculates the neighbourhood mean. Similarly, the median method applies an appropriate filter to calculate the neighbourhood median and assigns it as the threshold value. The results of applying these local thresholding methods are shown in Figure 4.3. It can be seen that all three local thresholds are effective in segmenting the comets (to varying degrees) but a significant amount of background and contamination remain. There is a significant "clear" area immediately around the comets, attributed to the block size used (99), which is of comparable size to the comet dimensions.

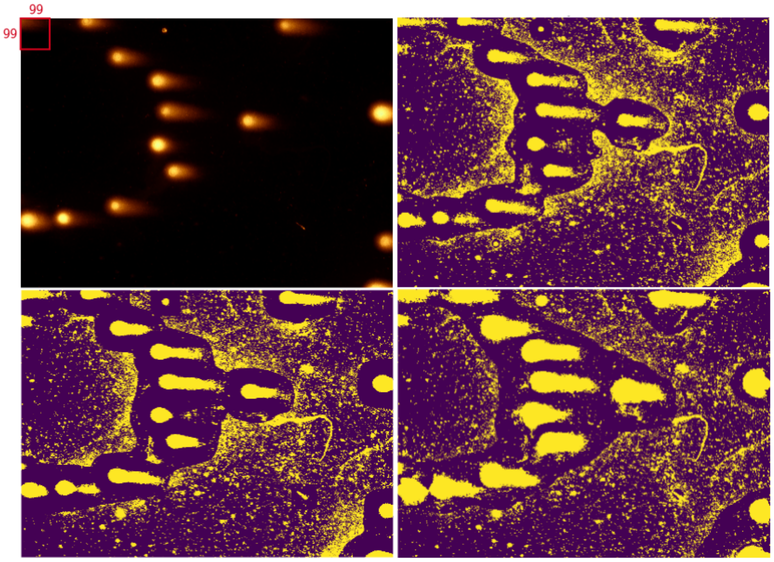


Figure 4.3: *The result of applying local thresholds to a raw image (top left) using the Gaussian (top right), mean (bottom left) and median (bottom right) methods with block size 99, depicted in red on the raw image for scale.*

The high background variation, revealed through local thresholding, suggests there is more background fluctuation than the human eye can process at these darker intensities.

This needs to be taken into account when performing the background corrections, discussed in Section 5.2.6. The block size was chosen after testing a range of values across different images, as shown in Figure 4.4. Lower values showed a greater sensitivity to the background variation, resulting in a much worse comet segmentation. Comets and clusters will be used interchangeably from here on as comet bodies are analysed and tracked as pixel clusters following the application of the watershed algorithm.

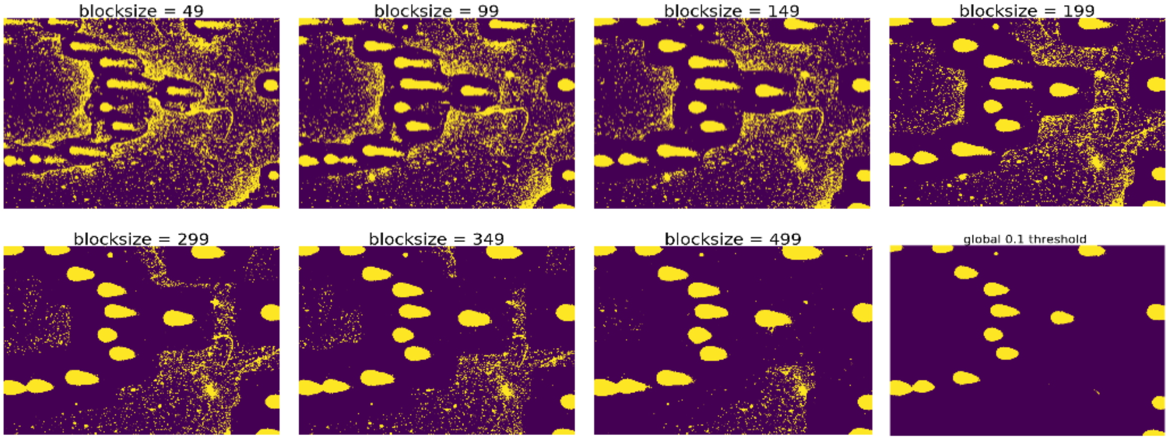


Figure 4.4: *Local mean thresholds applied to an image with varying block size values. The same image with a global threshold of 0.1 of the maximum pixel intensity is shown for comparison.*

Using a larger block size did decrease the background detected, however it was only a significant reduction for very large values (> 300) which tend toward the same results as when a global threshold is used, defeating the point of using a local threshold. These larger values are also more time intensive to implement, as the calculations are performed on a larger number of pixels, and would defeat the objective of speeding up the analysis process. It was therefore concluded that local thresholding is not suitable for comet assay images, due to the high inconsistent variation of the background. So a global threshold was used as a first step in comet identification.

4.1.2 Selecting a global threshold value using DBSCAN

A manual threshold was trialled, but of course posed the issue of how it should be selected, and how its variation affected comet identification. To determine the most suitable threshold value, the DBSCAN algorithm, with $minPoints = 4$ and $epsilon = 3$ was applied to an image thresholded at various values, as in Figure 4.5, to determine which provided the most accurate depiction of the clusters present. Separate clusters are distinguished with their own colour and assigned a numerical label. Clusters are labelled with a cross at their intensity weighted centre, making it easier to see the number of noise points that are defined as clusters, as individual pixels are not as clearly seen by the eye.

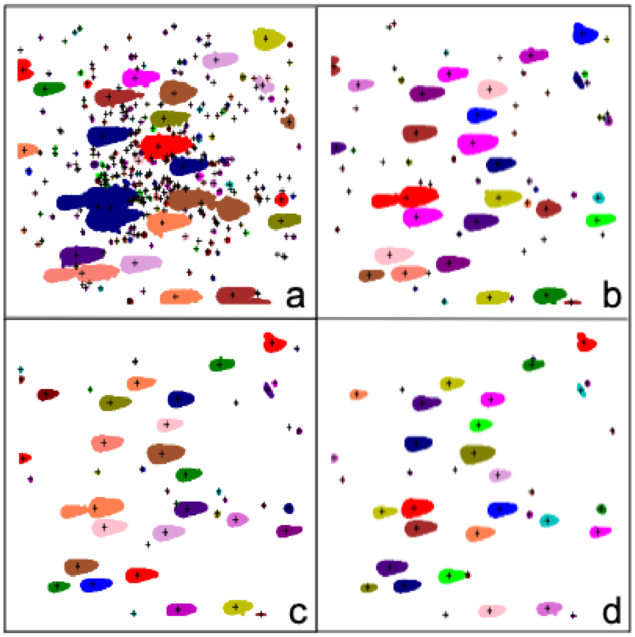


Figure 4.5: *The result of applying a range of threshold values, a) 4%, b) 6%, c) 8% and d) 10% of the image's maximum pixel value, to a comet assay image followed by the DBSCAN algorithm. Separate clusters are shown in different colours.*

Due to artefacts in the image from contamination, thresholding produces a considerable amount of noise in the image, particularly for lower values, which is grouped into its own set of clusters and needs to be filtered out before conducting any comet analysis. The methods for filtering out these unwanted clusters are defined in Section 4.1.3. This was again tested for a range of images and it was found that a threshold value of 10% of the maximum intensity value provided the best segmentation for a range of tested images. There were a lot of cases where a threshold increase from 8% to 10% did not provide any additional cluster separation and resulted in the loss of data around the cluster edge. The higher value was chosen despite this, as extracting as much information from each image is one of the goal's of this analysis and the lost comet data is compensated for by cluster expansion later on, discussed in Section 4.3.

The values of *minPoints* and *epsilon* were defined as 4 and 3, respectively. These were determined from the algorithm's recommendation for 2 dimensional data and from a plot of the average distance from each point to its k-nearest neighbours, as previously discussed in Section 3.3.

4.1.3 Watershed Analysis

The watershed algorithm is applied to an image, following a threshold at 10% of the maximum pixel intensity. As previously mentioned in Section 3.4, this algorithm requires markers to be defined. Here the cluster centres have been used, calculated by using the maximum distance from the background. The watershed algorithms then identifies separate clusters and assigns them a numerical label so that they can be individually identified, as shown in Figure 4.6. Here it can be seen that the many noise and contamination points have been identified as clusters (Figure 4.6b). This is because even a single pixel is identified as a cluster, unlike DBSCAN which requires a minimum number of points to be defined. Figure 4.6c shows that setting a pixel number filter as low as 20 for each identified cluster removes a lot of the tiny clusters, but a significant number still remain, such as clusters 20-25. These are of course not comets and are not pertinent in determining the DNA damage present, and need to be removed by some automated method.

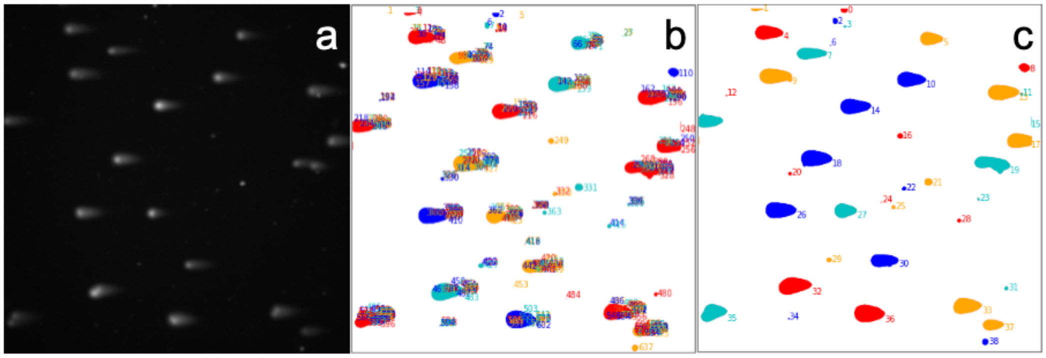


Figure 4.6: ***a**: The raw unprocessed image. **b**: Cluster identification after applying the watershed algorithm to a thresholded image at 10% of the maximum pixel value. Individual clusters and their assigned numerical label are shown in separate colours. **c**: Cluster identification after applying the watershed algorithm and setting a minimum number of pixels per cluster as 20.*

Filtering clusters

Filtering is essential to not only remove bodies of contamination but also exclude any “bad” comets from contributing to damage measurements. “Bad” comets are defined as those on the image edge and overlapping comets. Filtering objects on the image edge is the easiest to implement. A check is performed to see if any of a cluster’s points lie within a 10 pixel border around the image edge. If this is true, that particular cluster is removed from the saved clusters. The vast majority of contaminant objects are considerably smaller than comets, so a first parameter to filter on is the cluster size, defined by the number of pixels the cluster is comprised of. A lower constraint for size filters out contamination and noise, whilst an upper bound removes the cases of overlap where two or more comets have been identified as one object due to their close proximity. Analysis of 788 “good” comet clusters (those that do not encompass any contamination or overlap) across 50 images was performed to find the average, minimum and maximum area values (in pixels) which were 9 773.29, 3 134 and 26 835 respectively. These values were used to set initial lower and upper bound constraints for area as 3 000 and 28 000.

Similar analysis was performed to determine the average height, width and height to width ratio of comets, providing lower and uppers bound constraints as 70 and 270 for the comet width, 50 and 150 for the height and 1.0 and 3.5 for the height to width ratio for the dataset analysed. These constraints are applicable to images of dimensions 2048 × 2048 and 1040 × 1392. taken with microscope magnifications of 10x or 40x. Any dimensions or magnifications much larger or smaller than these would require the constraints to be scaled accordingly. To account for the high degree of variation seen in comet assay data, the constraints were revised by a 10% (of the average) subtraction/addition to the lower/upper bounds. This was done to make the analysis inclusive of new data it may be applied to. The value of 10% was chosen by no analytical means, but rather in a crude and rough manner, to include a reasonable tolerance for new data that these constraints are applied to going forward. The 10% addition and subtraction to the upper and lower bounds was not applied to the height to width ratio as it had already been applied to the height and width individually, and would be an unnecessary addition. A summary of the constraints, and the values used to determine them, are shown in Table 4.1.

Although these filtering parameters were effective in removing a lot of “bad” comet instances, a significant amount of contamination and noise object were still being identified as acceptable comets. The main reason for this is the high variability in comet dimensions. Overlapping comets in the horizontal or vertical configuration can still be within the area, height and width bounds defined, due to the large range needed to describe this highly variable data. To overcome this, a final constraint was applied to measure how many bright objects were present within each identified cluster. The underpinning logic here is that a good comet will have only one bright region; the comet head. Many occurrences of contamination or overlap will be identified by more than one bright region, or none at all, and will be filtered out accordingly.

Measure	Area	Width	Heigh	Height-Width ratio
Minimum	3 134	71	51	1.03
Maximum	26 835	269	150	2.90
Average	9 753	157	83	1.90
Standard Deviation	4 097	41	17	0.34
Lower Bound (LB)	3 000	70	50	1.00
Upper Bound (UB)	28 000	270	150	3.50
Adjusted LB	**2 022**	**54**	**41**	**1.00**
Adjusted UB	**28 977**	**285**	**158**	**3.50**

Table 4.1: *A summary of the constraints applied to comets identified by the watershed algorithm. All values are given to the nearest integer, except for the height-width ratio values which are rounded to 2 decimal places.*

To identify the head regions, another global threshold was applied to each image, of value 25% of the image's maximum intensity. Ideally, a much higher value would be used as the threshold in order to separate out objects of contamination and overlaps, that would cause two brighter region to be in close proximity. However, a significant amount of the data used in this work included images in which comet heads were only slightly brighter than the comet bodies, and were not detected at all at higher thresholds. A compromise was made to use a lower value which would not filter out all instances of overlap and contamination within comets, but would identify the vast majority of comets within an image. One of the aims of the automated program developed here is to maximise the data used and extracted from each image, so it was decided that the lower threshold for finding comet heads was sufficient. It was also deemed less of an issue as human analysis of the results is still required in this methodology, to ensure the measurements performed are on the correct objects (single comets and not overlaps or contamination). This is further discussed in Chapter 5, where comet measurement is explored.

4.1.4 Mask R-CNN Analysis

As discussed in Section 3.5.1 the Mask R-CNN architecture can be implemented for instance segmentation. A model was trained using this network by implementing transfer learning from Microsoft's COCO (Common Objects in COntext) model [154]. The dataset used was comprised of 60 images for training and a further 6 for validation of the model. These images were randomly selected across many different experiments and image sets, where varying camera equipment, magnifications and resulting image resolutions had been used. This was important in order to train a network on an expansive and diverse enough set of images to improve the results of performing inference on new, unseen images. The limitation of only using 66 images stems from two factors. Firstly, there was a limited amount of "good" data available to train from, whilst also leaving a sufficient amount of images to test the model. "Good" data refers to images that do not have very large contaminations present, as those that do have large contamination could skew the training process and cause abnormalities to be detected as comet objects. The second limitation is the time taken to annotate an image set prior to training. During training, the network takes a JavaScript Object Notation (JSON) file of annotated points as an input. These points are (x,y) co-ordinates of vertices that describe a polygon mask around each comet in an image, as shown in Figure 4.7. Defining these polygons for each comet in each image of the train and validation sets is a cumbersome process, taking up a lot of time.

An online software called VGG Image Annotator (VIA) [151] [150] was utilised to perform the annotations. The interface, shown in Figure 4.7, allows images to be uploaded from a local directory or from a website address on the internet. It allows a choice of annotations such as rectangular bounding boxes, circles, ellipses, individual points, polyline regions and polygon regions, which have been used in this case. The closing of a polygon region creates a mask, defining all pixels that are assigned to the object it is defining, which is what the network learns from.

The network architecture has been built in Python 3 using the tensorflow and keras packages. An example balloon model was used to create this comet model [149]. All code and configuration files were accessed from the matterport Mask RCNN github repository [156]. The model training required GPU computation due to the many, large images used and was performed in Google Colaboratory's online cloud architecture in a Jupyter notebook environment [157]. The model parameters were mainly kept as the configuration used in the example balloon model but some changes were made such as training for 5 epochs of 100 steps each and applying 50 validation steps at the end of each epoch.

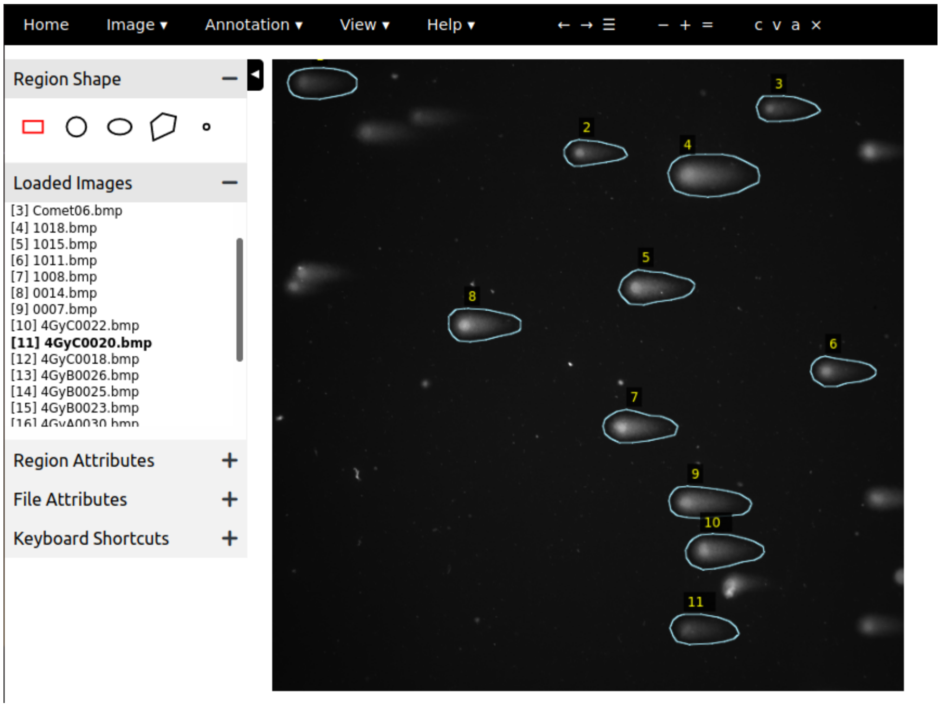


Figure 4.7: *A screenshot of the VIA software used to perform image annotations.*

The output of the model is a set of log files (one for each epoch of training) and a weights file (.h5) that stores all the learned relationships in order to perform inference. In this context, inference is the process of applying a trained model to new, unseen test data, to segment the object it has been trained to detect.

The result of applying the model is a mask of all segmented objects overlaid on the test image, as shown in Figure 4.8. The outputs of the model shown here for each identified object instance are the segmentation mask, bounding box, prediction confidence value and class name (comet). A threshold of at least 0.8 confidence was determined by inspecting a number of images. The vast majority of segmentation that had confidence values below 0.8 were on overlaps or did not define the comet region well. The output masks can be saved for later inspection, which was done to perform segmentation method comparison analysis, detailed in Section 4.2. The output is an array of dimensions rows × cols × n, which are the image height, width and number of objects segmented, respectively.

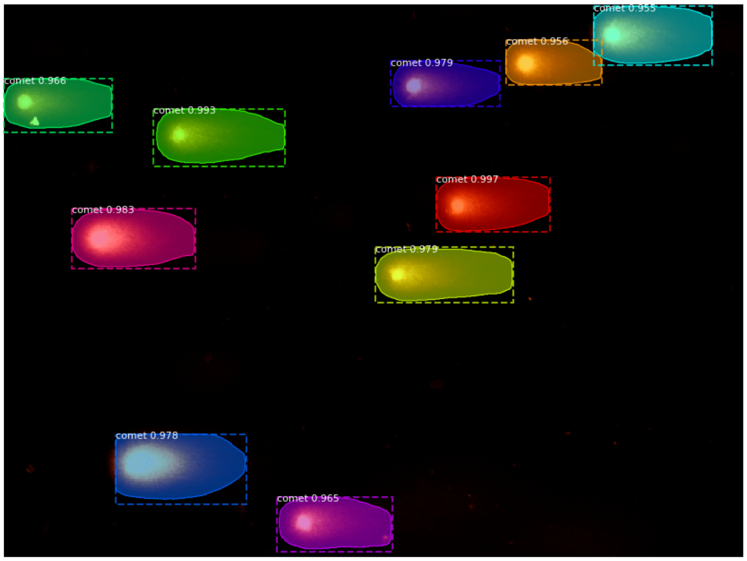


Figure 4.8: *A segmented image after performing inference from the trained Mask R-CNN comet model. Each identified comet is shown in an individually coloured mask and bounding box alongside the prediction confidence value and class name (comet).*

4.2 Comparing Comet Segmentation Methods

A dataset of 20 images was created for testing and comparing the two automated segmentation methods (watershed algorithm and Mask R-CNN). This image set was once again comprised of a broad range of images, captured from differing experiments using varying camera setups and configurations. This image set also needed to be exclusive of any images that were used in training the comet model and not so large as to make the manual annotating too time intensive of a process. Two images were chosen to compare the various methods, shown as examples in the remainder of this chapter and referred to as image A and B from here on. They were chosen as they are very different in regards to their resolution, the equipment used to capture them and the camera setting applied, and can demonstrate the effects of the two methods on different image types.

Image A is an example from a dataset acquired more recently, in 2021, from an experiment using UMSCC6 cells that have been irradiated with 100kV X-rays, using a higher resolution camera, in comparison to the dataset image B is from. This was acquired following the irradiation of HeLa cells with 60 MeV protons at Clatterbridge Cancer Centre. Images acquired similarly to image B are of a lower quality due to both their lower resolution and overexposure resulting in data loss. The histograms of these images are shown in Figure 4.9 to demonstrate their differences.

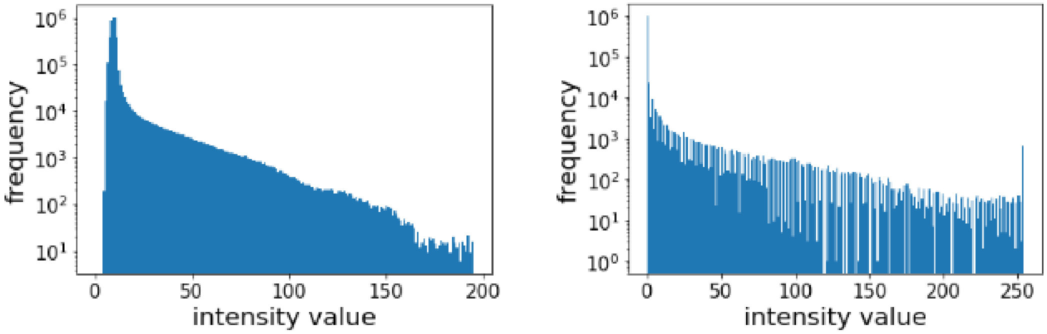


Figure 4.9: *Histograms of image A (left) and B (right) showing their intensity distributions. Frequency has been plotted on a logarithmic scale.*

Image A is of a higher resolution (2048 × 2048 pixels) than image B (696 × 520), accounting for its higher total frequency of intensities. The images also have different intensity ranges, with image A ranging from 4-196 whilst image B ranges from 0-255. The overexposure of image B is also apparent in Figure 4.9 from the spike in frequency in the last bin, suggesting that too long an exposure time was used during image acquisition. The format of these images is also very different. Image A was captured as a 16-bit TIFF image and converted to an 8-bit BMP image to significantly reduce its size for analysis. Image B was captured as an 8-bit TIFF format and converted to an 8-bit BMP, again, for file size reduction. However, the exact methods used to convert image B are unknown, as a single channel TIFF image (captured from the green light fluorescence of SYBR green stain) has become a three channel RGB image. This image was chosen for analysis here as a significant amount of data available was of this RGB form and all images were converted to greyscale prior to analysis, as single channel images are required for thresholding.

In order to have a baseline to compare each image to, and not just compare each method's results with one another, a ground truth data set was required. This was created by manually annotating segmentation masks for each image in the comparison set using the VIA software, as outlined for the training data annotations. As this is a manual process, it inherently creates some degree of human error which should be considered. In order to produce the most representative comet mask, encompassing as many pixels that should be categorised as comet pixels, each image was zoomed in on as much as possible during the annotating process. This made it easier to see where the comet boundary is, separating the comet from background. However, this was not as straight-forward a process as first thought. As shown in Figure 4.10, the described comet boundary is not a clearly defined region. When considering the individual human aspect of annotation, what one person would describe as the comet boundary may differ greatly to what another person would define it as.

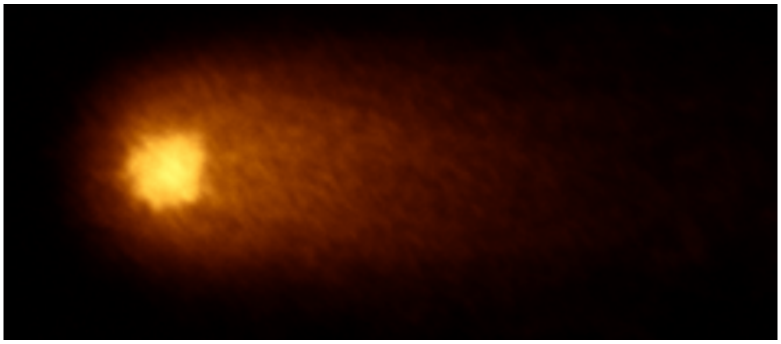


Figure 4.10: *Zooming in on a comet image to show the gradual decrease in intensity up to the comet boundary, which is not clearly defined.*

As mentioned in Section 2.1.1, the human brain has a greater sensitivity for dark intensity changes, but this sensitivity is still specific to each of our brain's functionality. This, coupled with varying settings of the monitor used to display the images during annotations, i.e. resolution, brightness, contrast and colour settings, all play a role in how an individual would define the comet boundary. To overcome this source of bias and error, a large set of people could each annotate the same set of images and the mask average could be used as the ground truth, but time constraints did not allow for this during the course of this work.

To compare the segmentations between the two automated methods (watershed and Mask RCNN) the identified comet masks were saved to file. This allowed a direct comparison with manually annotated images which were used as ground truth masks. The annotated polygon vertices (for manual annotations), extracted from the output JSON file, were used to create masks using the *points_in_poly* method from the skimage.measure python package [155]. This routine finds all the pixels within the polygon provided and outputs a mask. By plotting the masks produced from the different methods, a visual comparison was made, as shown for the two different images, A and B, in Figure 4.11 and Figure 4.12.

As can be seen, the results are varied. The watershed algorithm is much more consistent in the comets and shapes it finds, but hugely underestimates the comet region in some images. The large under-estimates occur in a considerable number of images, but not for all.

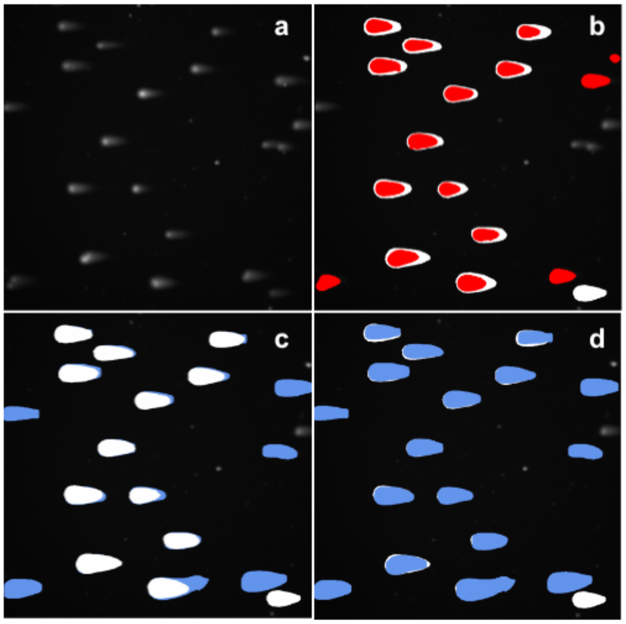


Figure 4.11: *A comparison of the comet masks produced from the watershed, Mask_RCNN and manual segmentations of image A.* ***a****: raw image,* ***b****: watershed segmented masks laid over manually annotated masks,* ***c****: manually annotated masks laid over Mask R-CNN segmented masks and* ***d****: Mask R-CNN segmented masks laid over manually annotated masks.*

Analysis was performed to explore if exposure, contrast or some other factor could be attributed to the amount a comet was under-segmented in comparison to the ground truth masks. Unfortunately, no such relationship was found, further stressing the highly varied nature of this data and the need for standardised image acquisition. The under-estimates can also be attributed to how the manual segmentations, used as ground truth, are defined. They may be exaggerating what a user would define as the comet boundary due to many factors. Firstly, the annotations were made by zooming into the images, revealing extended regions of low intensity that are not obvious at normal magnification. This extra intensity is of a much lower value than most of the comet body, and is therefore not detected when applying a threshold, prior to the application of the watershed algorithm.

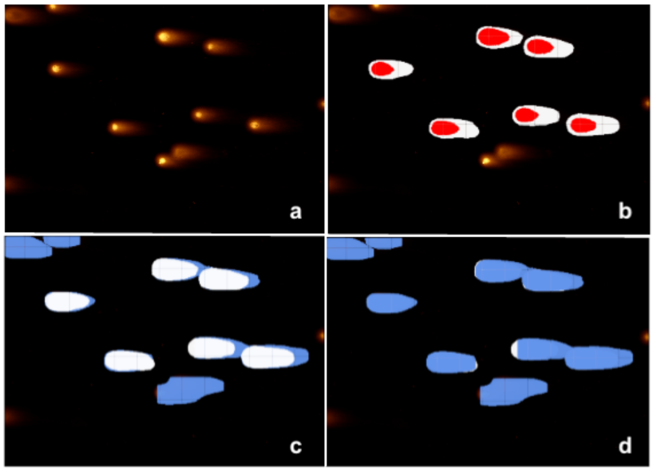


Figure 4.12: *A comparison of the comet masks produced from the watershed, Mask_ RCNN and manual segmentations of image B.* ***a****: raw image,* ***b****: watershed segmented masks laid over manually annotated masks,* ***c****: manually annotated masks laid over Mask R-CNN segmented masks and* ***d****: Mask R-CNN segmented masks laid over manually annotated masks.*

The Mask RCNN segmentations on the whole are a lot more accurate when compared to the ground truth masks, but are inconsistent in the segmentations. This is attributed to the very nature of machine learning and neural networks; we do not know exactly what the model has learned and can only see the results of this learning during inference. Some cases of incorrect segmentation could be related to a potential lack of appropriate training, such as identifying comets that are cut off at the edges or have some overlap. However, this was investigated by training more extensive models with 10 and 20 epochs in place of the original 5. The results showed that the models trained with higher epoch values performed worse on the same dataset, indicating over training and highlighting that a decrease in loss during training does not necessarily equate to better performance on an independent dataset. The identification of overlaps and objects of contamination by the neural network could be overcome by applying the same constraints to the objects as used for the clusters identified from the watershed algorithm.

The number of pixels identified correctly and incorrectly was analysed to perform a quantitative comparison of the different segmentation masks. The watershed/Mask R-CNN masks were treated as predictions and the manually annotated masks were the ground truth. In this way, confusion matrices were calculated for each image, defining the number of pixels identified as true positives (TP), false positives (FP), false negatives (FN) and true negatives (TN), as depicted in Figure 4.13.

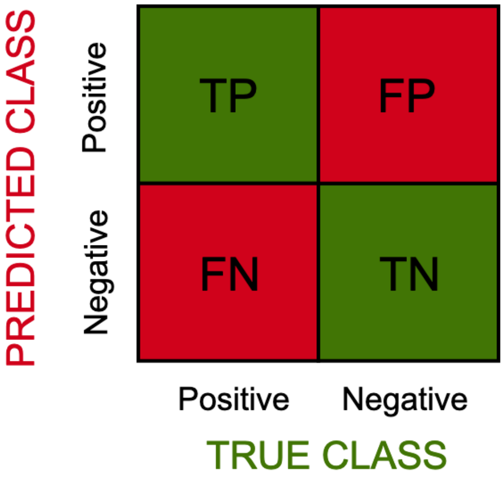


Figure 4.13: *An example confusion matrix showing the categorisation of positive/negative predictions against the true positive and negative classes.*

In this case, positive refers to pixels that are identified as belonging to a comet (and thus saved in the segmentation mask) and negative indicates not belonging to a comet, forming part of the image background. Using the same format as shown in Figure 4.13, confusion matrices were created for the two images analysed here, Figure 4.14 and Figure 4.15.

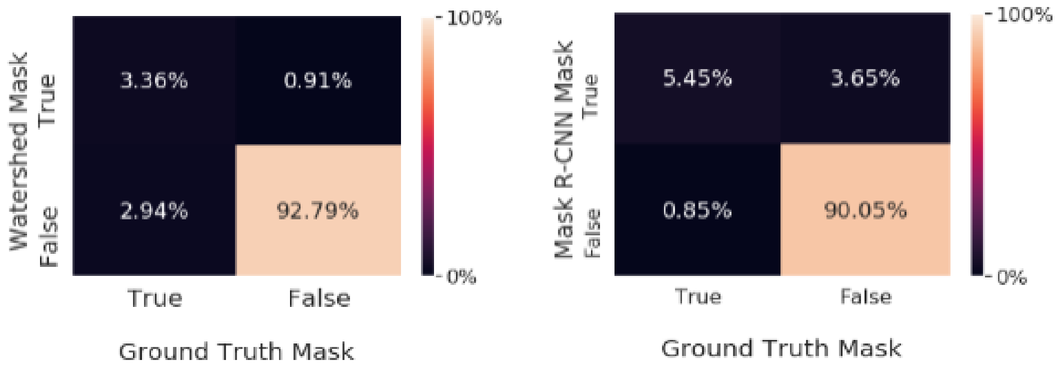


Figure 4.14: *Confusion matrices for the watershed method (left) and Mask R-CNN method (right) applied to image A.*

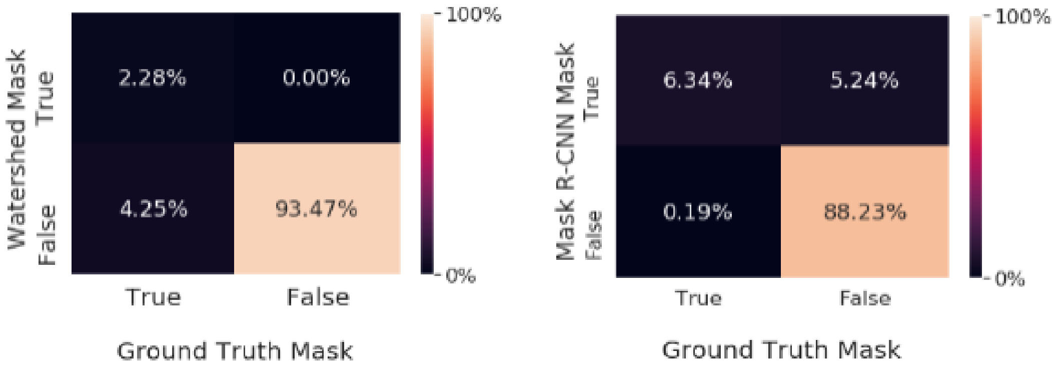


Figure 4.15: *Confusion matrices for the watershed method (left) and Mask R-CNN method (right) applied to image B.*

The very high values of the TN calculated for both images is expected as the majority of the image is composed of background pixels and not the comet class to be identified. The very low values calculated for the FP pixels watershed method illustrate how background pixels are very rarely labelled as belonging to a comet. This is because the watershed algorithm was applied to a slightly over thresholded image, and therefore the mask defined from this method will always be smaller than expected and is completely within the corresponding manual masks (as shown in Figures 4.11 and 4.12).

Average confusion matrices were calculated for the both the watershed and Mask R-CNN methods across all 20 images that were analysed, as shown in Figure 4.16.

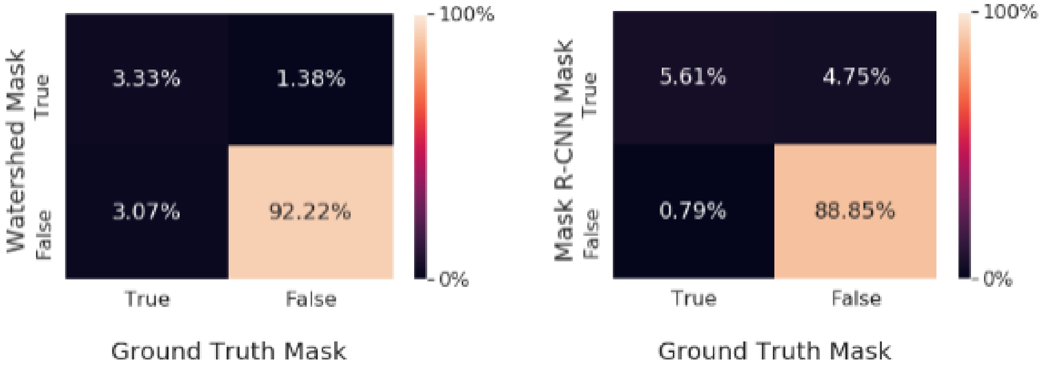


Figure 4.16: *Confusion matrices for the watershed method (left) and Mask R-CNN method (right) averaged for all 20 images.*

When evaluating binary classifications such as these, it is important to consider the makeup of the data. As mentioned, the vast majority of the data is in the negative class (not comet) so there is a large skew. This can be misleading depending on the evaluation metrics used. For example, classification accuracy is calculated as

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}. \tag{4.2}$$

So for image A we calculate the watershed and Mask R-CNN accuracies as 96.16% and 95.50% respectively, and for image B, we find the same values to be 95.75% and 94.57%. At first glance, these values indicate that both models are performing very well. However, as the data is predominantly background, it is heavily biased and we can only conclude that background pixels are accurately classified more than 90% of the time.
Precision is another measure, calculated as

$$Precision = \frac{TP}{TP + FP}. \tag{4.3}$$

This metric indicates how many of the correctly predicted cases were actually positive, removing the skew from the mostly background classification. It also acts as an indicator of the model's reliability. Precision was calculated to be 78.75% and 59.87% for the watershed and Mack R-CNN segmentations of image A and 100% and 54.77% for image B. Here we see that the watershed algorithm is much better at finding pixels that belong to the comet ground truth class than the neural network model.
Recall is used to determine the number of actual positive cases a model can predict, and is

calculated as

$$Recall = \frac{TP}{TP + FN}. \tag{4.4}$$

A summary of all accuracy, precision and recall values calculated for images A and B can be found in Table 4.2 alongside average values for all 20 images.

Segmentation Method and Image(s) Used	Accuracy (%)	Precision (%)	Recall (%)
Watershed Image A	96.2	78.8	53.4
Mask R-CNN Image A	95.5	59.9	86.5
Watershed Image B	95.8	100.0	35.0
Mask R-CNN Image B	94.6	54.8	97.1
Watershed Average	94.7	72.3	47.7
Mask R-CNN Average	94.1	56.6	90.2

Table 4.2: *A summary of the accuracy, precision and recall values calculated for images A and B for the watershed and Mask R-CNN segmentation methods. Average values are calculated from averaging each measurement for all 20 images that were analysed. All values have been rounded to 1 decimal place.*

4.3 Comet Expansions

As the watershed results are more consistent and the methods through which they are produced are better understood, options for improving the segmentation masks produced were explored. One consideration was to expand the identified clusters to overcome the underestimation problem, especially in the tail region. Two expansion methods were explored. The first is a whole comet expansion, in which the entire cluster boundary is extended by some assigned amount. The second is an expansion specifically of the right (tail) pixels to account for lost information in the segmentations. In order to perform either of these expansions, the comet edges need to first be defined. This is implemented using a routine that uses a shifting of all pixels in the left, right up and down direction by 1 pixel, to compare each shifted image with the original comet segmentation masks. Pixels that are not within the original mask once shifted image are defined as the image edge and stored as edge pixels. A visualisation of this methodology is shown in Figure 4.17, where the original shape is shown in orange; its vertices are at coordinates (2,2), (2,4), (5,4) and (5,2).

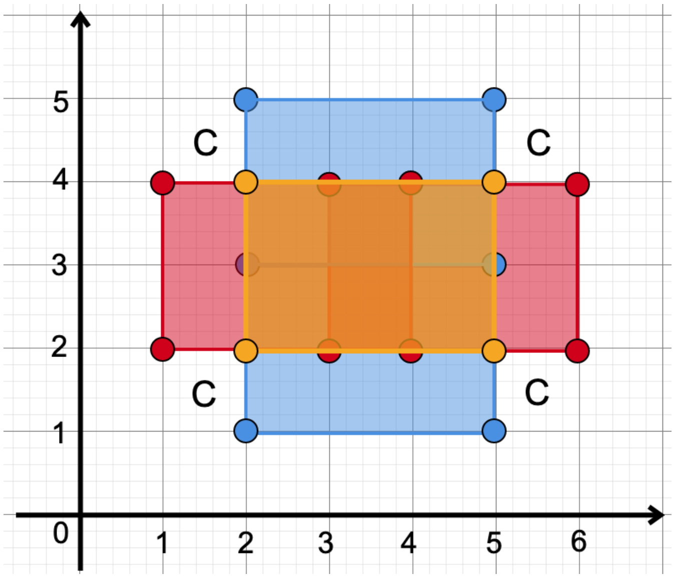


Figure 4.17: *A visual representation of the shift method used to find comet edges.*

The blue rectangle shows the original shifted up and down by 1 unit and the red shows the same for shifting in the right and left positions. We can use the pixel at position (3,3) as an example of a case that is not an edge. If we shift this pixel by 1 in any of the four directions (up down left or right) it will still be within the original orange rectangle, proving it is not part of an edge. The 1 pixel expansion of the orange object can be seen in the solely blue and red regions. The edges can then be used as positions to expand from, creating larger clusters to account for underestimating. An expansion range is required for both the overall and right tail expansions. For an overall expansion, the mask is shifted in all directions by 1 pixel each time, resulting in an expanded cluster that is the union of all shifted pixels. This method is repeated for the required expansion value number of times, but presents an issue. Each time the mask is shifted in the 4 directions, “corner” pixels, indicated by the letter c in Figure 4.17, are not included. These were accounted for by adding an extra right and left shift to the original up and down shifts for each iteration of the expansion process. The right expansion was then applied by shifting pixels to the right by the designated amount and finding the union of the original image and the shifted image.

The expansions required for different comets in different images was not a consistent value for either the whole comet expansion or the additional tail expansion. A compromise value for the most common expansion required was found to be a 10 pixel expansion for the whole comet followed by a further 20 pixel expansion for the tail pixels on the right side of the comet. This caused a significant number of comet segmentations to be overestimates, which would create additional error when comet measurements are performed. However, as the excess expansion is in background regions of the comet assay image, these pixels can be identified and removed prior to measurement, as is further discussed in Section 5.2.6. Another issue with the tail expansion in particular is loss of the tapered tail which is seen for the vast majority of comets. This is inevitable as the expansion is added in a uniform manner across the whole comet boundary. To overcome this, the expansion could be modified to add on fewer pixels the further right it was on the comet. Having said that, the parameters needed to determine the expansion reduction along the comet would vary between comets due to the high variability of not only the expansion required, but the comet shapes themselves, so this is therefore not a viable modification. Similar to the treatment of over-estimating the comet region, the over estimated tail pixels were handled before comet measurement.

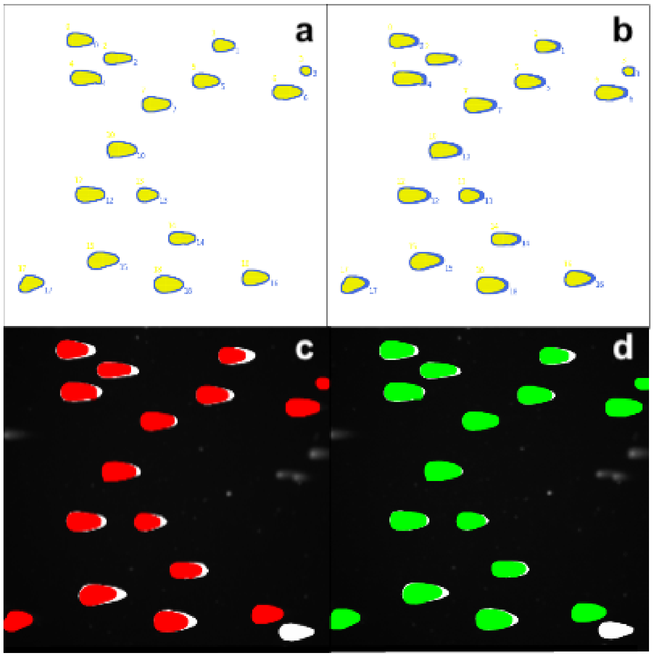


Figure 4.18: *Comparing the different expansions with the manual annotations.* ***a****: the original clusters found from the watershed algorithm (yellow) over expanded comets (blue),* ***b****: the original watershed clusters (yellow) over the overall plus tail expanded comets (blue),* ***c****: only the overall expanded comets (red) over the manual annotations (white) and* ***d****: the overall and tail expanded comets (green) over the manually annotated comets (white).*

Figures 4.18 and 4.19 show the mask variations between clusters identified by the watershed algorithm before and after expansion, and how the expanded clusters compare to the manually annotated masks for two images that have been taken under different parameters and camera equipment.

In Figure 4.18d we see that the expanded clusters plus the additional tail expansion provides a segmentation that is much closer to the ground truth mask. Many images tested in this manner showed similar results. However, a significant amount produced effects closer to those seen in Fig4.19d where the "blockiness" of the expanded clusters is much more apparent. These masks do not describe the ground truth mask as closely, highlighting the variation that is expected of such a broad test dataset of images.

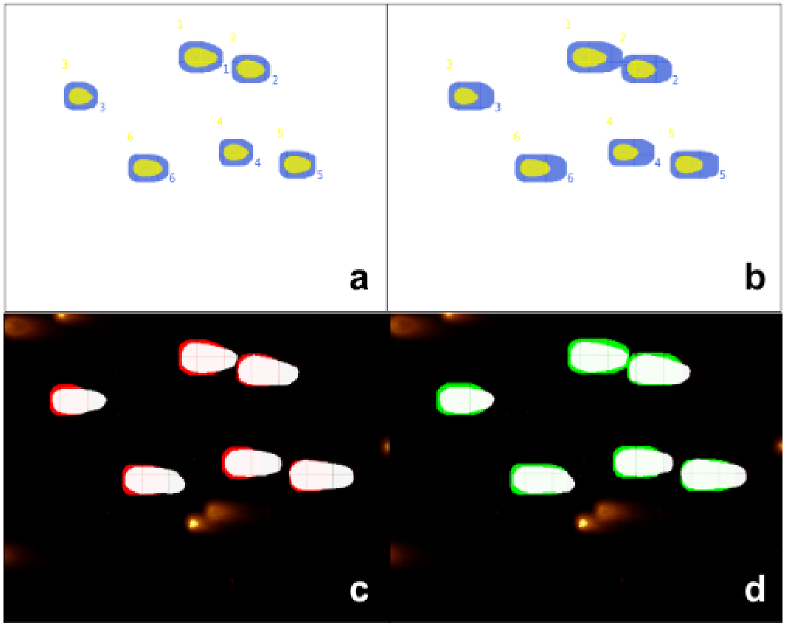


Figure 4.19: *Comparing the different expansions with the manual annotations.* ***a****: the original clusters found from the watershed algorithm (yellow) over expanded comets (blue),* ***b*** *the original watershed clusters (yellow) over the overall plus tail expanded comets (blue),* ***c*** *only the overall expanded comets (red) over the manual annotations (white) and* ***d*** *the overall and tail expanded comets (green) over the manually annotated comets (white).*

As discussed earlier, to quantify the addition the expansion provides, the number of pixels detected per mask was analysed, to produce confusion matrices. These can be found in Appendix B.1, and an overview of results for image A, B and an average of all 20 images is shown in Table 4.3.

We see that the TP values increase when the overall cluster expansion is applied, and increase again for the additional tail expansion. This is expected as the original clusters were underestimates, due to the higher threshold used, prior to applying the watershed algorithm. Unfortunately, the FP (top right) values also increase as the expansions are applied, indicating more pixels are identified as belonging to a comet by the expansions which are not part of the ground truth masks. Again, this was a predicted outcome as the expansions are based on an average image of average comets, and are not designed to account for all types of variability. If these FP pixels are obviously part of the image background, they will be dealt with accordingly, prior to comet measurement.

Watershed Method and Image(s) Used	TP (%)	FP (%)	FN (%)	TN (%)
Watershed Image A	3.4	0.9	2.9	92.8
Watershed Ex Image A	4.9	1.5	1.4	92.2
Watershed Rx Image A	5.4	1.8	0.9	91.9
Watershed Image B	2.3	0.0	4.2	93.5
Watershed Ex Image B	4.8	1.1	1.7	92.4
Watershed Rx Image B	6.0	1.7	0.5	91.8
Watershed Average	3.3	1.4	3.1	92.2
Watershed Ex Average	4.6	2.3	1.7	91.4
Watershed Rx Average	5.1	2.8	1.2	90.9

Table 4.3: *A summary of the TP, FP, FN and TN pixels found using the watershed masks for image A, B and an average across all 20 images. Watershed Ex and watershed Rx correspond to the whole expansion and whole plus tail expanded masks, respectively. All values have been rounded to 1 decimal place.*

Accuracy, precision and recall values were calculated for the expanded masks, for both the overall expansion only case and the additional tail expansion. An overview of all these values can be seen in Table 4.4. We see increases in both accuracy and recall for images A and B, indicating that the expanded watershed algorithm better describes the ground truth masks, and the tail expansion improves on this better approximation. However, the average values for accuracy do not reflect this. As previously discussed, this is due to the 10 pixel overall expansion and additional 20 pixel tail expansion not being optimised values for some images, causing over/under expansion. Due to the decreasing values for accuracy but not recall for the average image sets, this is attributed to an increase in FP pixels, highlighted in the confusion matrices, which is higher for the image set as a whole than for image A or

B individually. This is further emphasised when considering the precision values. As this is the proportion of TP pixels compared to TP + FP, an increase in FP pixels will reduce the precision calculated, which is what we see for all images as the expansions are applied.

The recall measure is particularly important in some cases such as medical classifications. It can be extremely dangerous to produce FN classifications, where conditions could be missed with serious consequences. FP classifications on the other hand can be distressing, whilst consuming a lot of time and resources, but are less detrimental to patient outcomes. In our case of classifying pixels in comet assay data, it is also more favourable to have more FP classifications than FN, as we do not want to miss actual comet pixels that contribute to measurements. FP pixels will not skew the results a significant amount, as long as there are only a small number of them, which is the case when comets have been expanded overall with the additional tail expansion. The effect these have on comet measurement is discussed in Section 5.2.3.

Segmentation Method and Image(s) Used	Accuracy (%)	Precision (%)	Recall (%)
Watershed Image A	96.2	78.8	53.4
Watershed Ex Image A	97.1	76.1	77.6
Watershed Rx Image A	97.3	74.5	85.6
Watershed Image B	95.8	100.0	35.0
Watershed Ex Image B	97.2	81.4	73.8
Watershed Rx Image B	97.8	78.3	92.0
Watershed Average	96.7	72.3	47.7
Watershed Ex Average	95.3	68.2	73.0
Watershed Rx	95.4	66.1	82.7

Table 4.4: *A summary of the accuracy, precision and recall values calculated for images A and B for the original masks found from the watershed algorithm and the expanded masks. Watershed Ex and watershed Rx correspond to the whole expansion and whole plus tail expanded masks, respectively. Average values are calculated from averaging each measurement for all 20 images that were analysed. All values have been rounded to 1 decimal place.*

4.4 Identification issues due to inconsistent data capture

The images analysed in this chapter have been specifically selected as they have preferable characteristics in regard to their exposure and contrast. This was required as a significant amount of the image data is not “ideal”, due to poor exposure, forced contrast ranges and additional background anomalies, that will be explored here to understand the effect they have on comet segmentation.

4.4.1 Exposure

Exposure and contrast have the biggest and most direct impact on determining damage, as mentioned in Section 2.4.2, due to the reliance on pixel intensity. Overexposure is much worse than under exposure as intensity data is lost and cannot be recovered. Figure 4.20 shows an example histogram of an overexposed image, where a spike can be seen in the final bin.

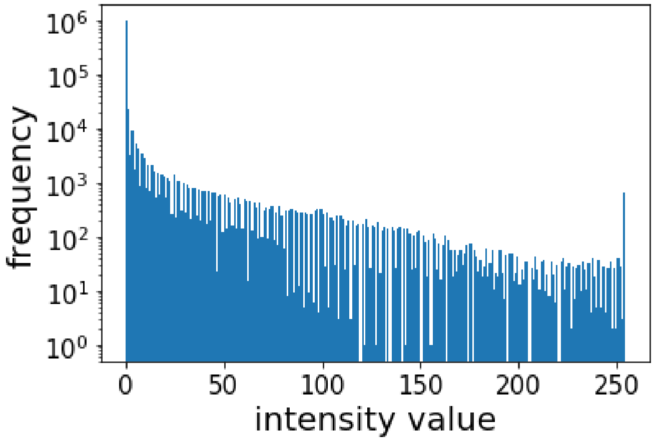


Figure 4.20: *The raw intensity histogram of an overexposed image. Frequency has been plotted on a logarithmic scale.*

This conveys that a too narrow contrast range was chosen during imaging, or too long an exposure time was used. Overexposure may appear to be a desired characteristic at first, as it better defines the comet head by representing it at a higher intensity. The comets are therefore much brighter against the image background, but this is at the cost of lost intensity data and needs to be avoided. This overexposure could be applied for internal analysis, comparing image sets where the same overexposure has been applied. However, unless these parameters are recorded and kept consistent even these internal comparisons are not accurate. Furthermore, this is bad scientific practice as it is not reproducible (from imaging the same slides) and can be misleading in published results. If fluorescence is to be used as a direct measure of the DNA present at a point, it needs to be accurately captured through calibrations and standardisation of camera equipment settings.

To test the effect varying exposure has on comet identification, slides were imaged at the same position whilst the exposure time was varied for each image captured. Areas on the slide containing bright contamination were avoided, so as to not skew the intensity scaling. This is because if a very bright contaminant is present in the image, its pixels will be assigned the top intensities and the rest of the image scaled according to their relative brightness. This causes the intensity of the comet, especially the head, to not appear as bright as it actually is and introduces error into any measurements that are made, and further stresses the need for avoiding contamination of the agarose during Comet Assay protocols. By purposefully avoiding such areas of contamination on a slide, bias was introduced into the imaging procedure, and is not good practice when conducting Comet Assay experiments. It was done in this case to avoid the scaling and allowed for the purpose of investigating exposure and contrast effects on imaging outcomes.

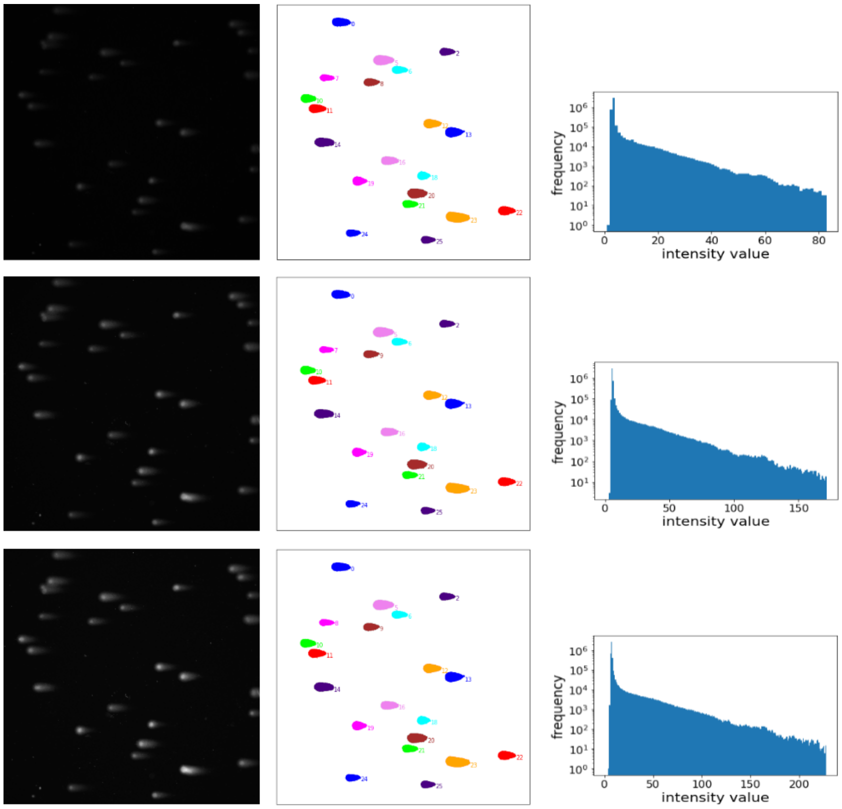


Figure 4.21: ***left****: Raw image,* ***middle****: watershed segmentations and* ***right****: the image histogram for images taken at exposure values 3 000, 7 000 and 10 000.*

As previously mentioned, exposure has units of lux seconds and measures the incident light over some time frame. The light itself does not change, unless imaging is carried out over long periods in which the stain can fade, so the variable parameter here is the exposure time. The micromanager software allows the user to vary the exposure used to take the comet images. The units of which relate to exposure time by a factor of 16. For example, setting the exposure variable to 10 000 (the maximum value possible) equated to an exposure time of 625 ms. A range of images were taken at 10 different exposure values ranging from 1 000 to 10 000, keeping the slide position constant. The raw image at some of these exposure values and the resulting watershed segmentations alongside the image histograms are shown in Figure 4.21.

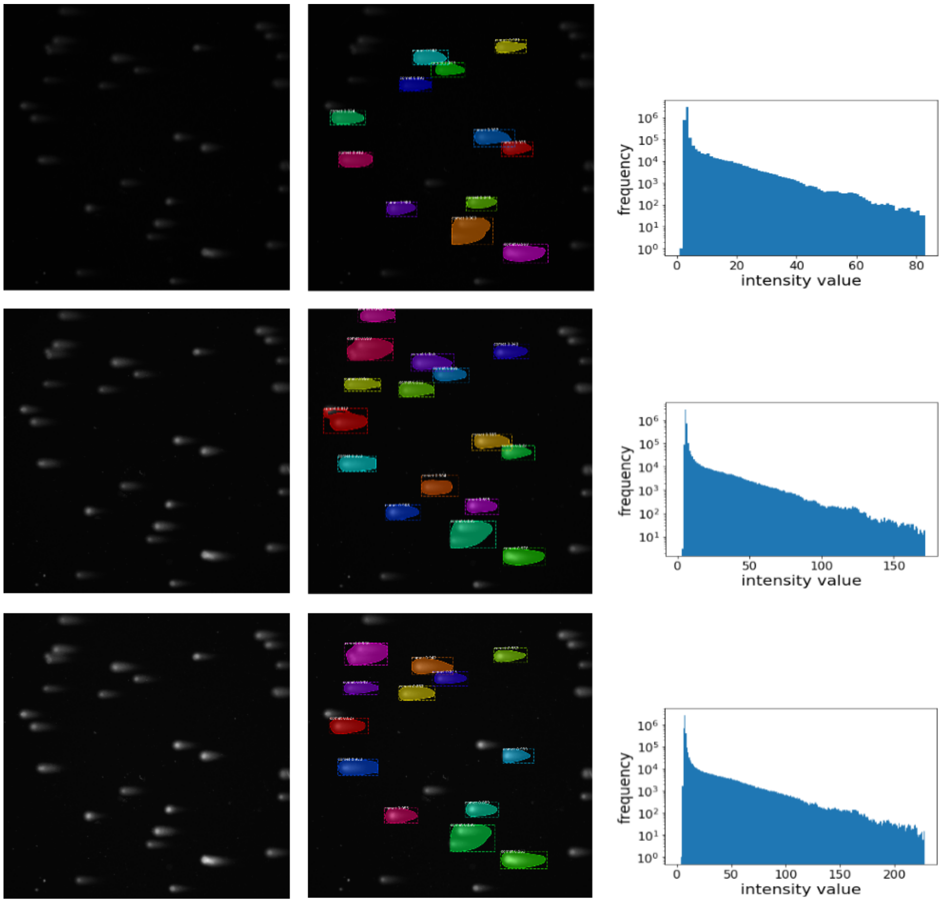


Figure 4.22: ***left***: *Raw image,* ***middle***: *Mask R-CNN segmentations and* ***right***: *the image histogram for images taken at exposure values 3 000, 7 000 and 10 000.*

As the threshold for identifying comets is set by a proportion of the maximum pixel value (10%) a negligible amount of difference was seen in comet segmentation at different exposures. Even when comparing the results between images taken at a 1 000 and 10 000 exposure value, only a slight increase in mask segmentation was seen, which would have very little effect in terms of comet identification. The effect exposure has on comet measurement is explored in Chapter 5. We do however see a difference in results when applying inference from the Mask R-CNN model. Figure 4.22 shows the mask predictions following inference from the trained Mask R-CNN model applied to images taken at varying exposures.

From analysing multiple image sets and as evidenced in Figure 4.22 the number of comets segmented increases up until around exposure values of 6 000 and 7 000 before declining at higher values where images are overexposed. This is expected as although the network was trained on a range of images of varying exposure, most were not at the extremes studied here, so the model has not learned to identify comets within these images.

4.4.2 Contrast

Contrast changes alter the top and bottom intensity values, defining the dynamic range used to describe an image, which are in turn used to normalise the acquired image values from 0 to the intensity range. The effect of contrast was explored, to see how it impacted comet identification. The micromanager software allows a user to set the contrast manually, use the full range defined by the bit-value selected or select an automatic function. The latter selects the top and bottom intensity values based on the minimum and maximum intensities currently detected by the live camera-microscope equipment. A set of images were captured in a similar manner as for the exposure analysis. The slide position and exposure was kept constant whilst multiple images were taken with varied contrast. This was first selected by using the automatic function within micromanager and then by using the full range of the 16-bit channel (0-65535). As for the varied exposure data, the watershed algorithm showed very little to no difference in the comet segmentations produced. This is not the case for the Mask R-CNN segmentations. The contrast variation plays a large role in how comets are detected, in a non-linear manner. For example, Figures 4.23, 4.24 and 4.25 show a comparison of the same cells imaged with the auto or full contrast range alongside their histograms.

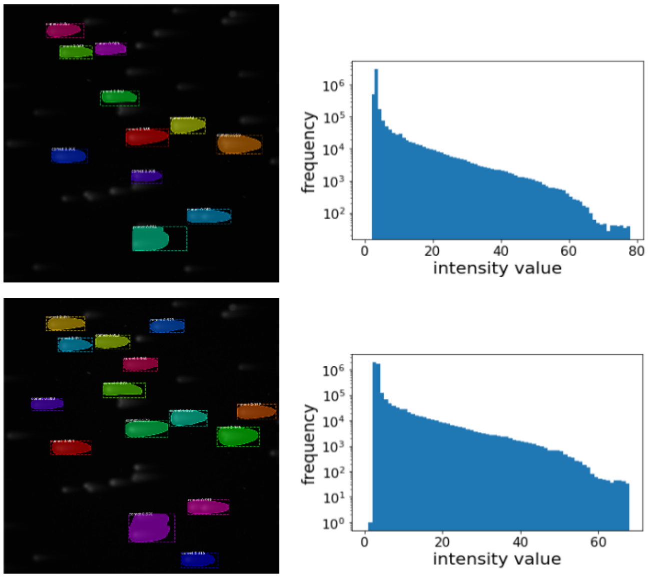


Figure 4.23: *Mask R-CNN segmented images alongside the image histogram. The top figures correspond to images taken using the auto contrast function whilst the bottom figures are from using the full contrast range at exposure value 3 000.*

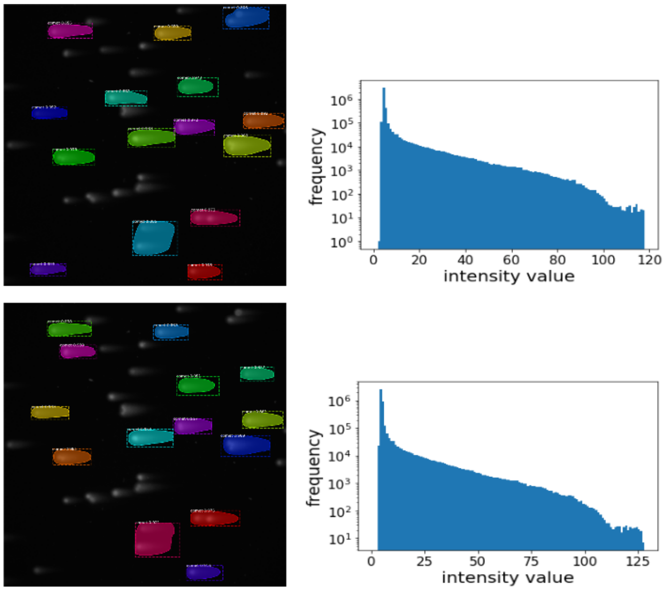


Figure 4.24: *Mask R-CNN segmented images alongside the image histogram. The top figures correspond to images taken using the auto contrast function whilst the bottom figures are from using the full contrast range at exposure value 5 000.*

A minimum detection threshold was set at 80% so only masks with this prediction confidence or higher were displayed. The first thing to note is that the masks detected changed when the auto contrast function was used as opposed to the full contrast range. This was not exclusive to certain exposure values as it was seen in all cases investigated. The prediction mask confidence also varied between the two contrast modes explored, with no consistent pattern seen for which produced a higher confidence mask segmentation. This further reiterates the point that we do not fully understand what neural network models learn and need to rigorously test them on a very wide set of data before using them in scientific analysis or safety-critical applications.

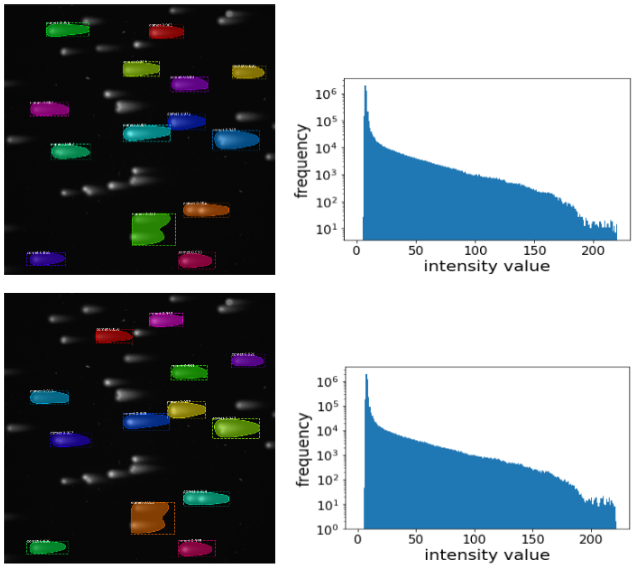


Figure 4.25: *Mask R-CNN segmented images alongside the image histogram. The top figures correspond to images taken using the auto contrast function whilst the bottom figures are taken using the full contrast range at exposure value 9 000.*

4.4.3 Background Vignette

When thresholding some images at very low values (<10% of the maximum intensity) a background vignette effect became apparent. Initially, this was thought to be due to some alignment issue of the microscope causing a shadow during imaging. However, further analysis showed that the central area is actually brighter and it is the image corners that are darker. It was ruled out that this was due to the microscope alignment causing a shadow as the location of brighter and darker regions was inconsistent across different images, as depicted in Figure 4.26.

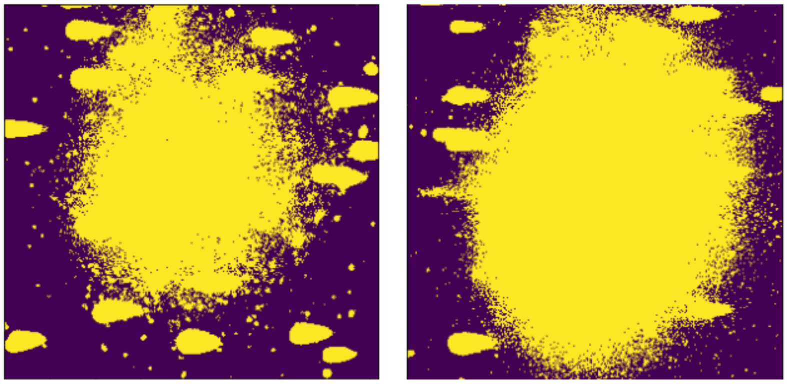


Figure 4.26: *A low threshold of 5% the maximum pixel value applied to two images to reveal the vignette effect causing darkened edges across some images.*

As this effect is only seen at very low thresholds, below what was used for comet identification, it was not an issue for anything discussed so far. Further background analysis is outlined and discussed in Chapter 5.

4.5 Conclusions

Global thresholds work best for Comet Assay data, as the high variability of the image background is not appropriate for the use of local thresholds. The watershed algorithm and Mask R-CNN image segmentation methods offer different advantages for the application of identifying comet objects in Comet Assay images. The watershed algorithm is much more robust, producing similarly shaped comets, once area, width and height constraints had been introduced. A compromise of a higher threshold was applied prior to the watershed algorithm in order to maximise comet identification, but this came at a cost of underestimating comet regions. By applying whole comet and an additional tail expansion, this under estimation was significantly reduced for a majority of cases. This method does however still identify a significant number of objects that are not good comets i.e. bodies of contamination or overlapping comets. The Mask R-CNN method is not as robust in identifying comets; its underpinning knowledge is unknown due to the black-box nature of neural networks, which is seen through the inconsistent segmentations it performs. On the whole, it does a much better job at ignoring contamination but it still identifies instances of comet overlap as belonging to the target comet class, as also seen in the watershed segmentations. This highlights the need for human input, even in automated methods, and is further discussed when comet measurement is explored in Chapter 5.

Results show that image exposure and contrast have negligible effect on watershed segmentations, attributed to using a proportion of the image's maximum pixel value as a threshold, rather than a set, rigid value. The Mask R-CNN model shows much variability in results when exposure and contrast are changed, but not in any linear or identifiable manner. Once again, this is attributed to the black-box nature of neural network learning that we do not entirely understand. The Mask R-CNN model should be tested on a much larger data set than was available here, in an attempt to further understand the relationships and patterns it has learned.

Chapter 5

Measuring the Detected Comets

5.1 Defining the Comet Head

Once comets have been segmented, measurements can be performed upon them. In order to implement these measurements, the pixels belonging to the comet head and tail need to be distinguished from one another. A comet's composition is dependent on many factors; the DNA damage the cell has undergone, whether the Comet Assay is of the alkaline (revealing single and double strand breaks) or neutral (revealing only double strand breaks) variety, agarose concentration, electrophoresis and DNA unwinding times. All data analysed in this work was produced from neutral Comet Assay experiments, following very similar protocols. To define the comet head, an assumption that the head contains the brightest pixels has been made. This is a valid assumption as it is only in cases of very high damage that the tail can appear brighter than the comet head, and this is usually seen in alkaline Comet Assay data. The data analysed here has been produced from cells irradiated at 4 or 8 Gy and therefore does not elicit a high enough level of damage to produce exceptionally bright comet tails.
As in Chapter 4, the two images, referred to as image A and B are used to demonstrate the effects of measurement in various contexts. Image A is an example of a correctly exposed image, whilst B is overexposed. These images were also taken using different imaging equipment, and therefore have different resolutions.

By assuming the head is the brightest part of the comet, the average position of pixels with the maximum intensity (within that comet) is found and taken to be the head centre, with co-ordinates (X_{cent}, Y_{cent}). The radius of the head is then chosen from the maximum between the distance from the head centre to the furthest left pixel of the comet, the furthest north pixel and the furthest south. All pixels that are contained with the circle of centre, (X_{cent}, Y_{cent}) and radius, r are assigned as head pixels, while the remainder belong to the comet tail. A depiction of the defining the comet head is shown in Figure 5.1. The largest radius of the three measured is chosen to ensure that the vast majority of head pixels are captured by this method. This does however sometimes cause an overestimate of pixels, as shown in Figure 5.1b, where the circle reaches beyond the top of the comet. This is a negligible number of pixels on the whole and for cases where it is not, the pixels will either need to be accounted for due to underestimates from comet identification or are background pixels and will be removed from measurement during background subtraction, as discussed in Section 5.2.6. Once head and tail pixels have been separated, comet measurements can be made.

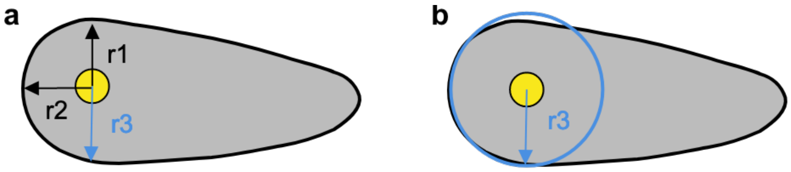


Figure 5.1: *The average position of the brightest pixels is indicated by the yellow circle (X_{cent}, Y_{cent}), **a**: r1, r2 and r3 are the radius lengths measured in the up, left and down positions from the head centre respectively. **b**: the longest radius, r3, is chosen and a circle of this radius with centre (X_{cent}, Y_{cent}) is used to define the comet head pixels.*

5.2 Intensity Measure

The pixel intensity is a proportional measure of the amount of DNA present at the location recorded by that pixel. By measuring the fluorescence (DNA) in the tail, we get an indicator of the migration that has occurred and thus the number of DNA strand breaks present. As a reminder, the common DNA measures used, mentioned in Section 2.3, are tail length, tail DNA and tail moment. Tail length is a pixel measure of the length of tail pixels, along the axis of the electric field, and tail DNA is the intensity proportion of DNA in the tail. Tail moment incorporates both these measures by multiplying tail length and tail DNA. Only tail length and tail DNA measures will be discussed in this chapter, as tail moment does not provide any further information on comet properties.

Comet measurements were performed following segmentations from the watershed algorithm, Mask R-CNN model and manually annotated comet masks. For comparison, comet measurements were also made using the widely used open source software OpenComet [120]. This software identifies comets by performing an adaptive threshold using Huang's method [158]. Comets are then removed if they are small objects, touch the image edge or overlap. The remaining regions are analysed to filter out irregular objects, and this analysis is based on two assumptions. The first is that comets have a convex shape, and the second is that comets are symmetrical. To check these assumptions for each remaining object, the convexity ratio and the displacement between the centre line from the front of the comet and the centre line from the back of the comet are calculated and discarded or kept according to the defined parameters.

The comet head is defined by one of two methods. The first assumes that the head contains the brightest pixels and calculates the average position of pixels with the highest intensity within that comet. This is taken to be the head centre from which a circle with radius equal to the length between the head centre and the comet's leftmost pixel. Pixels within this circle are then assigned as head pixels, whilst the rest belong to the tail. A second method is required for heavily damaged cells which may have brighter tails. If the first method fails, the comet's intensity profile along the horizontal axis is analysed. Following a convolutional smoothing of the intensity profile, the first maximum of the second differential of the profile is assigned as the boundary between head and tail pixels. The user is able to select either of these methods to define the comet head or choose an automatic function that tries the bright head method first and, in the case of its failure implements the profile analysis approach.

OpenComet produces a spread sheet of measurements and an accompanying annotated image for each image analysed, such as that shown in Figure 5.2. Prior to this, the user is able to review the annotations, where identified objects have been categorised in three ways; valid comet, outlier comet and invalid shape. The outlier comet label is usually assigned to instances where the comet detected may not be regular and the user is allowed to accept or reject all the identified measurements, choosing to remove any from the results spread sheet. An updated spread sheet is then produced if any of the measurements have been removed. Analysis via OpenComet was performed for all images analysed here, removing all incorrectly defined objects.

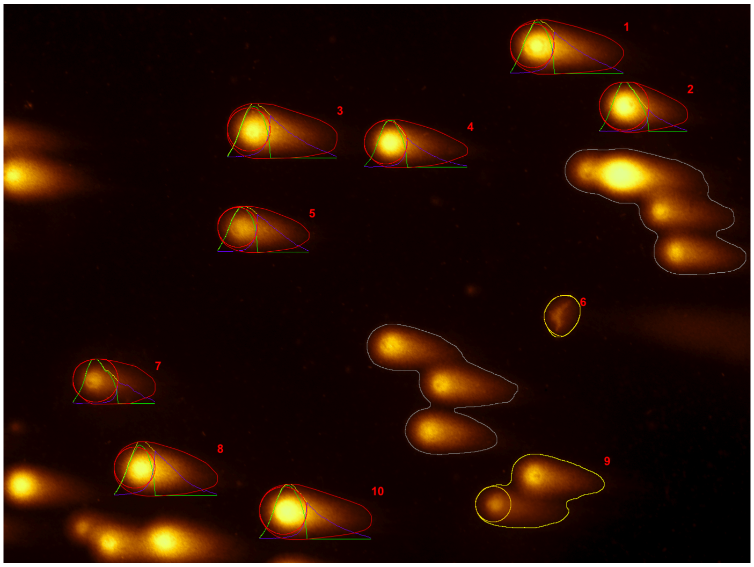


Figure 5.2: *An example annotated image produced from OpenComet. Objects outlined in red have been identified as valid comets, those in yellow are outliers and grey outlines indicate invalid shapes.*

5.2.1 Tail Length

Comparison plots of comet tail length produced from the watershed identified comets, Mask R-CNN comets and OpenComet for images A and B are shown in Figure 5.3 and Figure 5.4, respectively. Only comets that were identified by at least two of the methods are plotted, and a measure of 0 indicates that the comet was not identified for that method.

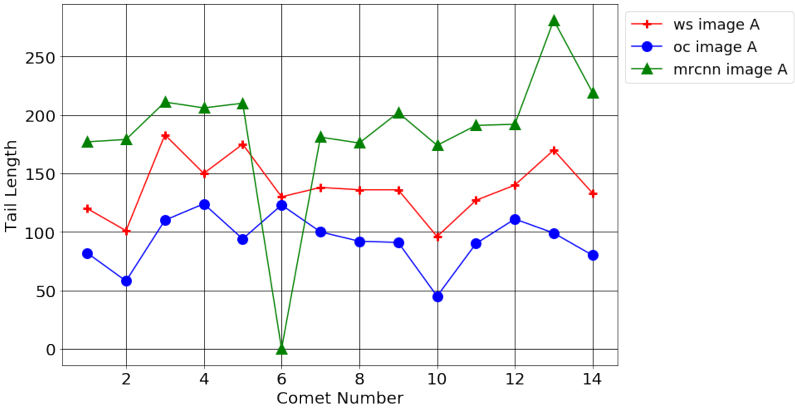


Figure 5.3: *Comparison of comet tail length measures found in image A via the watershed (ws), OpenComet (oc) and Mask R-CNN (mrcnn) methods.*

There is a clear discrepancy in the measurements performed, although we see a clearer trend across the comets for each method in image A. This may be attributed to the overexposure of image B, and how the different methods handle such an image. The Mask R-CNN model performs the worst of the three methods, as it was unable to identify a comet in both images, that the other two methods were able to find.

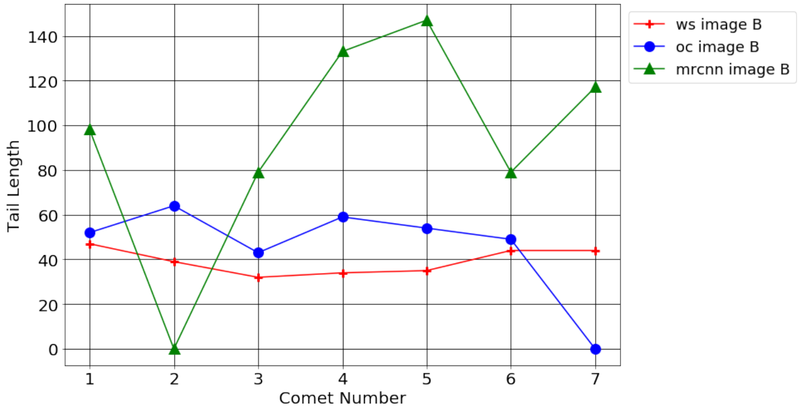


Figure 5.4: *Comparison of comet tail length measures found in image B via the watershed (ws), OpenComet (oc) and Mask R-CNN (mrcnn) methods.*

5.2.2 Tail DNA

A second measure of tail DNA was performed on each identified comet, to explore how the proportion of head and tail pixel intensities is measured by each method. This is shown in Figure 5.5 and Figure 5.6 for images A and B, respectively.

Here, a much closer agreement is seen for the watershed and Mask R-CNN methods, due to the comet head being defined in the same manner for both. The OpenComet measures are generally larger, much larger in some cases, due to the different head defining methods used. Note that, when the auto comet measure function is selected, OpenComet does not inform the user which of the two head defining methods was employed. It is suspected that for cases where the tail DNA is close to the values found using the watershed method, the bright head method has been utilised, whereas the measurements that are further away use the profile analysis method. Again, there is a much larger discrepancy in the tail DNA measured in the overexposed image B, especially for some OpenComet cases, where it suspected that the intensity profile method has been used to define the comet head.

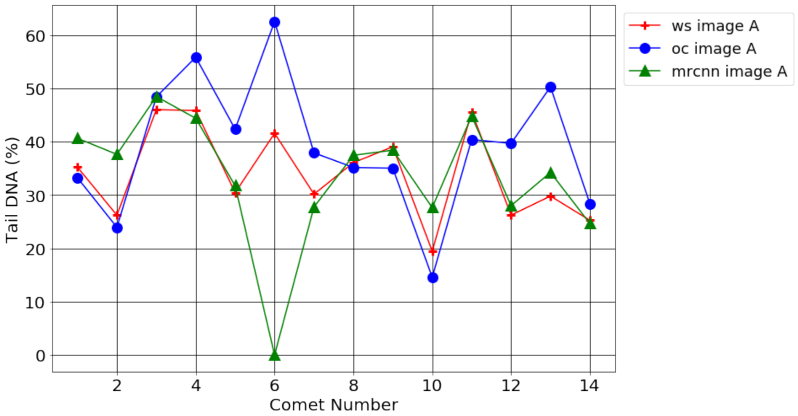


Figure 5.5: *Comparison of comet tail length measures found in image A via the watershed (ws), OpenComet (oc) and Mask R-CNN (mrcnn) methods.*

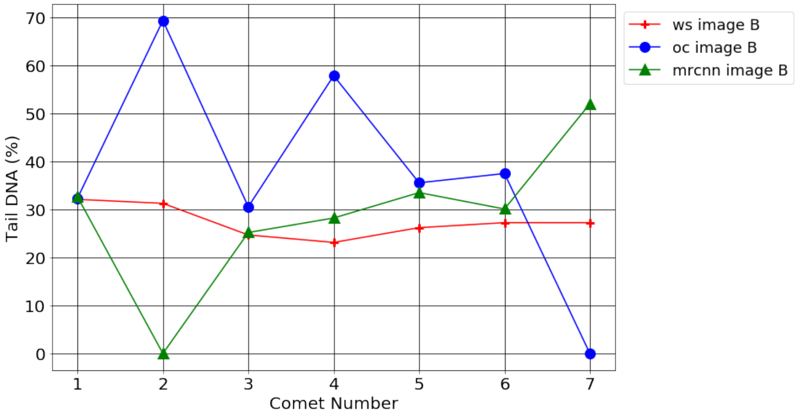


Figure 5.6: *Comparison of comet tail length measures found in image B via the watershed (ws), OpenComet (oc) and Mask R-CNN (mrcnn) methods.*

5.2.3 Effect of Expansion on Comet Measurement

In Section 4.3, comet expansion, particularly whole comet expansion plus an additional tail expansion, was shown to better approximate comet regions, but how do these expansions affect comet measurement? Analysis was performed to compare comet measurements made on the original comets identified by the watershed algorithm and their expanded comets with the equivalent OpenComet measurements.

A check was performed to ensure the expanded clusters did not now touch the image edge or overlap with any other comets in close proximity, and if so, they were removed from the accepted clusters to be measured. Figure 5.7 and Figure 5.8 show the tail length and tail DNA comparisons for images A and B respectively.

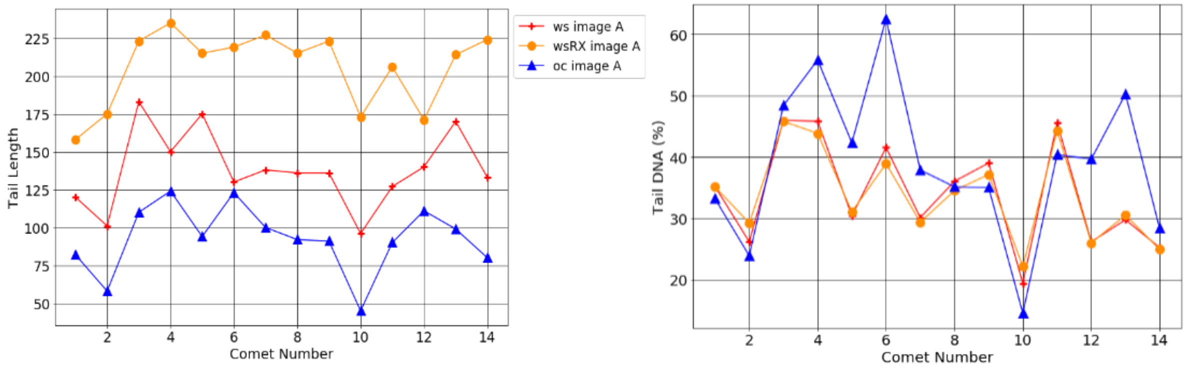


Figure 5.7: *Comparison of comet tail length and DNA found in image A via the watershed (ws), expanded watershed (wsRx) and OpenComet (oc) methods.*

As expected, we see a higher tail length for the expanded watershed clusters, however the increase is not always consistent. There are cases where an expanded cluster's tail length actually increases from the previous measurement, where it originally decreased for the original watershed clusters, such as comets 2, 4 and 6 in Figure 5.7. This indicates that the head centre and resulting comet head may have been defined differently for some expanded clusters.

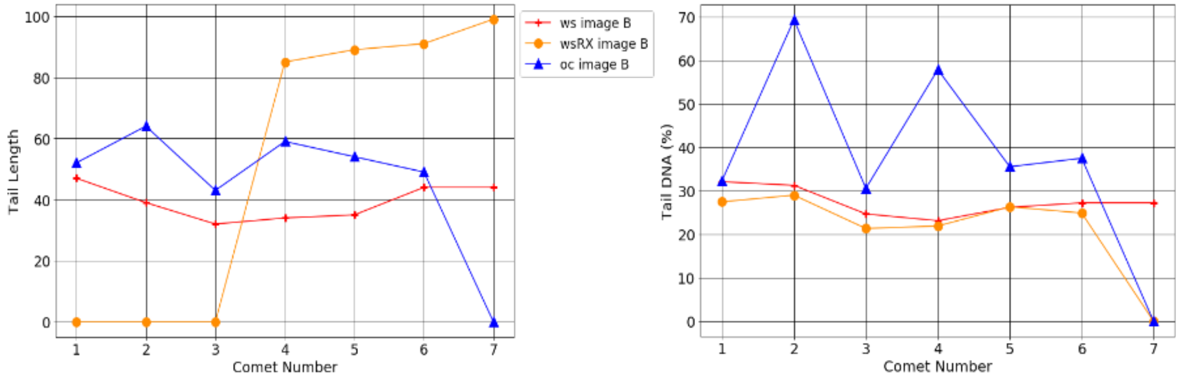


Figure 5.8: *Comparison of comet tail length and DNA found in image A via the watershed (ws), expanded watershed (wsRx) and OpenComet (oc) methods.*

However, upon inspection of the tail DNA measurements for image A, we see that the measures are much more in agreement. This tells us that the proportion of head and tail pixel intensities remains approximately the same for the original and expanded clusters, which is unexpected as a tail expansion has been applied in addition to a whole comet expansion, which should produce an increased tail DNA measure. A possible reason for this is that the additional tail pixels found by the comet expansions are of such low intensity that although they contribute to an increased tail length measure, their addition to tail DNA is minimal. This further stresses the need for reporting both measures (tail length and tail DNA), as a single measure can be misleading. As seen previously, OpenComet measurements for both tail length and DNA are approximately in agreement with the original watershed clusters for some comets and very different for others, attributed to the different methods used to determine the comet head.

5.2.4 Effect of Exposure on Comet Measurement

Changes in the exposure and contrast of an image have very little effect on comet segmentations when applying the watershed algorithm, but they do affect the Mask R-CNN segmentations, as shown in Sections 4.4.1 and 4.4.2. As both exposure and contrast have a direct relationship on an image's intensity spectrum, they are expected to also have a large effect on comet measurement.

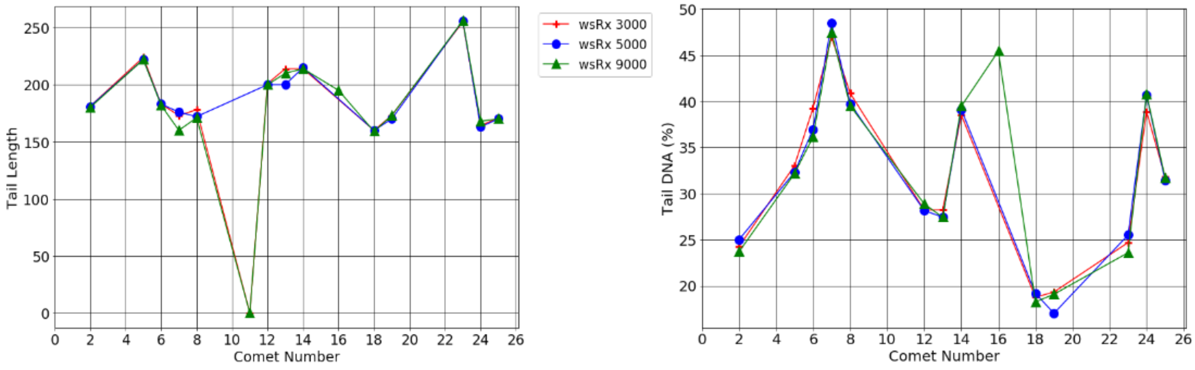


Figure 5.9: *Tail length and DNA measures for images taken at exposures 3 000, 5 000 and 9 000 where comets have been identified using the expanded watershed algorithm.*

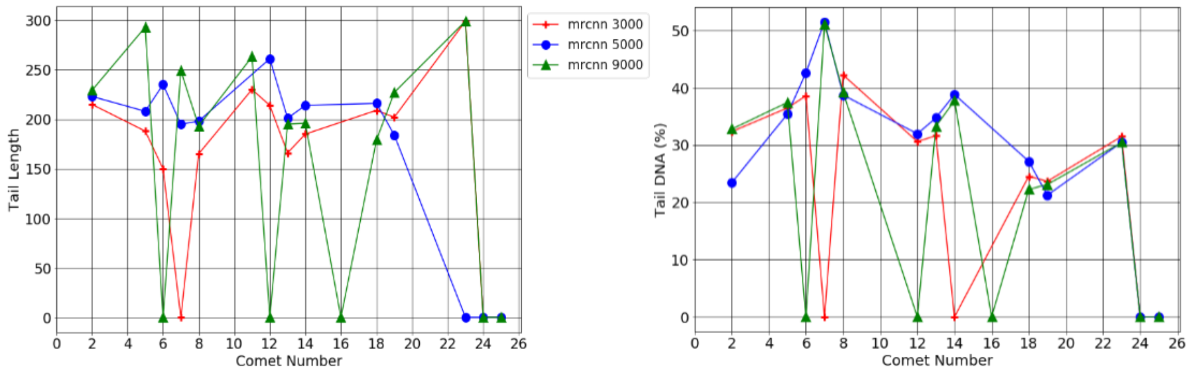


Figure 5.10: *Tail length and DNA measures for images taken at exposures 3 000, 5 000 and 9 000 where comets have been identified using the Mask R-CNN model.*

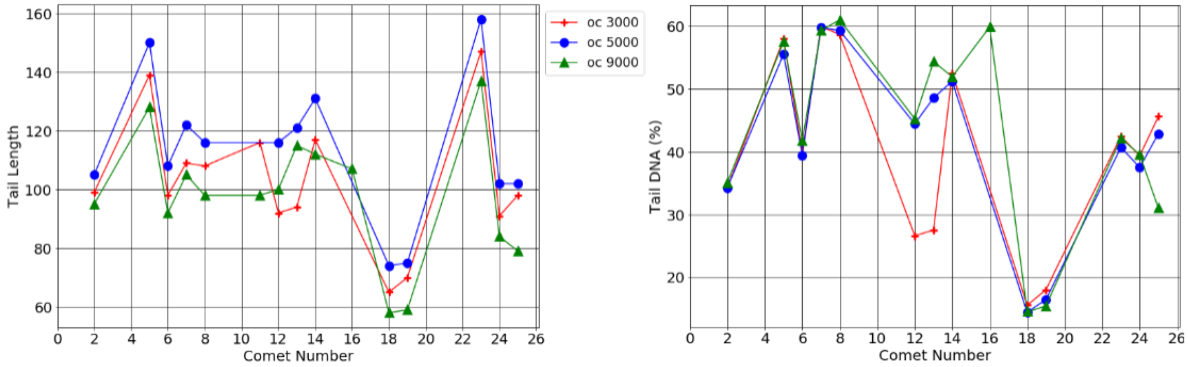


Figure 5.11: *Tail length and DNA measures for images taken at exposures 3 000, 5 000 and 9 000 where comets have been identified using OpenComet.*

Results from analysing comet measurements on an image set with varying exposures are shown in Figure 5.9, Figure 5.10 and Fig5.11, for comets identified via the expanded watershed method, Mask R-CNN model and OpenComet software, respectively.

As shown earlier in Section 4.4.1, the watershed algorithm shows negligible difference in comet segmentation as exposure is increased, from which follows that it also has negligible effect on comet measurement as shown in Figure 5.9 for both tail length and DNA. The outlier measurements at exposure 9 000 are cases where a comet was not identified or was only identifiable at this high exposure level. Otherwise, the measurements taken at 3 000 and 5 000 exposure values were approximately in agreement for all detected comets.

The Mask R-CNN and OpenComet measurements have more variability as exposure is changed. From Figure 5.10 it can be seen that the Mask R-CNN model struggles to find comets at the high exposure of 9000 for a few instances, and once for the low 3 000 exposure. This is similar to what was seen for comet identifications, where there was no distinguishable pattern for when a comet was or was not identified. When comets are found, they have some variation in tail length, but this is minimal when considering comet dimensions. As outlined in Section 4.1.3, comet widths are approximately 30-300 pixels. So a variation of 20 pixels due to exposure for a comet with a width of 200 pixels, assuming half of the comet is tail, can cause a 20% change in tail length recorded, which is significant when reporting results. This then translates into smaller variations in tail DNA as although the tail length may have increased, the proportion of intensity in the tail has not, stressing the need to report both measurements. A similar result can be seen for OpenComet, showing the need to calibrate camera equipment to ensure the appropriate exposure is used for imaging Comet Assay slides.

5.2.5 Effect of Contrast on Comet Measurement

Contrast is varied by altering the maximum and minimum intensity values. Two sets of images were taken, one using an auto function that automatically defines the minimum and maximum intensity values by analysing the images histogram prior to image capture and the second using the full intensity spectrum. Tail length varied little when comparing the auto and full contrast images, at exposure values 3 000, 5 000 and 9 000 and has been omitted here. More detail for these measurements can be found in Appendix B.2.

Tail DNA varied considerably more with contrast, as seen in Figure 5.12, Figure 5.13 and Figure 5.14. At all three exposure values, we see a larger discrepancy in tail DNA for the auto contrast images. The Mask R-CNN model in particular identifies fewer comets for the auto images than the full contrast ones. By narrowing the range of intensities used to describe an image, it is more difficult for the model to recognise comet objects, as it has not been trained on a significant number of images of this contrast type. OpenComet measurements also vary, by more than 10% in some cases. This may be attributed to the different head defining methods it implements, affecting comet analysis. However, there are variations in the image set in which the higher tail DNA measure is found, highlighting that varying contrast affects comet measurement in a non-linear way.

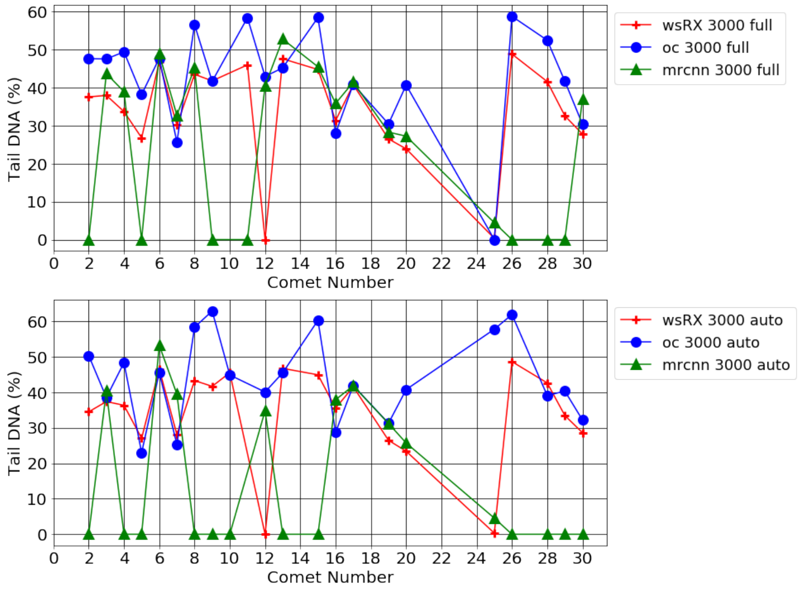


Figure 5.12: *Tail DNA measures for comets detected using the expanded watershed algorithm (wsRX), OpenComet (oc) and the Mask R-CNN (mrcnn) model using the auto and full contrast functions at exposure value 3000.*

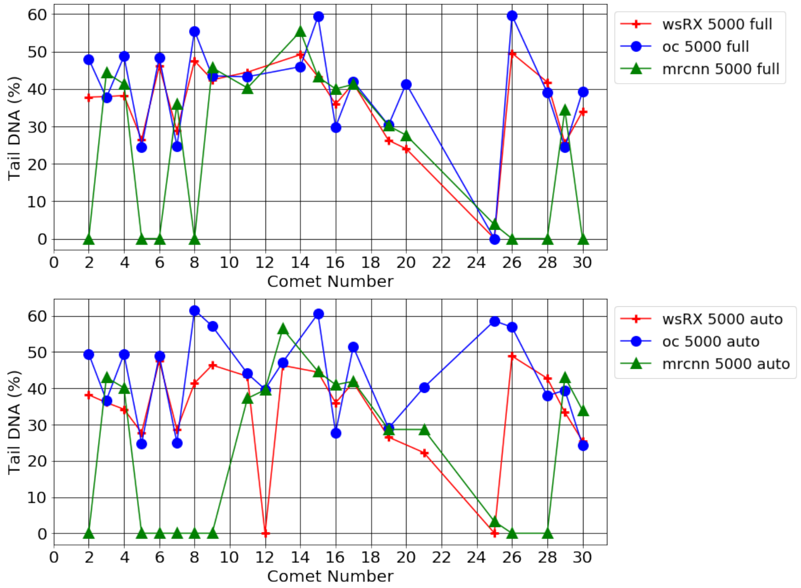


Figure 5.13: *Tail DNA measures for comets detected using the expanded watershed algorithm (wsRX), OpenComet (oc) and the Mask R-CNN (mrcnn) model using the auto and full contrast functions at exposure value 5000.*

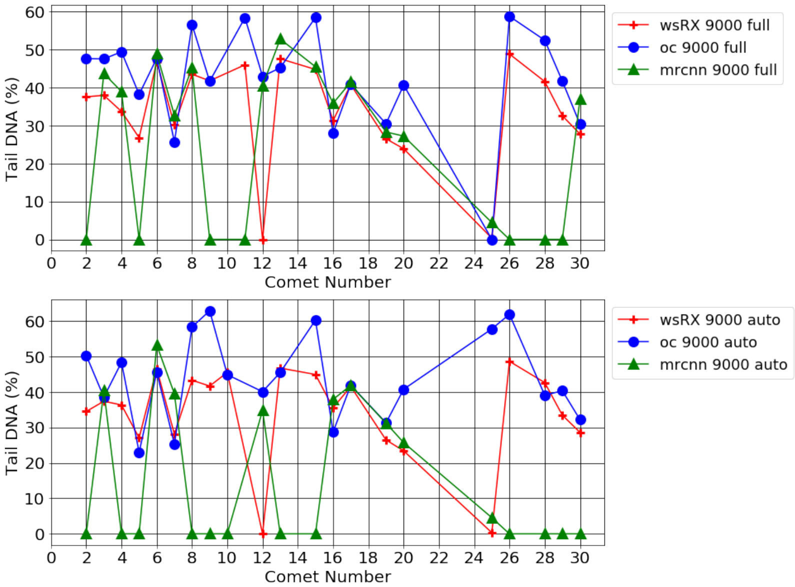


Figure 5.14: *Tail DNA measures for comets detected using the expanded watershed algorithm (wsRX), OpenComet (oc) and the Mask R-CNN (mrcnn) model using the auto and full contrast functions at exposure value 9000.*

5.2.6 Background Corrections

In order to perform accurate comet measurements, background corrections are necessary to ensure the true signal from the fluorescent stain is measured. In Section 4.4.3, a vignette effect was shown, causing darker edges in some images [159]. This did not have an effect on comet segmentation, as the threshold applied to images was higher than the intensities at which the darker regions appear. Analysis was performed to investigate darkened edge variation to determine if a more complex background correction was required to take this into account when measuring comets. Figure 5.15 shows the intensity profile taken across image A to visualise the background variation. The line segment taken across the image is also shown, followed by the corresponding line laid over the image thresholded at 5% and 10% of the maximum pixel value. As can be seen, the brighter central area is apparent at a 5% threshold, but completely disappears at 10%. After testing this across the whole data set, it was confirmed that the 10% threshold ensures that vignetting does not affect comet identification. However, we see from Figure 5.15b that the raw image does have an intensity variation across it that increases to a maximum around the centre of the image and decreases again. As this variation is minimal, only a maximum increase of 7 in intensity, no special consideration is needed for the background subtraction.

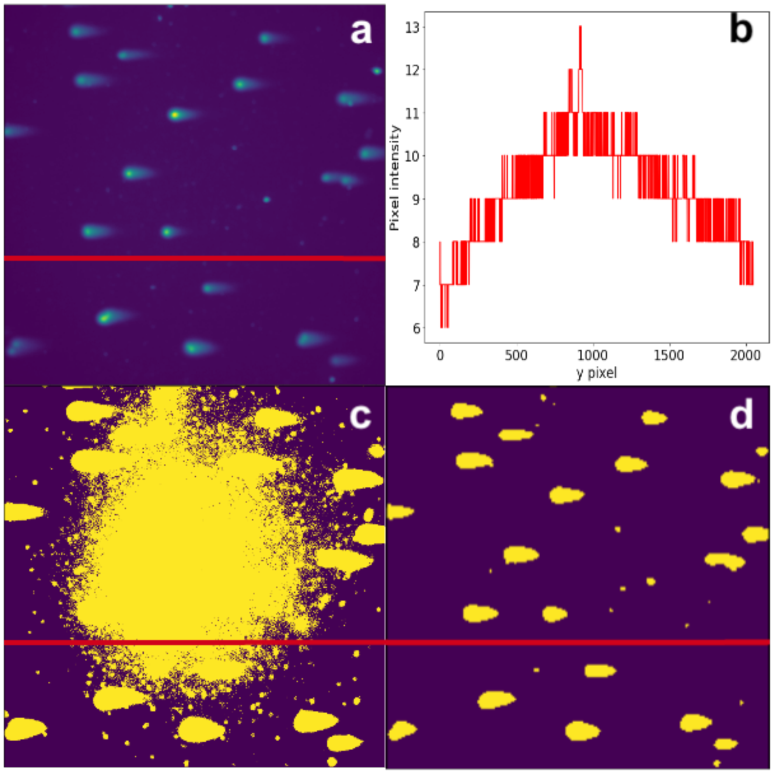


Figure 5.15: *Analysing the background of image A.* ***a:*** *raw image,* ***b:*** *intesnity profile of the red line shown in the other sub-figures,* ***c:*** *image thresholded at 0.05% the maximum pixel value and* ***d:*** *image thresholded at 10% the maximum pixel value.*

To correct the intensity, a local background was determined for each comet. This was conducted by applying a further two expansions of 10 pixels to each comet. The first expansion was necessary to create a large enough separation from the comet object to ensure only background pixels are detected. The second expansion was used, subtracting all the pixels from the first, leaving a background ring, shown in blue in Figure 5.16. This region was used as a background sample and its average intensity value was used as a background correction and subtracted from all intensity values within the comet. By using the local region around the comet, any background fluctuations, even if they are very small are taken into account and each comet is corrected according to its local environment rather than averaging the background correction out in a more global manner.

A large proportion of the values for average background were found to be 0, especially in the case of overexposed images. This was expected, as the comets in these images were often already overestimated due to the expansion applied. The nature of these images' intensity distributions often caused a near flat background of 0 pixel intensity on average, due to the reduced dynamic range, resulting in very bright comet regions and contrasting backgrounds of very low intensity. The absence of a background correction for overexposed images adds further error to the measurements performed upon them, resulting in misleading results. Where a background average > 0 was calculated, its intensity value was between 10-20, considerably less than that of a typical comet. The average background value found for each identified comet across all images is shown in Appendix B.3.

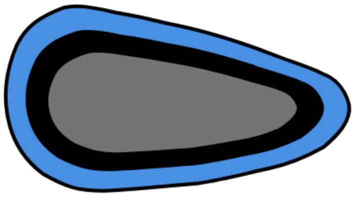


Figure 5.16: *A schematic of how the background subtraction is found for each comet. The original comet (following an overall and tail expansion) is shown in grey. The expanded comet region, to create a separation from the comet, is shown in black and the final expanded region, from which the background is calculated is in blue.*

As this expansion is some distance away from the clusters originally identified by the watershed algorithm, the likelihood of the new expansions overlapping with other comets, or incorporating bodies of contamination has significantly increased. Therefore, a second overlap check was performed following the final two expansions, in order to discard any comets that now overlapped or touched the image edge.

Background corrections do not affect tail length measurements so only tail DNA was explored for the corrected measures. Figure 5.17 and Figure 5.18 show these comparisons for image A and B. Correcting the comet intensity reduces the tail DNA measurements as expected, as the subtraction has the greatest effect on low intensity tail pixels, reducing the proportion of intensity in the tail. The correction did, however, create a greater difference between the OpenComet measurements, which was already large ($> 10\%$) in some cases.

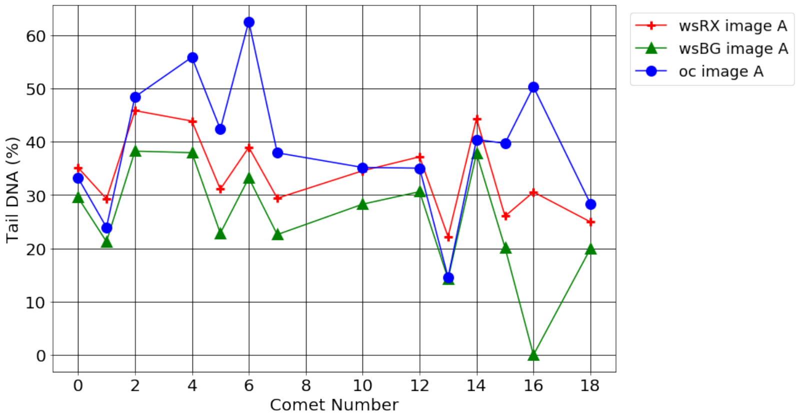


Figure 5.17: *Tail DNA measurements for the watershed comets expanded overall with the additional tail expansion with (wsBG) and without (wsRX) the background corrections compared to the OpenComet (oc) tail DNA measurements for image A.*

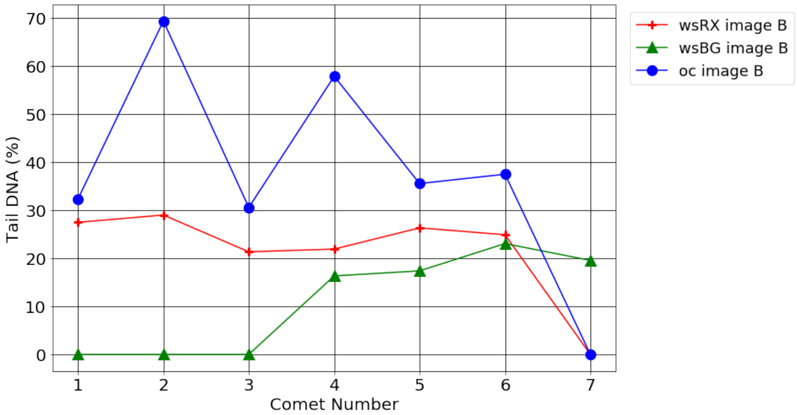


Figure 5.18: *Tail DNA measurements for the watershed comets expanded overall with the additional tail expansion with (wsBG) and without (wsRX) the background corrections compared to the OpenComet (oc) tail DNA measurements for image B.*

Figure 5.17 in particular highlights the high variation in tail DNA measurements of comets within the same image. Comets such as 0, 1 and 13 are examples where correcting the background of the expanded watershed comets brings the measurements much closer and even in approximate agreement to those of OpenComet, but there are still cases (such as comets 5, 7 and 15) that have a large discrepancy. This can be attributed to the different methods OpenComet uses to perform measurements on comets, or to the variation of cells, even from the same cell line, that have different radiosensitivities, affecting the damage they incur from the same level of radiation/amount of damaging agent.

5.2.7 Abnormal Cells

Upon analysing the image dataset, it was found that some comets not only appeared larger but also brighter, having 1.5-2 times higher total intensity within the comet. Two hypotheses can explain this. The first is that overlapping comets are the cause, the result of cells sticking together when they were embedded in the agarose. The second relates to the origin of HeLa cells. As previously mentioned, the HeLa cell line was first cultured over 70 years ago, It has been continually grown internationally since then, allowing enough time for DNA mutations to occur. Genetic mutations have led to an increased and varied number of chromosomes in cultured HeLa cells, ranging from 76-80 as opposed to the usual 46 for humans [160]. This would translate into a higher level of fluorescence in comet assay data of such cells, as seen in some cases.

Evidence points towards the second hypothesis being the more plausible of the two. Firstly, if cells sticking together were the cause of this higher intensity, it would follow that three or four cells could stick together, translating to measurements of up to four times the expected intensity, which was not observed. However, it may be that the limited datasets analysed did not contain a large enough number of images to enable this rare clumping of cells to be seen. Additionally, during the addition of cells into the molten agarose, trypsin is used [161]. This chemical allows a mono-layer of cells to be isolated and lifted from their growth medium, which further emphasises that cell overlaps in the z-direction due to cells sticking together is unlikely.

5.3 Comet Asymmetry as a New Damage Measure

Although tail length, tail DNA and tail moment are widely calculated and published as Comet Assay DNA damage results, there is an intrinsic error associated with these measures that is not considered and therefore not reported. This stems from the methods through which the comet head is defined by different algorithms, which of course impacts the results produced. A new asymmetry measure is proposed, which aims to reduce this source of inconsistency by removing the need to define comet head and tail pixels prior to measurement. This is a very simple measure that calculates the difference in pixels about some central point. The point is chosen in the same way the circle centre was found for defining the comet head, taking the average position of the comet's brightest pixels. As mentioned previously, this is based on the assumption that the brightest pixels, corresponding to the most dense region of DNA, are located in the head and is a reasonable assumption for the dataset analysed here. Other images, depicting more damaged cells or with very different experimental setups would not be appropriate for measurement via this method.

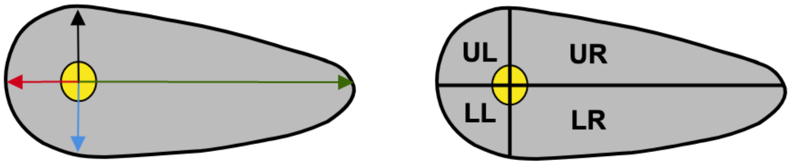


Figure 5.19: *Defining comet quadrants according to the head centre, which is taken as the average brightest pixel.*

Once the head centre is defined, a pixel count above, below, left and right of it is taken. This produces 4 sets of pixels, each assigned to a comet "quadrant", and referred to as the upper left (UL), upper right (UR), lower left (LL) and lower right (LR), as shown in Figure 5.19. Asymmetry is then measured for the longitudinal and transverse positions ($asymm_x$ and $asymm_y$), in relation to the electric field axis, to find the imbalance in pixel intensity. Using Figure 5.19 as a reference for quadrant names, the two asymmetry measures are calculated as

$$Left = \sum_{i=0}^{i=n_{UL}} UL_i + \sum_{i=0}^{i=n_{LL}} LL_i \qquad Right = \sum_{i=0}^{i=n_{UR}} UR_i + \sum_{i=0}^{i=n_{LR}} LR_i \tag{5.1}$$

$$Upper = \sum_{i=0}^{i=n_{UL}} UL_i + \sum_{i=0}^{i=n_{UR}} UR_i \qquad Lower = \sum_{i=0}^{i=n_{LL}} LL_i + \sum_{i=0}^{i=n_{LR}} LR_i \tag{5.2}$$

$$asymm_x = \frac{Left - Right}{Left + Right} \tag{5.3}$$

$$asymm_y = \frac{Upper - Lower}{Upper + Lower} \tag{5.4}$$

where n_{UL}, n_{UR}, n_{LL} and n_{LR} refer to the number of pixels in each quadrant.

Values of 0 for $asymm_x$ would indicate symmetry about the central head pixel, described by a vertical centre line, and communicate that there is no comet tail i.e. the cell has undergone no damage. This is unlikely as even control cells (those that have not been subjected to any damaging agents) exhibit some level of comet tail, which may be due to the migration of unbroken strands or damage from background radiation. A value of $asymm_y \sim 0$ would be expected as the intensity about the horizontal centre line should in principal be equal. However, experimental factors such as the parallel alignment of slides in the electrophoresis tank and contaminants can skew this measure.

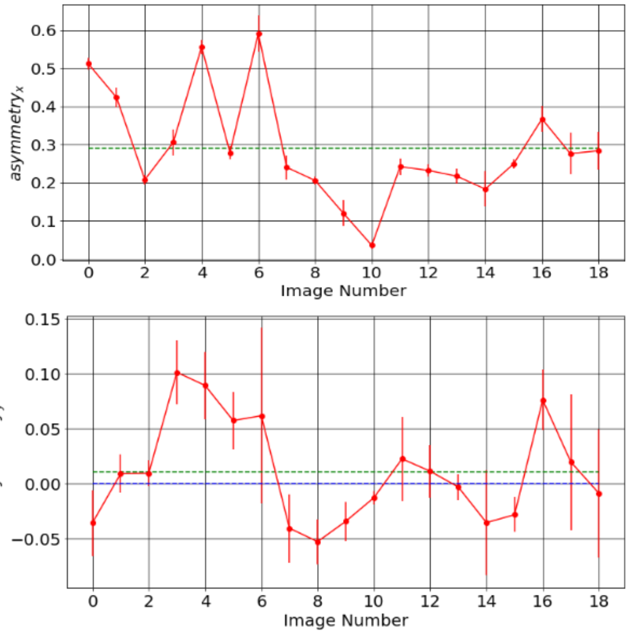


Figure 5.20: *Average asymmetry measurements in the x and y direction for the comparison dataset. The average asymmetry value is indicated by the green line and the blue line shows the expected 0 value for $asymm_y$.*

Asymmetry analysis was performed on all 20 images of the comparison dataset, for which the results are shown in Figure 5.20. It can be seen that the asymmetry values in both the x and y direction vary significantly across the images. This was expected in the x direction as the tail DNA and tail length measures have shown high variation, but it was unexpected in the y direction. It may be accounted for by the method used to approximate the head centre, or could be indicative of misalignment of the microscope slides within the electrophoresis tank, so that the slides are not exactly aligned with the electric field direction. Analysis of $assym_y$ measurements from the expected 0 produced a χ^2 value of 0.079 per degree of freedom. By removing the average values, and their associated errors (which are much larger than the others) from images 6, 17 and 18 the χ^2 value per degree of freedom became 0.092. Images 6, 17 and 18 had large errors on their average asymmetry as fewer comets were identified in these images. The low χ^2 value shows that the variation from 0 is not as large as first thought, but also that a larger sample of comets is needed per image to reduce the larger errors and produce a more concrete conclusion from this measure.

To verify the asymmetry measurements, they were compared, comet by comet, for images A and B, shown in Figure 5.21 and Figure 5.22 respectively.

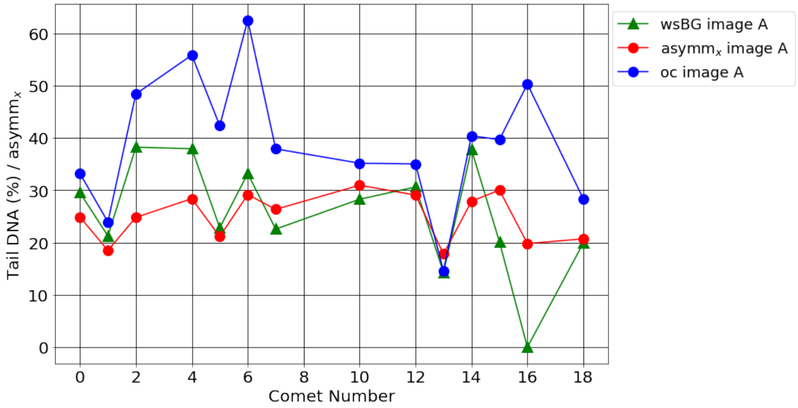


Figure 5.21: *Comparing tail DNA measurements calculated from the background corrected watershed comets (wsBG) with the OpenComet measurements (oc) and the asymm$_x$ values for comets identified in image A.*

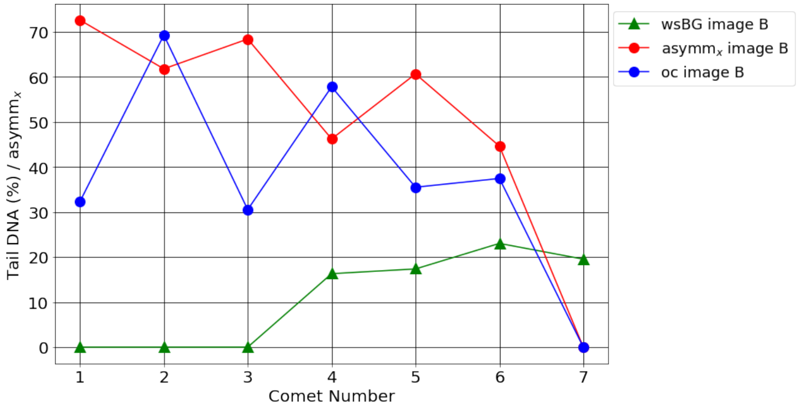


Figure 5.22: *Comparing tail DNA measurements calculated from the background corrected watershed comets (wsBG) with the OpenComet measurements (oc) and the asymm$_x$ values for comets identified in image B.*

There is a similar trend in how measurements vary, cluster to cluster, between the asymmetry values and the corresponding background corrected values, but this is only seen for image A. This may be due to the fewer comets in image B, but as the variation is much larger, it can be concluded that more than this is causing such a high variation. As discussed previously, image B is an overexposed image, which may be causing its tail DNA and $asymm_x$ measures to behave peculiarly. The head centre is the same for both, so it is the definition of head and tail pixels that is causing the discrepancy. The significantly lower tail DNA values for the background corrected comets indicate that too few tail pixels are being defined, due to overestimating the comet head. This is probably because overexposure can significantly shift the head centre to the right, resulting in a circle of larger radius being used in defining head pixels.

Although the $asymm_x$ and tail DNA values for the background corrected comets are in some degree of agreement for image A, the asymmetry measure does not take the measurement much closer to the OpenComet tail DNA values on the whole. There are two conclusions that can be made here. The first is that the identification of comets using the watershed algorithm is not robust enough to successfully identify the majority of pixels belonging to the comet and should not be used for comet assay image analysis. The second is that the methods used to perform comet measurements by OpenComet are not transparent enough to definitively compare with. This is further emphasised by the high level of variation between the OpenComet measurements and the corresponding asymmetry values in Figure 5.21. Comets 1, 10 and 13 are much more in agreement than the others, showing that similar DNA damage conclusions can be reached by both methods of analysis, but this is not reliable for all comet instances.

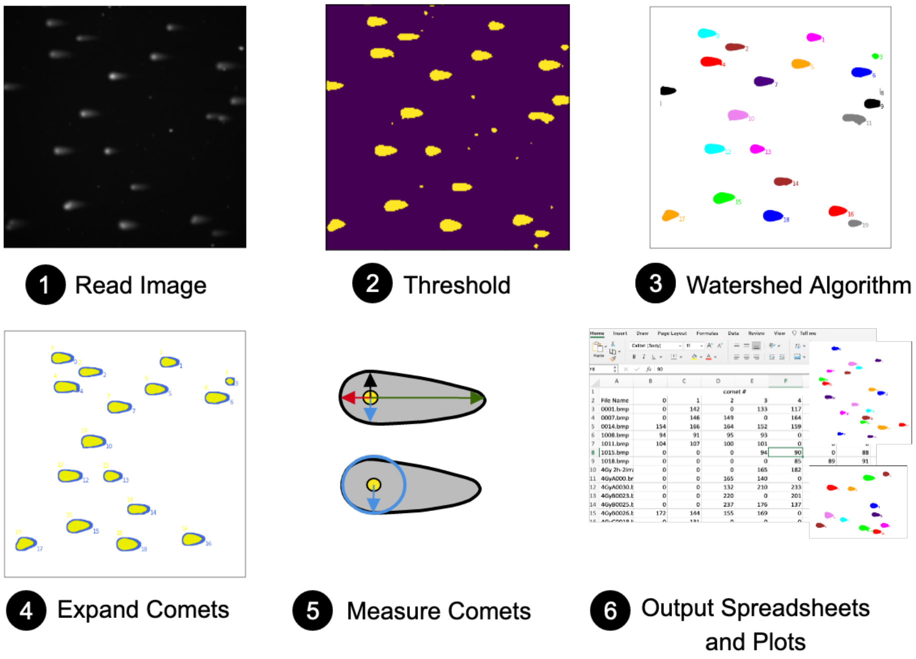
5.4 Conclusions

As in the case of comet identification, the varied nature of Comet Assay data affects comet measurement. The widely used metrics of tail length and tail DNA can be unreliable and potentially misleading if the precise method of head definition is unknown, as seen through the application of OpenComet.

This software does not immediately discard small regions of contamination, although there is a check for small comet areas. This must mean that these regions either pass the bright head analysis check, or the comet intensity profile is deemed appropriate, highlighting the

issue with such methods that do not have robust enough constraints on what is and is not a comet region.
The Mask R-CNN architecture measurements are generally similar to those of the watershed identified comets, as the same method of measurement is implemented in both. It is the identification of comets that is inconsistent, making this an unreliable method in its current form.

A new method of measuring asymmetry removes the need for background corrections from extra comet expansions, allowing more comets to be measured. Asymmetry may be seen as a more favourable measure as it does not require head and tail pixels to be defined, a source of high variability. It does however rely on the premise that the head contains a comet's brightest pixels and is not applicable for use on highly damaged cells that have a high proportion of DNA in the comet tail. This means that the extra expansions do not need to be applied, enabling more comets to be measured per image, especially in the case of more crowded comet images.



	A	B	C	D	E	F
1				comet #		
2	File Name	0	1	2	3	4
3	0001.bmp	0	142	0	133	117
4	0007.bmp	0	146	149	0	164
5	0014.bmp	154	166	164	152	159
6	1008.bmp	94	91	95	93	0
7	1011.bmp	104	107	100	101	0
8	1015.bmp	0	0	0	94	90
9	1018.bmp	0	0	0	0	85
10	4Gy 2h-2im	0	0	0	165	182
11	4GyA000.br	0	0	165	140	0
12	4GyA0030.b	0	0	132	210	233
13	4GyB0023.b	0	0	220	0	201
14	4GyB0025.b	0	0	237	176	137
15	4GyB0026.b	172	144	155	169	0
16	4GyC0018.b	0	131	0	0	0

Figure 5.23: *An overview of the six main components that form the analysis pipeline for measuring DNA damage in Comet Assay data.*

A complete, automated analysis pipeline has been developed that combines the comet identification and measurement processes. It is able to read in a directory of Comet Assay images and output a spreadsheet of comet measurements, accompanied with reference plots for each analysed image, as shown in Figure 5.23. The measurements performed are tail length, tail DNA and comet asymmetry, and are indicative of the damage present in the imaged cells. The optimum values applied at each stage of the analysis pipeline are outlined in Table 5.1.

Analysis Pipeline Step	**Optimum Conditions**
1: Read Image	N/A
2: Threshold	10% of the maximum pixel intensity value
3: Watershed Algorithm	Constraints applied to filter clusters as in Table 4.1 and comet heads found by thresholding at 25% of the maximum intensity value
4: Expand Comets	10 pixel overall comet expansion and an additional 20 pixel tail expansion
5: Measure Comets	Comet head radius definition found from maximum of head centre to either northernmost, southernmost or westernmost comet pixel
6: Output spreadsheet and plots	Comet measurements of tail length, tail DNA and asymmetry

Table 5.1: *An overview of the optimum conditions used at each step of the automated comet analysis pipeline shown in Figure 5.23.*

Chapter 6

Simulated Comet Assay Images

Monte Carlo (MC) simulations are useful in a wide range of applications. Here, they allow us to create "test" data, mimicking what we would find experimentally. In the MC used in this work, random number generators are utilised to simulate how DNA strands are damaged, leading to various different strand lengths, and also where in the agarose coated microscope slide the cells are embedded. There are two main reasons that a Monte Carlo model is useful in the image analysis of Comet Assay data.

Firstly, there is a lack of data on which the Mask R-CNN instance segmentation model could be trained. Although it seems to still have a high degree of accuracy in some of the segmentations it performs, without sufficient data it was unable to be tested in a rigorous and extensive manner. A further reason such a Monte Carlo would be useful is that it would provide ground truth data to compare the segmentation outputs with. A simulated image provides all pixels associated with the DNA. This could then be compared directly with the segmentation masks to see the true accuracy and precision of the methods, within the limitations imposed by the quality of the simulation.

6.1 Overview of the model and assumptions made

The Monte Carlo simulates the random breaking of strands, followed by their drift through agarose during electrophoresis. It then models the effects of the camera optics and sensor by applying Gaussian smearing to produce an image with sharpness similar to those produced experimentally following a Comet Assay, as shown in Figure 6.1.

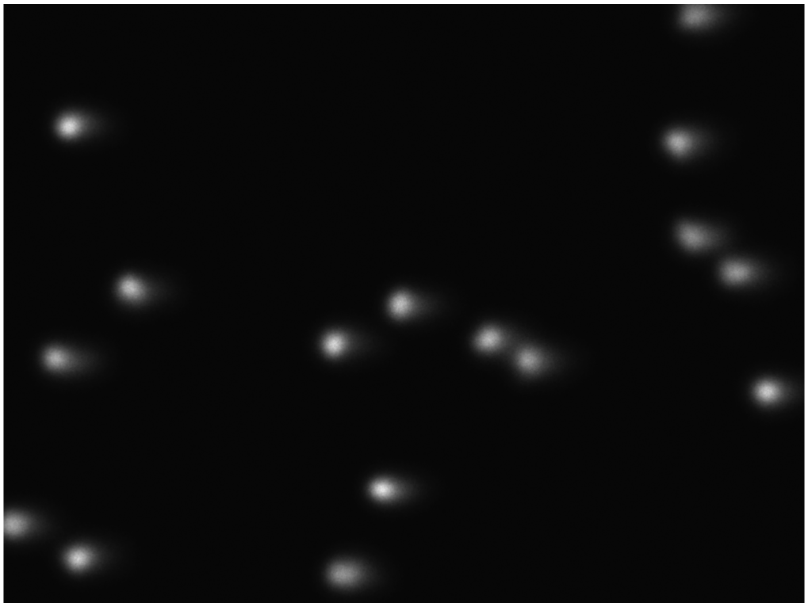


Figure 6.1: *An example simulated image produced from the Monte Carlo model.*

Due to resource constraints, the DNA was simulated as much less granular than it actually is, with each strand comprised of 5 000 parts rather than the $\sim 3 \times 10^9$ base pairs in a single strand of DNA. Large chunks of what should be individual base pairs are grouped together in order to reduce the computing resources required. Each strand is assigned a random break value, $pBreak$, which determines the number of breaks the strand suffers. The higher the value of $pBreak$, the higher probability a strand has of being broken into multiple parts, relating to the dose given to cells in an experimental Comet Assay context.

The tuning of parameters for this model was done by trial and error in order to produce the most experimental-like images, and are summarised in Table 6.1. This is not a robust method but as the aim of the model was to produce images as visually close to experimental data, and not one that models the physical mobility of the DNA strands through agarose, it was deemed appropriate. The parameter $maxStretch$ was defined to set the maximum possible longitudinal length of the comet. Two further parameters, $smearBreak$ and $smearWhole$ were set as the standard deviation of the Gaussian smearing for the broken strands and all strands respectively. As broken strands are able to move more easily than unbroken strands, the value of *smearBreak* was set higher, to reflect the increase in variance of their position under electrophoresis.

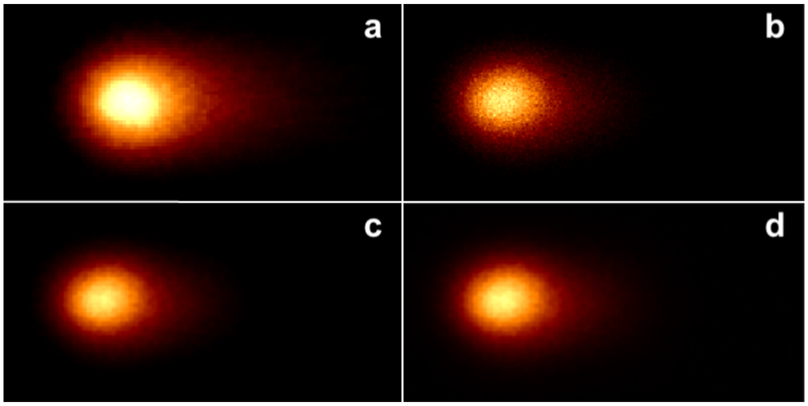


Figure 6.2: ***a**: an experimental comet, **b**: a raw simulated comet, **c**: a Gaussian smoothed simulated comet and **d**: a Gaussian smoothed comet with additional noise.*

The co-ordinates of each strand (and its encompassing 5 000 parts) then formed a histogram, showing the final comet intensity. As shown in Figure 6.2, the comet image produced from this is grainy, and not smoothed out as seen in experimental images. This is due to the intrinsic smoothing and smearing that occurs during image acquisition, where adjacent pixels are correlated. To mimic this process, the comet pixels were smoothed using a Gaussian filter with $\sigma = 1.03$. Again, this standard deviation parameter was applied following trial and error to replicate the experimental smoothing of pixels during imaging. A final addition of noise was applied to imitate image background, as without it, the entire background region was of intensity value 0, which is not seen in real images.

To create a whole image containing comets, a random number was found from a normal distribution of mean 13 and standard deviation 5.60, and assigned as the number of cells to be simulated. These distribution parameters were set after analysing the number of comets present in a sample of 50 experimental images. A cell position was first defined from a random uniform distribution within the image dimension constraints (1040 × 1392), labelled x_{cell} and y_{cell}. The cell centre was then built upon, to describe the x and y positions of each strand by including random variables for r, $theta$ and phi from uniform distributions with range 0 to $cellradius^2$, -1 to 1 and 0 to 2π, respectively. An example distribution of strands within a cell of radius 25 pixels is shown in Figure 6.3.

The final position of each strand and its subsequent parts were found by applying the whole strand and broken strands smearing. Broken fragments were allowed to drift up to a random position between their starting x position and $maxStrecth$.

Parameter	Value
cell radius	25 (pixels)
$maxStretch$	3.75 × cell radius
$smearBreak$	0.65 × cell radius
$smearWhole$	0.45 × cell radius

Table 6.1: *A summary of the parameters used in the Monte Carlo to produce simulated Comet Assay images.*

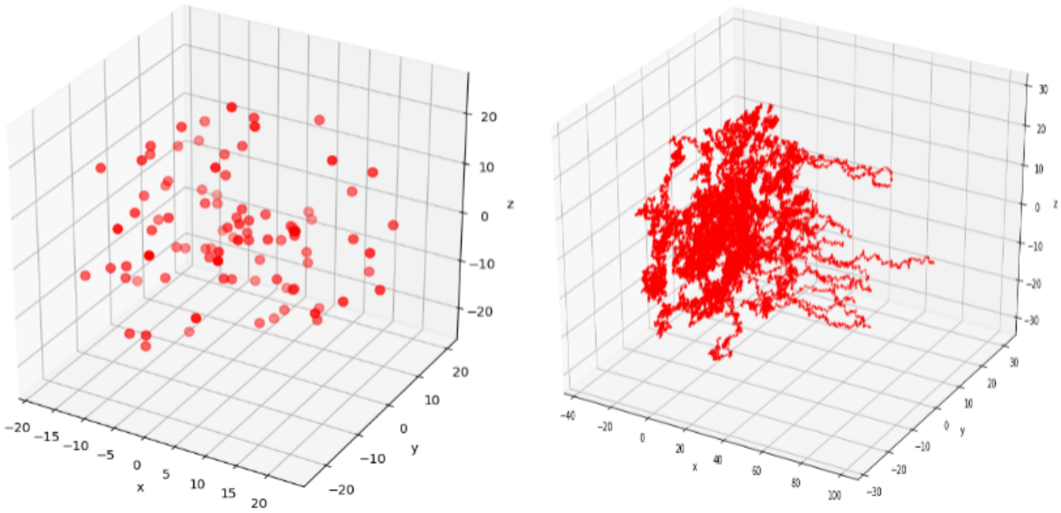


Figure 6.3: *The initial random distribution of the beginning of all strands within a cell volume with radius 25 pixels (left) and the depiction of the entire strand, comprised of 5 000 parts with some drift of broken strands in the positive x direction (right).*

Figure 6.4 depicts another example of modelling a single cell, showing the raw positions of strands alongside the cell following a Gaussian smoothing of broken and unbroken strands according to parameters $smearBreak$ and $smearWhole$.

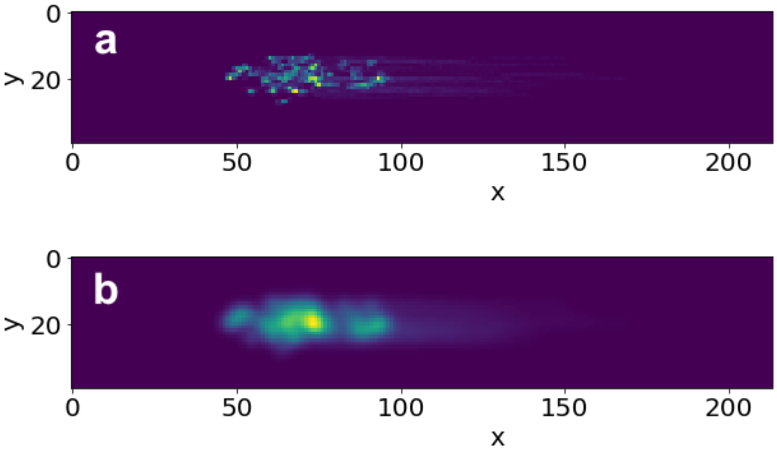


Figure 6.4: *Modelling a single cell.* ***a****: the raw positions of the strand segments and* ***b****: the Gaussian smoothed positions providing an image closer to that seen in experimental data.*

The code used to create simulated Comet Assay images can be found in the github repository shown in Appendix C.1.

6.2 Evaluating the Model

To compare an experimental comet with a simulated one, their comet intensity histograms were compared, as shown in Figure 6.5. We can see that a very similar intensity range and distribution is seen for the experimental and simulated smoothed comets. The addition of noise shifts the smoothed distribution, so that the majority of the pixels are not of value 0.

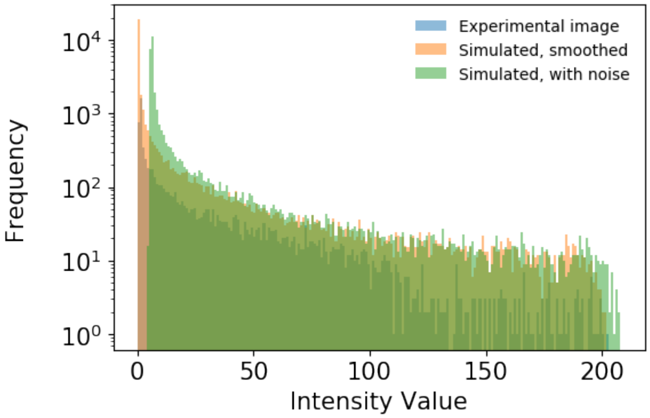


Figure 6.5: *Comparison of the intensity distribution of a simulated comet with one from image B. Frequency has been plotted on a logarithmic scale.*

The whole image histograms were also compared to images A and B (the same image examples used in Chapters 4 and 5), to see how the intensity distributions differed on a global level, shown in Figure 6.6. The distribution of the simulated image is much closer to that of image A than the overexposed image B, indicating that the intensity of the simulated images was reasonably modelled.

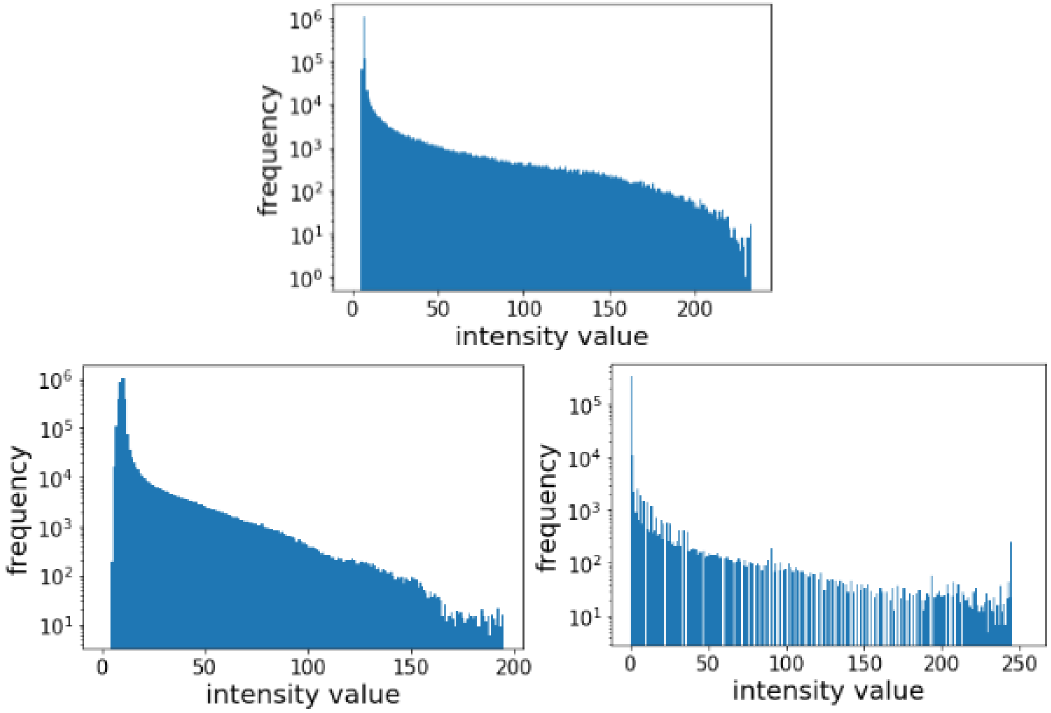


Figure 6.6: *Comparison of the histograms of a simulated image (top), image A (bottom left) and image B (bottom right). Frequency has been plotted on a log scale.*

6.3 Measuring Asymmetry to Predict Strand Breaks

By varying the parameter $pBreak$, the level of damage cells exhibit within an image can be controlled. A set of simulated images was produced, consisting of 40 images for p_{break} values 0, 0.1.... 0.9. From this, average tail DNA and asymmetry values were found from a large range of images using the methods outlined in Chapter 5. The relationship between p_{break} and tail DNA and p_{break} and $asymm_x$ are shown in Figure 6.7 and Figure 6.8 respectively.

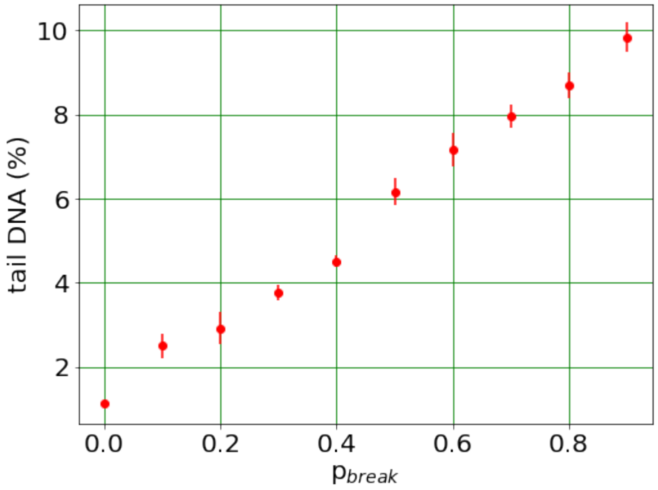


Figure 6.7: *The relationship between p_{break} and tail DNA for simulated images.*

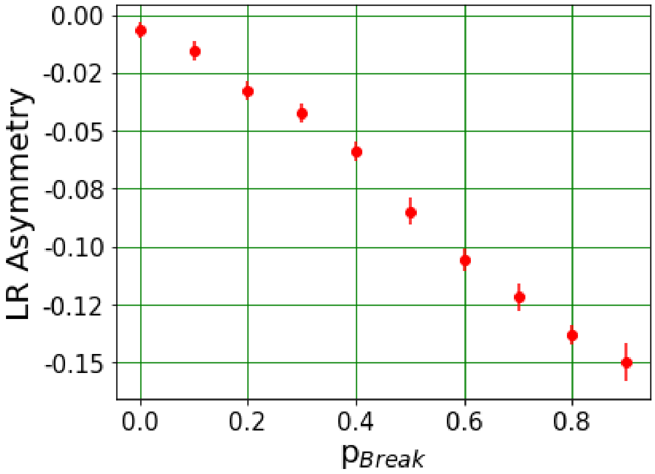


Figure 6.8: *The relationship between p_{break} and asymmetry for simulated images.*

As can be seen, there is a slightly quadratic relationship between p_{break} and both tail DNA and $asymm_x$. A linear fit was also explored on the data, however a quadratic model had a better agreement with the data. This analysis can be seen in Appendix B.4.
The LR asymmetry values are negative, indicating that the right pixel intensities sum, that are subtracted from the left are larger, and show a comet tail.

The relationship between p_{break} and $asymm_x$ was parameterised as

$$p_{break} = a + bA + cA^2, \tag{6.1}$$

and the values of the coefficients, a, b and c, were determined using a least squares fit, as shown in Figure 6.9. The co-coefficient values found for the model are shown in Table 6.2. The inverse function was used to determine p_{break} values from $asymm_x$ measurements. The error decreases with asymmetry magnitude, as the error factors in the asymmetry value in the calculation. The χ^2 value per degree of freedom is 1.08, indicating the model represents the data well.

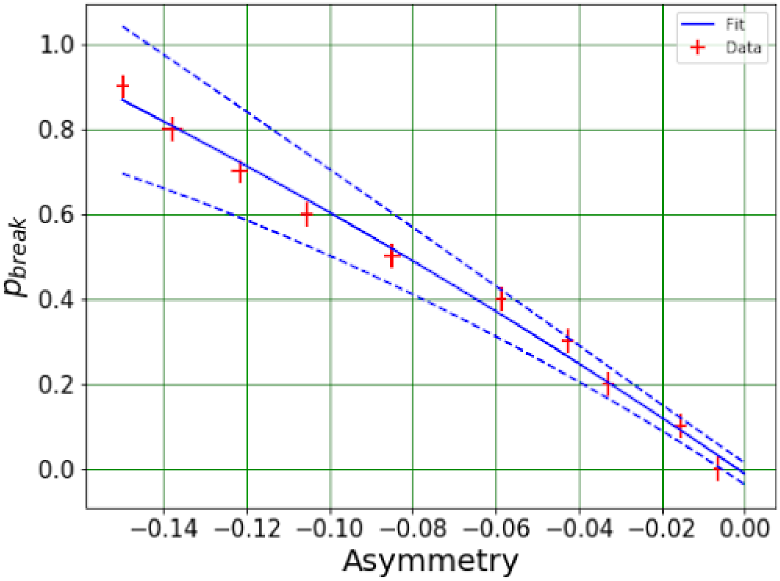


Figure 6.9: *The model and its associated error plotted with the asymmetry data used to create it.*

Parameter	Value	Error
a	-0.012	0.026
b	-6.720	0.835
c	-5.685	5.216

Table 6.2: *A summary of the parameters found when applying a linear fit to the asymmetry values of the MC data. All values have been given to 3 decimal places.*

Further simulated data, created with p_{break} values of 0.45 to 0.85 was used to test the function. Again, average asymmetry values for each set of images was calculated, alongside their errors, and used as inputs to the inverse function to find the corresponding p_{break} value, as shown in Figure 6.10. The line $y = x$ is also shown in blue to compare the expected and calculated values. We see that the values of p_{break} obtained from the inverse function are in agreement with the expected values on the whole, but only marginally within the error tolerance. The values calculated for p_{break} values of 0.45 and 0.85 are not in agreement. It is not known what caused this disagreement between the calculated and expected values, indicating teh MC may not be working as expected at the lower and higher values of p_{break}, and further investigation would be needed.

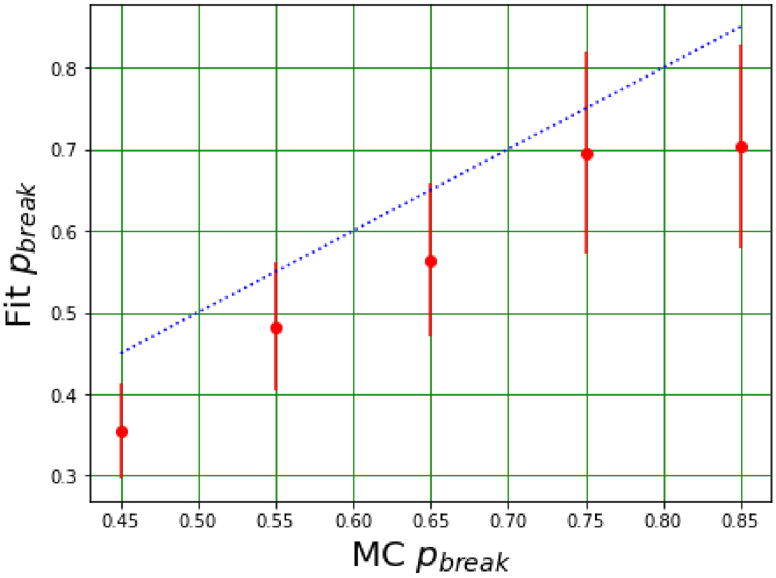
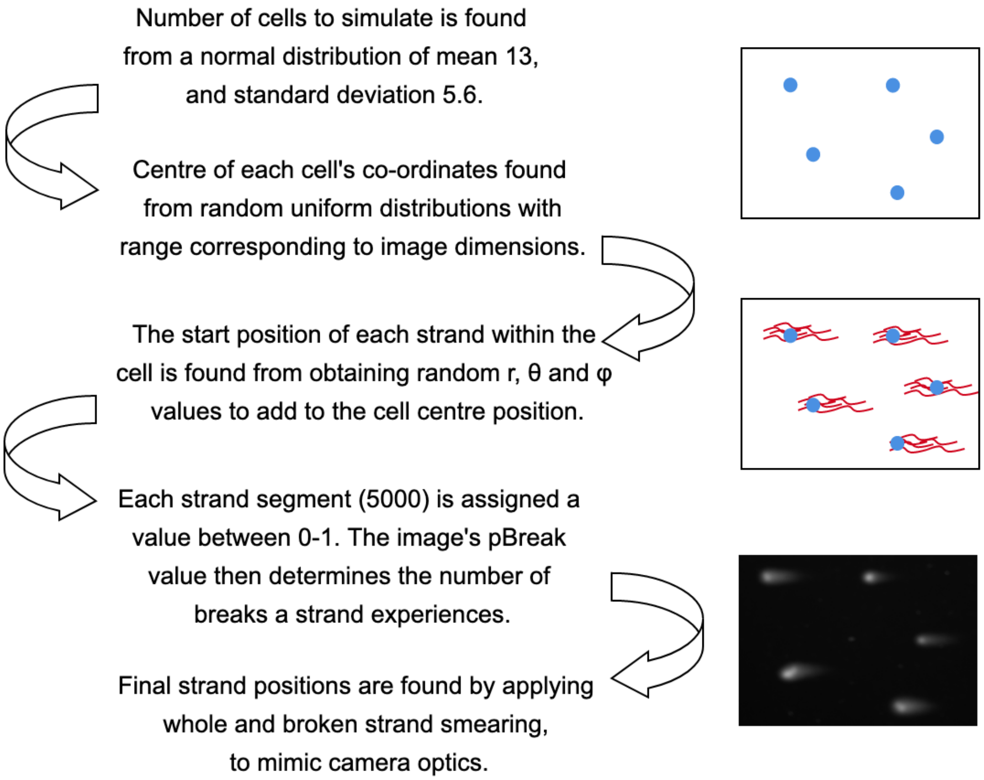


Figure 6.10: *Calculated p_{break} values from the inverse model for measured asymmetry values. The line $y = x$ is also shown in blue to indicate the expected values.*

6.4 Conclusions

The MC model outlined is able to produce images visually similar to those seen experimentally. It is based on a p_{break} parameter that varies the number of strand breaks possible within an image and is a model of cells suffering from varying degrees of DNA damage. Similarities are seen in the intensity distribution of simulated and experimental images, on both the local comet level and the level of the whole image. An overview of the steps taken by the MC to produce a simulated Comet Assay image is shown in Figure 6.11.



Number of cells to simulate is found from a normal distribution of mean 13, and standard deviation 5.6.

Centre of each cell's co-ordinates found from random uniform distributions with range corresponding to image dimensions.

The start position of each strand within the cell is found from obtaining random r, θ and φ values to add to the cell centre position.

Each strand segment (5000) is assigned a value between 0-1. The image's pBreak value then determines the number of breaks a strand experiences.

Final strand positions are found by applying whole and broken strand smearing, to mimic camera optics.

Figure 6.11: *An overview of the MC steps to produce a simulated Comet Assay image.*

However, the asymmetry and tail DNA measurements, even for the highest p_{break} values, are much lower than those measured in experimental images, shown in Chapter 5. This suggests that the simulation is not modelling the comet tail well, as is also seen by eye. Very basic and crude assumptions were made in the modelling of DNA here. In order to produce images closer to those seen experimentally, a deeper understanding and modelling of the DNA mobility would need to be developed through further studies.

Such studies could include atomic force microscopy imaging of agarose to better understand its composition, aiding the MC model. Data collated from exploring the mobility of DNA at different strand lengths through agarose could also aid the model, as the data available, and the details of how it was acquired are very limited [162]. This would enable more accurate migration values to be included, producing more realistic comet compositions.

Chapter 7

Conclusions

The Comet Assay is a useful tool for determining the DNA damage present in cells. It is widely used and relatively easy to implement on a wide range of cell lines. Comet Assay analyses feed into cancer research and are instrumental in measuring the efficacy of potential new treatment options.

A high level of variability is seen in Comet Assay data, stemming from both the chemical and physical experimental protocols and the image acquisition process. Even small variation in the experimental procedure can lead to a considerable degree of variation in results, and this variability is not restricted to one stage of Comet Assay protocols. It has been shown that all steps can lead to some degree of variation which can be compounded throughout the process. To overcome the many points of variability, a standardised protocol could be designed to act as a base, from which researchers can clearly outline any alterations or additions that are made for their studies. The field as a whole needs to move towards detailing experimental methods in a thorough manner, enabling the context of results to be understood in a more complete way.

With regard to the imaging process, the true extent of the measurement uncertainties is not as well researched and was explored in this work. No mention of imaging system calibration is seen in the vast majority of published literature. As there is such a range of equipment, the data storage format and analysis methods used, it is important to outline all these factors as they can affect results obtained.

There are many methods for image processing and various of these for image segmentation have been explored here. The best method is entirely dependent on the image composition and outcome desired. “Classical” clustering algorithms (DBSCAN and watershed) were compared with neural network instance segmentation, using Mask R-CNN architecture. Both methods have their advantages, however for segmenting comets in Comet Assay data, the watershed algorithm proved to be more robust and consistent. Mask R-CNN was able to cope with contamination within images better in some cases, but the high degree of variability made the method unreliable in the form presented here.

The automated analysis pipeline developed here succeeds in minimising human input, only requiring the user to define the image directory to be analysed and the output location for the spread sheet of measurements and corresponding plots. This removes human bias in the comet selection and identification processes.
Steps were taken to maximise the number of comets measured per image. This is important when considering the high variability seen in comets, even within a single image. By using all available data, the accuracy of results is improved upon. Removing contaminated comets was also prioritised, as these can skew the averaged measurements causing inaccurate conclusions to be drawn. This was achieved by applying constraints to the identified comet dimensions and topology, allowing background and instances of comet overlap to be removed.
As expected, due to the high variation in comet dimensions, some contaminated comets still made it to the measurement stage of analysis, making it crucial that plots of the analysed comets in each image were available for a user to review. This essential step unfortunately remains necessary, especially in cases where results may steer treatment plans and affect patient safety.

The desire to produce completely automated systems is universal and artificial intelligence and machine learning technologies continue to advance, making this increasingly possible. But the implementation of such processes in medical settings requires deep consideration and should always incorporate the unrivalled experience of healthcare professionals. Although rigorous testing is undertaken prior to incorporating automated systems, the “black-box” nature of artificial intelligence algorithms mean caution should always be taken. The function of AI systems is not yet fully understood and can only be inferred through testing, so any conclusions made cannot be conclusive. By outputting plots of measured comets alongside the measurements themselves, the system developed here allows the user to ascertain the reliability of results and whether certain measurements should be omitted.

Although changes in image contrast and exposure did not affect comet identification, due to the methods chosen in this work, variations in measurement were seen, particularly for comets identified by the neural network and the OpenComet software. Tail length in particular has a higher variation, compared to tail DNA and highlights the need for both measures to be stated in results, as conclusions made from a single measure can be misleading.

As comet measurements are directly based upon how head and tail pixels are defined, the commonly used measures are difficult to repeat, unless the exact head defining methods are known. A novel asymmetry measure is introduced, removing the need to explicitly define the two categories of pixels in each comet. As currently implemented, it does, however, only apply in circumstances where the head contains the comet's brightest pixels, and therefore is not recommended for use in analysing heavily damaged cells.

A recurring issue with evaluating and analysing Comet Assay data is the volume of data available. Once an experiment has been conducted, it is possible in theory to acquire many images across a single slide providing a large dataset for analysis. However, as both the imaging and analysis processes can be cumbersome and monotonous, fewer images are often taken and analysed. Although efforts should, and usually are, taken to image at various locations across a microscope slide, this is often still not a large enough sample to capture the high variability of comets, as demonstrated in this thesis. Therefore a reliable system that can be left to analyse a large number of images is crucial.

The Mask R-CNN segmentation method that was explored was not as successful in identifying comets compared to both the automated process outlined and the OpenComet software. It was found that this was not due to insufficient computing resources, as training models for longer using a larger number of epochs actually reduced model performance. It may be that a larger image set is required for the network to learn a more extensive set of features that belong to comet assay data, especially due to its highly varied nature. The Monte Carlo model was able to produce images in which comets appeared similar to their experimental counterparts. However, the simulated images did not reproduce the correct elongation of comets, according to the relative damage they represent, although an increase in both tail DNA and asymmetry was observed for increasing strand breaks. This highlights that the crude assumptions the model is built on are a good starting point, but further work is required to more accurately model the movement of DNA through agarose, in order to produce more accurate comet tail elongations. The addition of calibration damage curves from known dose would particularly help benchmark the expected strand break damage for a particular dose, and would aid the improvement of the Monte Carlo.

Bibliography

[1] Nadia Benmoussa et al. “Chimney-sweeps’ cancer—early proof of environmentally driven tumourigenicity”. In: *The Lancet Oncology* 20.3 (2019), p. 338. ISSN: 1470-2045. DOI: https://doi.org/10.1016/S1470-2045(19)30106-8. URL: https://www.sciencedirect.com/science/article/pii/S1470204519301068.

[2] Rudolf Virchow. “As Based upon Physiological and Pathological Histology”. In: *Nutrition Reviews* 47.1 (Jan. 1989), pp. 23–25. ISSN: 0029-6643. DOI: 10.1111/j.1753-4887.1989.tb02747.x. eprint: https://academic.oup.com/nutritionreviews/article-pdf/47/1/23/24085628/nutritionreviews47-0023.pdf. URL: https://doi.org/10.1111/j.1753-4887.1989.tb02747.x.

[3] William S. Halsted. “I. The Results of Radical Operations for the Cure of Carcinoma of the Breast”. In: *Annals of Surgery* 46.1 (1907), pp. 1–19. DOI: 10.1097/00000658-190707000-00001.

[4] R F Mould. “Röntgen and the discovery of X-rays”. In: *The British Journal of Radiology* 68.815 (1995). PMID: 8542220, pp. 1145–1176. DOI: 10.1259/0007-1285-68-815-1145. eprint: https://doi.org/10.1259/0007-1285-68-815-1145. URL: https://doi.org/10.1259/0007-1285-68-815-1145.

[5] Abraham A Sherman. “Translation of an Historic Paper: On a New, Strongly Radioactive Substance, Contained in Pitchblende: By MP Curie, Mme P. Curie and MG Bémont; Presented by M. Becquerel”. In: *Journal of Nuclear Medicine* 11.6 (197), pp. 269–270.

[6] Theodor Boveri. “Concerning the Origin of Malignant Tumours by Theodor Boveri. Translated and annotated by Henry Harris”. In: *Journal of Cell Science* 121.Supplement$_1$ (Jan. 2008), pp. 1–84. ISSN: 0021-9533. DOI: 10.1242/jcs.025742. URL: https://doi.org/10.1242/jcs.025742.

[7] Cancer Research UK. *Cancer incidence statistics*. URL: http://www.cancerresearchuk.org/health-professional/cancer-statistics/incidence#heading-Zero. (accessed: 08.05.2018).

[8] Ralf Dahm. “Friedrich Miescher and the discovery of DNA”. In: *Developmental Biology* 278.2 (2005), pp. 274–288. ISSN: 0012-1606. DOI: https://doi.org/10.1016/j.ydbio.2004.11.028. URL: http://www.sciencedirect.com/science/article/pii/S0012160604008231.

[9] J. D. Watson and F. H. C. Crick. “Molecular Structure of Nucleic Acidas: A Structure for Deoxyribose Nucleic Acid”. In: *Nature* 171 (1953), pp. 737–738. DOI: 10.1038/171737a0.

[10] R. E. Franklin and R. G. Gosling. “The structure of sodium thymonucleate fibres. I. The influence of water content”. In: *Acta Crystallographica* 6.8-9 (1953), pp. 673–677. DOI: 10.1107/S0365110X53001939. URL: https://doi.org/10.1107/S0365110X53001939.

[11] A. Klug. “Rosalind Franklin and the Discovery of the Structure of DNA”. In: *Nature* 219.5156 (1968), pp. 808–810. DOI: 10.1038/219808a0. URL: https://doi.org/10.1038/219808a0.

[12] ROSALIND E. FRANKLIN and R. G. GOSLING. “Molecular Configuration in Sodium Thymonucleate”. In: *Nature* 171.4356 (1953), pp. 740–741. DOI: 10.1038/171740a0. URL: https://doi.org/10.1038/171740a0.

[13] J. Thompson et al. “Rosalind’s X-ray photo of DNA as an undergraduate optical diffraction experiment”. In: *American Journal of Physics* 86.2 (2018), pp. 95–104. DOI: 10.1119/1.5020051. eprint: https://doi.org/10.1119/1.5020051. URL: https://doi.org/10.1119/1.5020051.

[14] D.E. Rooney. *Human Cytogenetics: Constitutional Analysis : A Practical Approach*. Human Cytogenetics. Oxford University Press, 2001. ISBN: 9780199638390. URL: https://books.google.co.uk/books?id=Ohl4WApBeUUC.

[15] Bruce Alberts et al. *Chapter 4: DNA and Chromosomes*. Fourth Edition. New York: Garland Science, 2002, pp. 191–196. ISBN: 0-8153-3218-1.

[16] Cristina M. Alberini and Eric Klann. “Chapter 5 - Regulation of Neuronal Gene Expression and Protein Synthesis”. In: *From Molecules to Networks (Third Edition)*. Ed. by John H. Byrne, Ruth Heidelberger, and M. Neal Waxham. Third Edition. Boston: Academic Press, 2014, pp. 149–174. ISBN: 978-0-12-397179-1. DOI: https://doi.org/10.1016/B978-0-12-397179-1.00005-1. URL: https://www.sciencedirect.com/science/article/pii/B9780123971791000051.

[17] Bruce Alberts et al. *Chapter 17 - The Cell Cycle and Programmed Cell Death*. Fouth Edition. New York: Garland Science, 2002, pp. 983–990. ISBN: 0-8153-3218-1.

[18] Bruce Alberts et al. *Chapter 5 - DNA Replication, Repair and Recombination*. Fouth Edition. New York: Garland Science, 2002, pp. 935–939. ISBN: 0-8153-3218-1.

[19] John F. Ward. “Ionizing Radiation Damage to DNA”. In: *Advances in DNA Damage and Repair: Oxygen Radical Effects, Cellular Protection, and Biological Consequences*. Ed. by Miral Dizdaroglu and Ali Esat Karakaya. Boston, MA: Springer US, 1999, pp. 431–439. ISBN: 978-1-4615-4865-2. DOI: 10.1007/978-1-4615-4865-2-35.

[20] Wei Han and K. N. Yu. “Ionizing Radiation, DNA Double Strand Break and Mutation”. In: *Advances in Genetics Research*. Ed. by Kevin B. Urbano. Vol. 4. New York: Nova Science Publishers, Inc., 2010. Chap. 7. ISBN: 978-1-61728-764-0.

[21] Robert B Daroff et al. *Clinical Neurogenetics*. Elsevier Health Sciences, 2012.

[22] Bruce Alberts et al. *Chapter 6 - How Cells Read the Genome: From DNA to Protein*. Fouth Edition. New York: Garland Science, 2002, pp. 302–364. ISBN: 0-8153-3218-1.

[23] Bruce Alberts et al. *Chapter 23 - Cancer*. Fourth Edition. New York: Garland Science, 2002, pp. 1333–1340. ISBN: 0-8153-3218-1.

[24] Cancer.Net. *The Genetics of Cancer*. URL: https://www.cancer.net/navigating-cancer-care/cancer-basics/genetics/genetics-cancer. (accessed: 03.12.2021).

[25] Amir Mehrgou and Mansoureh Akouchekian. "The importance of BRCA1 and BRCA2 genes mutations in breast cancer development". In: *Medical journal of the Islamic Republic of Iran* 30 (May 2016), pp. 369–369.

[26] M.E. Lomax, L.K. Folkes, and P. O'Neill. "Biological Consequences of Radiation-induced DNA Damage: Relevance to Radiotherapy". In: *Clinical Oncology* 25.10 (2013). Advances in Clinical Radiobiology, pp. 578–585. ISSN: 0936-6555. DOI: https://doi.org/10.1016/j.clon.2013.06.007. URL: http://www.sciencedirect.com/science/article/pii/S0936655513002471.

[27] Keith W. Caldecott. "Single-strand break repair and genetic disease". In: *Nature Reviews Genetics* 9.8 (2008), pp. 619–631. URL: https://doi.org/10.1038/nrg2380.

[28] Caleb Ellicott Finch and Edward L Schneider. *Handbook of the Biology of Aging*. Van Nostrand Reinhold, 1985.

[29] Keith W. Caldecott. "DNA Single-Strand Break Repair and Spinocerebellar Ataxia". In: *Cell* 112.1 (2003), pp. 7–10. ISSN: 0092-8674. DOI: https://doi.org/10.1016/S0092-8674(02)01247-3. URL: https://www.sciencedirect.com/science/article/pii/S0092867402012473.

[30] Arne Klungland et al. "Base Excision Repair of Oxidative DNA Damage Activated by XPG Protein". In: *Molecular Cell* 3.1 (1999), pp. 33–42. ISSN: 1097-2765. DOI: https://doi.org/10.1016/S1097-2765(00)80172-0.

[31] A.A.W.M. van Loon et al. "Induction and repair of DNA single-strand breaks and DNA base damage at different cellular stages of spermatogenesis of the hamster upon in vitro exposure to ionizing radiation". In: *Mutation Research/DNA Repair* 294.2 (1993), pp. 139–148. ISSN: 0921-8777. DOI: https://doi.org/10.1016/0921-8777(93)90022-9.

[32] Nimrat Chatterjee and Graham Walker. "Mechanisms of DNA damage, repair, and mutagenesis". In: *Environmental and molecular mutagenesis* 58.5 (June 2017), pp. 235–263. DOI: 10.1002/em.22087. URL: https://pubmed.ncbi.nlm.nih.gov/28485537.

[33] Tyler L. Fowler et al. "High-throughput detection of DNA double-strand breaks using image cytometry". In: *BioTechniques* 58.1 (2015). PMID: 25605579, pp. 37–39. DOI: 10.2144/000114248. eprint: https://www.future-science.com/doi/pdf/10.2144/000114248. URL: https://www.future-science.com/doi/abs/10.2144/000114248.

[34] Michael R. Lieber. "The Mechanism of Double-Strand DNA Break Repair by the Nonhomologous DNA End-Joining Pathway". In: *Annual Review of Biochemistry* 79.1 (2010). PMID: 20192759, pp. 181–211. DOI: 10.1146/annurev.biochem.052308.093131. eprint: https://doi.org/10.1146/annurev.biochem.052308.093131.

[35] Michael M. Vilenchik and Alfred G. Knudson. “Endogenous DNA double-strand breaks: Production, fidelity of repair, and induction of cancer”. In: *Proceedings of the National Academy of Sciences* 100.22 (2003), pp. 12871–12876. ISSN: 0027-8424. DOI: 10.1073/pnas.2135498100. eprint: https://www.pnas.org/content/100/22/12871.full.pdf. URL: https://www.pnas.org/content/100/22/12871.

[36] Aroumougame Asaithamby and David J. Chen. “Mechanism of cluster DNA damage repair in response to high-atomic number and energy particles radiation”. In: *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 711.1 (2011). From chemistry of DNA damage to repair and biological significance. Comprehending the future, pp. 87–99. ISSN: 0027-5107. DOI: https://doi.org/10.1016/j.mrfmmm.2010.11.002. URL: http://www.sciencedirect.com/science/article/pii/S0027510710003064.

[37] Alan E Tomkinson and Annahita Sallmyr. “Structure and function of the DNA ligases encoded by the mammalian LIG3 gene”. In: *Gene* 531.2 (Dec. 2013), pp. 150–157.

[38] Wendy J Cannan and David S Pederson. “Mechanisms and Consequences of Double-Strand DNA Break Formation in Chromatin”. In: *Journal of cellular physiology* 231.1 (Jan. 2016), pp. 3–14.

[39] NHS. *overview Chemotherapy*. URL: https://www.nhs.uk/conditions/chemotherapy/. (accessed: 03.09.2021).

[40] Wei Han and K. N. Yu. “Response of cells to ionizing radiation”. In: *Advances in Biomedical Science and Engineering*. Ed. by S. C. Tjong. Illnois: Bentham Science Publishes Ltd., 2009. Chap. 6, pp. 204–262.

[41] Philip P Connell and Samuel Hellman. “Advances in radiotherapy and implications for the next century: a historical perspective”. In: *Cancer research* 69.2 (2009), pp. 383–392.

[42] W. C. Röntgen. “Ueber eine neue Art von Strahlen”. In: *Annalen der Physik* 300.1 (1898), pp. 12–17. DOI: https://doi.org/10.1002/andp.18983000103. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/andp.18983000103.

[43] Cancer Treatment Centers of America. *Why does radiation therapy cause skin reactions?* URL: https://www.cancercenter.com/community/blog/2020/11/radiation-therapy-skin-issues. (accessed: 15.12.2021).

[44] Lester J. Peters and K. Kian Ang. “Accelerated Fractionation”. In: *Innovations in Radiation Oncology*. Ed. by H. Rodney Withers and Lester J. Peters. Berlin, Heidelberg: Springer Berlin Heidelberg, 1988, pp. 231–238. ISBN: 978-3-642-83101-0. DOI: 10.1007/978-3-642-83101-0_21. URL: https://doi.org/10.1007/978-3-642-83101-0_21.

[45] H. R. Withers. “Hyperfractionation”. In: *Current Topics in Clinical Radiobiology of Tumors*. Ed. by Hans-Peter Beck-Bornholdt. Berlin, Heidelberg: Springer Berlin Heidelberg, 1993, pp. 45–50. ISBN: 978-3-642-84918-3. DOI: 10.1007/978-3-642-84918-3_4. URL: https://doi.org/10.1007/978-3-642-84918-3_4.

[46] BE Hickey et al. "Hypofractionation for clinically localized prostate cancer". In: *Cochrane Database of Systematic Reviews* 9 (2019). ISSN: 1465-1858. DOI: 10.1002/14651858.CD011462.pub2. URL: https://doi.org//10.1002/14651858.CD011462.pub2.

[47] Heinrich Hertz. "Ueber einen Einfluss des ultravioletten Lichtes auf die electrische Entladung". In: *Annalen der Physik* 267.8 (1887), pp. 983–1000.

[48] Einstein Albert. "On a Heuristic Viewpoint Concerning the Production and Transformation of Light". In: *Annalen der Physik* 17 (1905), pp. 132–148.

[49] Arthur H. Compton. "A Quantum Theory of the Scattering of X-rays by Light Elements". In: *Phys. Rev.* 21 (5 May 1923), pp. 483–502. DOI: 10.1103/PhysRev.21.483. URL: https://link.aps.org/doi/10.1103/PhysRev.21.483.

[50] Patrick Maynard Stuart Blackett, Giuseppe Paolo Stanislao Occhialini, and Ernest Rutherford. "Some photographs of the tracks of penetrating radiation". In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 139.839 (1933), pp. 699–726. DOI: 10.1098/rspa.1933.0048. eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rspa.1933.0048. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rspa.1933.0048.

[51] Gopal B Saha. *Physics and radiobiology of nuclear medicine*. Springer Science & Business Media, 2012.

[52] Steve Webb. *The physics of three dimensional radiation therapy: Conformal radiotherapy, radiosurgery and treatment planning*. CRC Press, 1993.

[53] Robert R. Wilson. "Radiological use of fast protons". In: *Radiology* 47 (1946), pp. 487–491. DOI: 10.1148/47.5.487.

[54] R. E. Zirkle. "Biological Effectiveness of Alpha Particles as a Function of Ion Concentration Produced in their Paths". In: *The American Journal of Cancer* 23.3 (1935), pp. 558–567. DOI: 10.1158/ajc.1935.558. eprint: http://cancerres.aacrjournals.org/content/23/3/558.full.pdf.

[55] Wayne D Newhauser and Rui Zhang. "The physics of proton therapy". In: *Physics in Medicine & Biology* 60.8 (2015), R155. URL: http://stacks.iop.org/0031-9155/60/i=8/a=R155.

[56] W. H. Bragg M.A. and R. Kleeman B.Sc. "XXXIX. On the α particles of radium, and their loss of range in passing through various atoms and molecules". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 10.57 (1905), pp. 318–340. DOI: 10.1080/14786440509463378. eprint: https://doi.org/10.1080/14786440509463378. URL: https://doi.org/10.1080/14786440509463378.

[57] Waldemar SJMAA Ulmer and Evangelos Matsinos. "Theoretical methods for the calculation of Bragg curves and 3D distributions of proton beams". In: *Eur. Phys. J. ST* 190 (2010), pp. 1–81. DOI: 10.1140/epjst/e2010-01335-7. arXiv: 1008.3645 [physics.med-ph].

[58] 2003. URL: https://physicsworld.com/a/how-particles-can-be-therapeutic/.

[59] Rachel J. Carter et al. “Complex DNA Damage Induced by High Linear Energy Transfer Alpha-Particles and Protons Triggers a Specific Cellular DNA Damage Response”. In: *International Journal of Radiation Oncology*Biology*Physics* 100.3 (2017), pp. 776–784. ISSN: 0360-3016. DOI: https://doi.org/10.1016/j.ijrobp.2017.11.012. URL: http://www.sciencedirect.com/science/article/pii/S0360301617341007.

[60] Kyuseok Im et al. “An Introduction to Performing Immunofluorescence Staining”. In: *Methods in molecular biology (Clifton, N.J.)* 1897 (2019), pp. 299–311.

[61] abcam. *Direct vs indirect immunofluorescence*. URL: https://www.abcam.com/secondary-antibodies/direct-vs-indirect-immunofluorescence. (accessed: 9.3.2020).

[62] Catherine M Nickson et al. “Misregulation of DNA damage repair pathways in HPV-positive head and neck squamous cell carcinoma contributes to cellular radiosensitivity”. In: *Oncotarget* 8.18 (May 2017), pp. 29963–29975. ISSN: 1949-2553. DOI: 10.18632/oncotarget.16265. URL: http://europepmc.org/articles/PMC5444717.

[63] Andrew R. Collins. “The comet assay for DNA damage and repair”. In: *Molecular Biotechnology* 26.3 (Mar. 2004), p. 249. ISSN: 1559-0305. DOI: f.

[64] Sambrook J and Russell DW. *Molecular Cloning*. Third Edition. Cold Spring Harbor Laboratory Press, 2001. ISBN: 0-87969-577-3.

[65] Jaaky Narayanan, Jun-Ying Xiong, and Xiang-Yang Liu. “Determination of agarose gel pore size: Absorbance measurements vis a vis other techniques”. In: *Journal of Physics: Conference Series* 28 (Jan. 2006), pp. 83–86. DOI: 10.1088/1742-6596/28/1/017. URL: https://doi.org/10.1088/1742-6596/28/1/017.

[66] G. Toncheva, D. Hadjikinov, and I. Panchev. “Investigation of syneresis of agar jellies with sorbitol”. In: *Food Chemistry* 49.1 (1994), pp. 29–31. ISSN: 0308-8146. DOI: https://doi.org/10.1016/0308-8146(94)90228-3. URL: https://www.sciencedirect.com/science/article/pii/0308814694902283.

[67] Peggy L Olive and Judit P Banáth. “The comet assay: a method to measure DNA damage in individual cells”. In: *Nature protocols* 1.1 (2006), pp. 23–29. ISSN: 1754-2189. DOI: 10.1038/nprot.2006.5. URL: https://doi.org/10.1038/nprot.2006.5.

[68] Peggy L. Olive, Danuta Wlodek, and Judit P. Banáth. “DNA Double-Strand Breaks Measured in Individual Cells Subjected to Gel Electrophoresis”. In: *Cancer Research* 51.17 (1991), pp. 4671–4676. ISSN: 0008-5472. eprint: https://cancerres.aacrjournals.org/content/51/17/4671.full.pdf. URL: https://cancerres.aacrjournals.org/content/51/17/4671.

[69] Peter Möller et al. “Minimum Information for Reporting on the Comet Assay (MIRCA): recommendations for describing comet assay procedures and results”. In: *Nature Protocols* 15.12 (2020), pp. 3817–3826.

[70] O. Ostling and K.J. Johanson. “Microelectrophoretic study of radiation-induced DNA damages in individual mammalian cells”. In: *Biochemical and Biophysical Research Communications* 123.1 (1984), pp. 291–298. ISSN: 0006-291X. DOI: https://doi.org/10.1016/0006-291X(84)90411-X. URL: http://www.sciencedirect.com/science/article/pii/0006291X8490411X.

[71] Narendra P. Singh et al. “A simple technique for quantitation of low levels of DNA damage in individual cells”. In: *Experimental Cell Research* 175.1 (1988), pp. 184–191. ISSN: 0014-4827. DOI: https://doi.org/10.1016/0014-4827(88)90265-0. URL: http://www.sciencedirect.com/science/article/pii/0014482788902650.

[72] Jiang Lei, Wang Yan, and Li Shaoshan. “Application of the comet assay to measure DNA damage induced by UV radiation in the hydrophyte, Spirodela polyrhiza”. In: *Physiologia Plantarum* 129.3 (), pp. 652–657. DOI: 10.1111/j.1399-3054.2006.00820.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1399-3054.2006.00820.x.

[73] T.S. Kumaravel and Awadhesh N. Jha. “Reliable Comet assay measurements for detecting DNA damage induced by ionising radiation and chemicals”. In: *Mutation Research/Genetic Toxicology and Environmental Mutagenesis* 605.1 (2006), pp. 7–16. ISSN: 1383-5718. DOI: https://doi.org/10.1016/j.mrgentox.2006.03.002. URL: http://www.sciencedirect.com/science/article/pii/S1383571806000519.

[74] Johns Hopkins Medicine. *The Importance of HeLa Cells*. URL: https://www.hopkinsmedicine.org/henriettalacks/importance-of-hela-cells.html. (accessed: 2.2.2022).

[75] Rebecca Skloot. *The Immortal Life of Henrietta Lacks*. London, England: Picador, 2019.

[76] Gurvinder Kaur and Jannette M Dufour. “Cell lines: Valuable tools or useless artifacts”. In: *Spermatogenesis* 2.1 (Jan. 2012), pp. 1–5.

[77] NHS. *The Clatterbridge Cancer Centre*. URL: https://www.clatterbridgecc.nhs.uk/. (accessed: 19.10.2022).

[78] Andrzej Kacperek. “Protontherapy of eye tumours in the UK: a review of treatment at Clatterbridge”. In: *Applied Radiation and Isotopes* 67.3 (2009), pp. 378–386.

[79] C. M. Kortman. “Redundancy reduction—A practical method of data compression”. In: 1967. DOI: DOI:10.1109/PROC.1967.5479.

[80] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*. Third Edition. USA: Pearson Prentice-Hall, Inc., 2008. ISBN: 013168728X.

[81] Oge Marques. “Image Compression and Coding”. In: *Encyclopedia of Multimedia*. Ed. by Borko Furht. Boston, MA: Springer US, 2008, pp. 318–323. ISBN: 978-0-387-78414-4. DOI: 10.1007/978-0-387-78414-4_342. URL: https://doi.org/10.1007/978-0-387-78414-4_342.

[82] Rafael C. Gonzalez and Richard E. Woods. “Chapter 8: Image Compression”. In: *Digital Image Processing*. Third Edition. USA: Pearson Prentice-Hall, Inc., 2008, pp. 525–529. ISBN: 013168728X.

[83] J Wang and S Langer. “A brief review of human perception factors in digital displays for picture archiving and communications systems”. In: *Journal of digital imaging* 10.4 (Nov. 1997), pp. 158–168.

[84] Hamed Rahimi-Nasrabadi et al. "Image luminance changes contrast sensitivity in visual cortex". In: *Cell Reports* 34.5 (2021), p. 108692. ISSN: 2211-1247. DOI: https://doi.org/10.1016/j.celrep.2021.108692. URL: https://www.sciencedirect.com/science/article/pii/S221112472100005X.

[85] Richard H. Wiggins et al. "Image File Formats: Past, Present, and Future". In: *RadioGraphics* 21.3 (2001). PMID: 11353125, pp. 789–798. DOI: 10.1148/radiographics.21.3.g01ma25789. eprint: https://doi.org/10.1148/radiographics.21.3.g01ma25789. URL: https://doi.org/10.1148/radiographics.21.3.g01ma25789.

[86] N. Ahmed, T. Natarajan, and K.R. Rao. "Discrete Cosine Transform". In: *IEEE Transactions on Computers* C-23.1 (1974), pp. 90–93. DOI: 10.1109/T-C.1974.223784.

[87] Nasir Ahmed. "How I came up with the discrete cosine transform". In: *Digit. Signal Process.* 1 (1991), pp. 4–5.

[88] John Miano. *Compressed Image File Formats: JPEG, PNG, GIF, XBM, BMP*. USA: ACM Press/Addison-Wesley Publishing Co., 1999. ISBN: 0201604434.

[89] Aldus Developers Desk. "TIFF". In: (1986).

[90] Sonal Chawla, Meenakshi Beri, and Ritu Mudgil. "Image compression techniques: a review". In: *International Journal of Computer Science and Mobile Computing* 3.8 (2014), pp. 291–296.

[91] Arthur Edelstein et al. "Computer Control of Microscopes Using µManager". In: *Current Protocols in Molecular Biology* 92.1 (2010), pp. 14.20.1–14.20.17. DOI: https://doi.org/10.1002/0471142727.mb1420s92. eprint: https://currentprotocols.onlinelibrary.wiley.com/doi/pdf/10.1002/0471142727.mb1420s92. URL: https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/0471142727.mb1420s92.

[92] T. S. Kumaravel et al. "Comet Assay measurements: a perspective". In: *Cell Biology and Toxicology* 25.1 (Feb. 2009), pp. 53–64. ISSN: 1573-6822. DOI: 10.1007/s10565-007-9043-9. URL: https://doi.org/10.1007/s10565-007-9043-9.

[93] Peter Møller et al. "On the search for an intelligible comet assay descriptor". In: *Frontiers in Genetics* 5 (2014), p. 217. ISSN: 1664-8021. DOI: 10.3389/fgene.2014.00217. URL: https://www.frontiersin.org/article/10.3389/fgene.2014.00217.

[94] hCOMET. *The comet assay as a human biomonitoring tool.* URL: https://www.hcomet.eu/. (accessed: 21.10.2021).

[95] Susan Elmore. "Apoptosis: a review of programmed cell death". In: *Toxicologic pathology* 35.4 (June 2007), pp. 495–516.

[96] Ezgi Eyluel Bankoglu, Carolin Schuele, and Helga Stopper. "Cell survival after DNA damage in the comet assay". In: *Archives of Toxicology* 95.12 (2021), pp. 3803–3813.

[97] Takahiro Kyoya et al. "The effect of different methods and image analyzers on the results of the in vivo comet assay". In: *Genes and Environment* 40.1 (2018), p. 4.

[98] R. R. Tice et al. “Single cell gel/comet assay: Guidelines for in vitro and in vivo genetic toxicology testing”. In: *Environmental and Molecular Mutagenesis* 35.3 (2000), pp. 206–221. doi: https://doi.org/10.1002/(SICI)1098-2280(2000)35:3<206::AID-EM8>3.0.CO;2-J. url: https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291098-2280%282000%2935%3A3%3C206%3A%3AAID-EM8%3E3.0.CO%3B2-J.

[99] Amaya Azqueta et al. “Towards a more reliable comet assay: Optimising agarose concentration, unwinding time and electrophoresis conditions”. In: *Mutation Research/Genetic Toxicology and Environmental Mutagenesis* 724.1 (2011), pp. 41–45. issn: 1383-5718. doi: https://doi.org/10.1016/j.mrgentox.2011.05.010. url: https://www.sciencedirect.com/science/article/pii/S1383571811001665.

[100] José Manuel Enciso et al. “Does the duration of lysis affect the sensitivity of the in vitro alkaline comet assay?” In: *Mutagenesis* 30.1 (Dec. 2014), pp. 21–28. issn: 0267-8357. doi: 10.1093/mutage/geu047. eprint: https://academic.oup.com/mutage/article-pdf/30/1/21/12783619/geu047.pdf. url: https://doi.org/10.1093/mutage/geu047.

[101] Clara Johansson et al. “An ECVAG† trial on assessment of oxidative damage to DNA measured by the comet assay”. In: *Mutagenesis* 25.2 (Nov. 2009), pp. 125–132. issn: 0267-8357. doi: 10.1093/mutage/gep055. eprint: https://academic.oup.com/mutage/article-pdf/25/2/125/3354526/gep055.pdf. url: https://doi.org/10.1093/mutage/gep055.

[102] M.V.M. Lafleur et al. “Alkali-labile Sites and Post-irradiation Effects in Single-stranded DNA Induced by H Radicals”. In: *International Journal of Radiation Biology and Related Studies in Physics, Chemistry and Medicine* 33.3 (1978), pp. 273–281. doi: 10.1080/09553007814550151. eprint: https://doi.org/10.1080/09553007814550151. url: https://doi.org/10.1080/09553007814550151.

[103] Clara Ersson and Lennart Möller. “The effects on DNA migration of altering parameters in the comet assay protocol such as agarose density, electrophoresis conditions and durations of the enzyme or the alkaline treatments”. In: *Mutagenesis* 26.6 (July 2011), pp. 689–695. issn: 0267-8357. doi: 10.1093/mutage/ger034. eprint: https://academic.oup.com/mutage/article-pdf/26/6/689/3862690/ger034.pdf. url: https://doi.org/10.1093/mutage/ger034.

[104] Nikolai P. Sirota et al. “Some causes of inter-laboratory variation in the results of comet assay”. In: *Mutation Research/Genetic Toxicology and Environmental Mutagenesis* 770 (2014), pp. 16–22. issn: 1383-5718. doi: https://doi.org/10.1016/j.mrgentox.2014.05.003. url: https://www.sciencedirect.com/science/article/pii/S1383571814001508.

[105] Judit P. Banáth, Andrew Kim, and Peggy L. Olive. “Overnight Lysis Improves the Efficiency of Detection of DNA Damage in the Alkaline Comet Assay”. In: *Radiation Research* 155.4 (2001), pp. 564–571. issn: 00337587, 19385404. url: http://www.jstor.org/stable/3580359.

[106] Arlene L Oei et al. “Effects of hyperthermia on DNA repair pathways: one treatment to inhibit them all”. In: *Radiation oncology (London, England)* 10 (Aug. 2015), pp. 165–165.

[107] E. Dikomey and J. Franzke. "Effect of Heat on Induction and Repair of DNA Strand Breaks in X-irradiated CHO Cells". In: *International Journal of Radiation Biology* 61.2 (1992), pp. 221–233. DOI: 10.1080/09553009214550851. eprint: https://doi.org/10.1080/09553009214550851. URL: https://doi.org/10.1080/09553009214550851.

[108] Sergey A Shaposhnikov et al. "Single-cell gel electrophoresis (the comet assay): Loops or fragments?" In: *electrophoresis* 29.14 (2008), pp. 3005–3012.

[109] Nancy C Stellwagen and Earle Stellwagen. "Effect of the matrix on DNA electrophoretic mobility". In: *Journal of chromatography. A* 1216.10 (Mar. 2009), pp. 1917–1929.

[110] Peggy L. Olive. "The Comet Assay". In: *In Situ Detection of DNA Damage: Methods and Protocols*. Ed. by Vladimir V. Didenko. Totowa, NJ: Humana Press, 2002, pp. 179–194. ISBN: 978-1-59259-179-4. DOI: 10.1385/1-59259-179-5:179. URL: https://doi.org/10.1385/1-59259-179-5:179.

[111] Gunnar Brunborg, Linn Rolstadaas, and Kristine B. Gutzkow. "Electrophoresis in the Comet Assay". In: *Electrophoresis*. Ed. by Oana-Maria Boldura and Cornel Baltă. Rijeka: IntechOpen, 2018. Chap. 4. DOI: 10.5772/intechopen.76880. URL: https://doi.org/10.5772/intechopen.76880.

[112] A. Hartmann et al. "Recommendations for conducting the in vivo alkaline Comet assay". In: *Mutagenesis* 18.1 (Jan. 2003), pp. 45–51. ISSN: 0267-8357. DOI: 10.1093/mutage/18.1.45. eprint: https://academic.oup.com/mutage/article-pdf/18/1/45/9895053/180045.pdf. URL: https://doi.org/10.1093/mutage/18.1.45.

[113] Hubert Zipper et al. "Investigations on DNA intercalation and surface binding by SYBR Green I, its structure determination and methodological implications". In: *Nucleic Acids Research* 32.12 (Jan. 2004), e103–e103. ISSN: 0305-1048. DOI: 10.1093/nar/gnh101. eprint: https://academic.oup.com/nar/article-pdf/32/12/e103/7038026/gnh101.pdf. URL: https://doi.org/10.1093/nar/gnh101.

[114] Robert Tebbs et al. "Modification of the Comet assay for the detection of DNA strand breaks in extremely small tissue samples". In: *Mutagenesis* 14 (Aug. 1999), pp. 437–8.

[115] S. Braafladt, Vytas Reipa, and D. H. Atha. "The Comet Assay: Automated Imaging Methods for Improved Analysis and Reproducibility". In: *Scientific Reports* 6 (2016). DOI: 10.1038/srep32162. URL: https://www.nature.com/articles/srep32162.

[116] Ralph Jacobson et al. *Manual of Photography*. Routledge, 2013.

[117] Andrew R. Collins et al. "The comet assay: topical issues". In: *Mutagenesis* 23.3 (2008), pp. 143–151. DOI: 10.1093/mutage/gem051. URL: http://dx.doi.org/10.1093/mutage/gem051.

[118] W. Frieauff, A. Hartmann, and W. Suter. "Automatic analysis of slides processed in the Comet assay". In: *Mutagenesis* 16.2 (2001), pp. 133–137. DOI: 10.1093/mutage/16.2.133. URL: http://dx.doi.org/10.1093/mutage/16.2.133.

[119] Oxford Instruments Andor. *Komet 7 and 7-GLP*. URL: http://www.andor.com/scientific-software/komet-software. (accessed: 13/05/2018).

[120] Benjamin M. Gyori et al. "OpenComet: an automated tool for comet assay image analysis". In: *Redox Biology* 2 (2014), pp. 457–465.

[121] Yiyu Hong et al. "Deep learning method for comet segmentation and comet assay image analysis". In: *Scientific Reports* 10.1 (2020), p. 18915.

[122] Srikanth Namuduri et al. "Automated quantification of DNA damage via deep transfer learning based analysis of comet assay images". In: *Applications of Machine Learning*. Ed. by Michael E. Zelinski et al. Vol. 11139. International Society for Optics and Photonics. SPIE, 2019, pp. 256–262. DOI: 10.1117/12.2529352. URL: https://doi.org/10.1117/12.2529352.

[123] Ümit Atila et al. "Classification of DNA damages on segmented comet assay images using convolutional neural network". In: *Computer Methods and Programs in Biomedicine* 186 (2020), p. 105192. ISSN: 0169-2607. DOI: https://doi.org/10.1016/j.cmpb.2019.105192. URL: https://www.sciencedirect.com/science/article/pii/S0169260719300707.

[124] Riccardo Rosati et al. "Faster R-CNN approach for detection and quantification of DNA damage in comet assay images". In: *Computers in Biology and Medicine* 123 (2020), p. 103912. ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2020.103912. URL: https://www.sciencedirect.com/science/article/pii/S0010482520302559.

[125] Nobuyuki Otsu. "A Threshold Selection Method from Gray-Level Histograms". In: *IEEE Transactions on Systems, Man, and Cybernetics* 9.1 (1979), pp. 62–66. DOI: 10.1109/TSMC.1979.4310076.

[126] C.H. Li and C.K. Lee. "Minimum cross entropy thresholding". In: *Pattern Recognition* 26.4 (1993), pp. 617–625. ISSN: 0031-3203. DOI: https://doi.org/10.1016/0031-3203(93)90115-D. URL: https://www.sciencedirect.com/science/article/pii/003132039390115D.

[127] Reuven Y. Rubinstein and Dirk P. Kroese. "A Tutorial Introduction to the Cross-Entropy Method". In: *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*. New York, NY: Springer New York, 2004, pp. 29–58. ISBN: 978-1-4757-4321-0. DOI: 10.1007/978-1-4757-4321-0_2. URL: https://doi.org/10.1007/978-1-4757-4321-0_2.

[128] G W Zack, W E Rogers, and S A Latt. "Automatic measurement of sister chromatid exchange frequency." In: *Journal of Histochemistry & Cytochemistry* 25.7 (1977). PMID: 70454, pp. 741–753. DOI: 10.1177/25.7.70454. URL: https://doi.org/10.1177/25.7.70454.

[129] C.H. Li and P.K.S. Tam. "An iterative algorithm for minimum cross entropy thresholding". In: *Pattern Recognition Letters* 19.8 (1998), pp. 771–776. ISSN: 0167-8655. DOI: https://doi.org/10.1016/S0167-8655(98)00057-9. URL: https://www.sciencedirect.com/science/article/pii/S0167865598000579.

[130] Haniza Yazid et al. "Performance analysis of entropy thresholding for successful image segmentation". In: *Multimedia Tools and Applications* (2022).

[131] T. W. Ridler and S. Calvard. "Picture Thresholding Using an Iterative Selection Method". In: *IEEE Transactions on Systems, Man, and Cybernetics* 8.8 (1978), pp. 630–632. DOI: 10.1109/TSMC.1978.4310039.

[132] Jui-Cheng Yen, Fu-Juay Chang, and Shyang Chang. "A new criterion for automatic multilevel thresholding". In: *IEEE Transactions on Image Processing* 4.3 (1995), pp. 370–378. DOI: 10.1109/83.366472.

[133] Payel Roy et al. "Adaptive thresholding: A comparative study". In: *2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*. 2014, pp. 1182–1186. DOI: 10.1109/ICCICCT.2014.6993140.

[134] Martin Ester et al. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. KDD'96. Portland, Oregon, 1996, pp. 226–231.

[135] SP Lloyd. "Least square quantization in PCM. Bell Telephone Laboratories Paper. Published in journal much later: Lloyd, SP: Least squares quantization in PCM". In: *IEEE Trans. Inform. Theor.(1957/1982)* 18 (1957), p. 11.

[136] James MacQueen et al. "Some methods for classification and analysis of multivariate observations". In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1. 14. Oakland, CA, USA. 1967, pp. 281–297.

[137] Nadia Rahmah and Imas Sukaesih Sitanggang. "Determination of Optimal Epsilon (Eps) Value on DBSCAN Algorithm to Clustering Data on Peatland Hotspots in Sumatra". In: *IOP Conference Series: Earth and Environmental Science* 31 (Jan. 2016), p. 012012. DOI: 10.1088/1755-1315/31/1/012012. URL: https://doi.org/10.1088/1755-1315/31/1/012012.

[138] Jadwiga Rogowska. "5 - Overview and Fundamentals of Medical Image Segmentation". In: *Handbook of Medical Imaging*. Ed. by ISAAC N. BANKMAN. Biomedical Engineering. San Diego: Academic Press, 2000, pp. 69–85. ISBN: 978-0-12-077790-7. DOI: https://doi.org/10.1016/B978-012077790-7/50009-6. URL: https://www.sciencedirect.com/science/article/pii/B9780120777907500096.

[139] Bernhard Preim and Charl Botha. "Chapter 4 - Image Analysis for Medical Visualization". In: *Visual Computing for Medicine (Second Edition)*. Ed. by Bernhard Preim and Charl Botha. Second Edition. Boston: Morgan Kaufmann, 2014, pp. 111–175. ISBN: 978-0-12-415873-3. DOI: https://doi.org/10.1016/B978-0-12-415873-3.00004-3. URL: https://www.sciencedirect.com/science/article/pii/B9780124158733000043.

[140] Yingbo Liang and Jian Fu. "Watershed Algorithm for Medical Image Segmentation Based on Morphology and Total Variation Model". In: *International Journal of Pattern Recognition and Artificial Intelligence* 33.05 (2019), p. 1954019. DOI: 10.1142/S0218001419540193. eprint: https://doi.org/10.1142/S0218001419540193. URL: https://doi.org/10.1142/S0218001419540193.

[141] Warren S. McCulloch and Walter Pitts. "A logical calculus of the ideas immanent in nervous activity". In: *The bulletin of mathematical biophysics* 5.4 (1943), pp. 115–133.

[142] Frank Rosenblatt. "The perceptron: a probabilistic model for information storage and organization in the brain." In: *Psychological review* 65.6 (1958), p. 386.

[143] Laveen N. Kanal. “Perceptron”. In: *Encyclopedia of Computer Science*. GBR: John Wiley and Sons Ltd., 2003, pp. 1383–1385. ISBN: 0470864125.

[144] Kaiming He et al. *Mask R-CNN*. 2018. arXiv: 1703.06870 [cs.CV].

[145] Shaoqing Ren et al. *Faster R-CNN:s Towards Real-Time Object Detection with Region Proposal Networks*. 2016. arXiv: 1506.01497 [cs.CV].

[146] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV].

[147] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2015. arXiv: 1409.1556 [cs.CV].

[148] Di Tian et al. “Review of object instance segmentation based on deep learning”. In: *Journal of Electronic Imaging* 31.4 (2021), pp. 1–18. DOI: 10.1117/1.JEI.31.4.041205. URL: https://doi.org/10.1117/1.JEI.31.4.041205.

[149] Waleed Abdulla. *Splash of Color: Instance Segmentation with Mask R-CNN and TensorFlow*. URL: https://engineering.matterport.com/splash-of-color-instance-segmentation-with-mask-r-cnn-and-tensorflow-7c761e238b46. (accessed: 11.2.2019).

[150] Abhishek Dutta and Andrew Zisserman. “The VGG Image Annotator (VIA)”. In: *arXiv preprint arXiv:1904.10699* (2019).

[151] A. Dutta, A. Gupta, and A. Zissermann. *VGG Image Annotator (VIA)*. Version: 1.0.6, Accessed: Feb 2019. 2016. URL: http://www.robots.ox.ac.uk/~vgg/software/via/.

[152] Bryan C. Russell et al. “LabelMe: A Database and Web-Based Tool for Image Annotation”. In: *International Journal of Computer Vision* 77.1 (2008), pp. 157–173.

[153] Sinno Jialin Pan and Qiang Yang. “A survey on transfer learning”. In: *IEEE Transactions on knowledge and data engineering* 22.10 (2009), pp. 1345–1359.

[154] Tsung-Yi Lin et al. “Microsoft COCO: Common Objects in Context”. In: *CoRR* abs/1405.0312 (2014). arXiv: 1405.0312. URL: http://arxiv.org/abs/1405.0312.

[155] Stéfan van der Walt et al. “scikit-image: image processing in Python”. In: *PeerJ* 2 (June 2014), e453. ISSN: 2167-8359. DOI: 10.7717/peerj.453. URL: https://doi.org/10.7717/peerj.453.

[156] Waleed Abdulla. *Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow*. 2017. URL: https://github.com/matterport/Mask%5C_RCNN.

[157] Ekaba Bisong. “Google Colaboratory”. In: *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*. Berkeley, CA: Apress, 2019, pp. 59–64. ISBN: 978-1-4842-4470-8. DOI: 10.1007/978-1-4842-4470-8_7. URL: https://doi.org/10.1007/978-1-4842-4470-8_7.

[158] Liang-Kai Huang and Mao-Jiun J. Wang. “Image thresholding by minimizing the measures of fuzziness”. In: *Pattern Recognition* 28.1 (1995), pp. 41–51. ISSN: 0031-3203. DOI: https://doi.org/10.1016/0031-3203(94)E0043-K. URL: https://www.sciencedirect.com/science/article/pii/0031320394E0043K.

[159] D.B. Goldman and Jiun-Hung Chen. "Vignette and exposure calibration and compensation". In: *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*. Vol. 1. 2005, pp. 899–906. DOI: 10.1109/ICCV.2005.249.

[160] Jonathan J M Landry et al. "The genomic and transcriptomic landscape of a HeLa cell line". In: *G3 (Bethesda, Md.)* 3.8 (Aug. 2013), pp. 1213–1224.

[161] S Y Brendler-Schwaab et al. "Cells of different tissues for in vitro and in vivo studies in toxicology: Compilation of isolation methods." In: *Toxicol In Vitro* 8.6 (Dec. 1994), pp. 1285–1302.

[162] C.R. Calladine et al. "A study of electrophoretic mobility of DNA in agarose and polyacrylamide gels". In: *Journal of Molecular Biology* 221.3 (1991), pp. 981–1005. ISSN: 0022-2836. DOI: https://doi.org/10.1016/0022-2836(91)80187-Y. URL: https://www.sciencedirect.com/science/article/pii/002228369180187Y.

[163] Fisher Scientific. *Fisherbrand Power Supply*. URL: https://www.fishersci.ie/shop/products/power-supply-29/12613546. (accessed: 19.8.2021).

[164] Olympus. *Industrial Microscopes BX61*. URL: https://www.olympus-ims.com/en/microscope/bx61-2/#!cms[focus]=cmsContent13561. (accessed: 19.8.2021).

[165] Hamamatsu Photonics. *ORCA-Flash4.0 LT PLUS C11440-42U30 Specifications*. URL: https://www.hamamatsu.com/content/dam/hamamatsu-photonics/sites/documents/99_SALES_LIBRARY/sys/SCAS0122E_C11440-42U30.pdf. (accessed: 19.8.2021).

Appendix A

Comet Assay Experimental Details

A.1 Comet Assay protocol

The following is an example protocol for conducting an alkaline Comet Assay study on WI-38 lung fibroblast cells, to observe SSB damage.

Agarose

- Normal melting point for 1% agarose in dH_2O (1 g in 100 ml)
- Low melting point agarose (1%) in PBS (1g in 100 ml)

Lysis Buffer

- Solution of dH_2O (800 ml) containing NaCl (146.1 g, 2.5 M), EDTA disodium salt (37.2 g, 100mM) and Tris base (1.2g, 10mM). Set to pH 10.5 by the addition of NaOH (8 g) and 10 M NaOH dropwise. Adjust to 1 L and store at 4° C.
- Prepare complete solution just prior to use by adding a mixture of 1 ml DMSO and 1 ml Triton X-100 (detergent that breaks cells down) to 98 ml cold lysis buffer

Electrophoresis Buffer

- 300 mM NaOH, 1 mM EDTA and 1% DMSO, prepared just prior to use.
- NaOH (100 g in 500 ml, 5M) and EDTA (14.8896 g in 200 ml, 200 mM)

- 75 ml 5 M NaOH, 6.25 ml 0.2 M EDTA and 12.5 ml DMSO make up to 1.25 L with dH_2O. Check pH > 13 and store at 4° C until required.

Neutralisation Buffer

- Tris-HCl (12.11 g in 200 ml, 500 mM), pH 8.0 (5M HCl)

Staining Solution

- Dilute SYBR Gold 1 in 20,000 in dH_2O, pH 8.0 (1 $\mu m/20ml$)

Comet Assay Protocol

- Prepare slides by adding 800 μl molten normal melting point agarose to a slide, add a coverslip and leave agarose to set. Remove coverslip and air dry slides overnight.
- Trypsinise exponentially growing cells and dilute 100,000 cells/ml. (Trypsin lifts a mono-layer of cells off of the plastic they have been grown on).
- Prepare 1 mM H_2O_2 stock (1.14 μl) 8.8 M H_2O_2 in 10 ml PBS). H_2O_2 causes damage that is repaired by BER (base excision repair), which is the main repair mechanism of oxidative damage.
- Add 250 μl cell suspension per well of a 24-well plate on ice to prevent repair (damage happens faster with no repair) and adhesion. Add 2.5 μl of 1 mM H_2O_2 (10 μM dose) and incubate on ice for 5 minutes.
- Add 1 ml low melting point agarose (37° C) to cells, mix and add 1 ml to NMPA coated slide. Add coverslip and place on ice tray for 2 mins.

For repair studies place slides in a humidified chamber (so cells don't dry out) at 37° *C for required time points.*

- Remove coverslip and place in coplin jars containing fresh cold lysis buffer and lyse cells for at least 1 h (or can be left overnight if needed) at 4° C.
- Transfer slides to electrophoresis tank, cover with 1.25 L fresh electrophoresis buffer (4° C) and allow DNA to unwind for 30 mins.
- Electrophorese at 25 V for 25 mins, adjusting current to 300 mA by adding or removing buffer (sometimes needs to go up to 400 mA as more buffer cannot be removed without covering slides).

- Remove slides from tank and cover with 1 ml neutralisation buffer for 5 min. Repeat twice.
- Allow slides to dry overnight.
- Rehydrate cells in dH_2O (pH 8.0) for 30 mins and add 1 ml SYBR Gold to each slide for 30 mins.
- Remove excess stain (pipetted back into stain bottle - cells may be put into bottle, potentially contaminating the stain) and dry slides overnight prior to analysis/storage.

A.2 Comet Assay Imaging Equipment

A following is a list of imaging equipment used to collect the majority of data in this work. Part of the data was acquired with a similar setup, but with a lower resolution camera.

- Power supply unit (Fisher Scientific, Loughborough, UK; Cat. No.: 12613546) [163]
- Fluorescent microscope Olympus BX61 (Olympus, Hamburg, Germany) [164]
- Digital sCmos camera Hamamatsu ORCA-Flash 4.0 LT C11440-42U (Hamamatsu Photonics UK Limited, Hertfordshire, UK) [165]
- Komet 6.0 image analysis software (Andor Technology, Belfast, Northern Ireland) [119]

Appendix B

Additional Measurements and Data

B.1 Confusion Matrices

The following figures show confusion matrices for the original watershed , overall expanded and overall plus additional tail expanded masks for image A, image B and an average of all comparison images that were omitted in section 4.3.

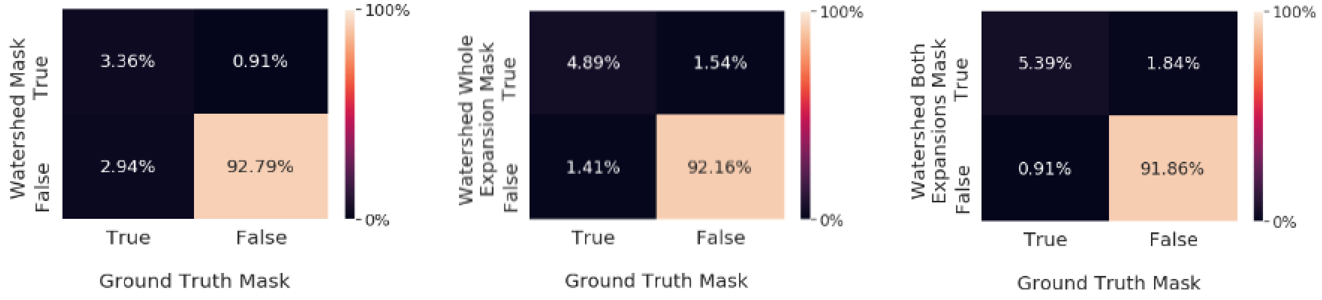


Figure B.1: *The confusion matrices for the original watershed masks, the whole comet expanded masks and the whole comet plus tail expanded masks of image A.*

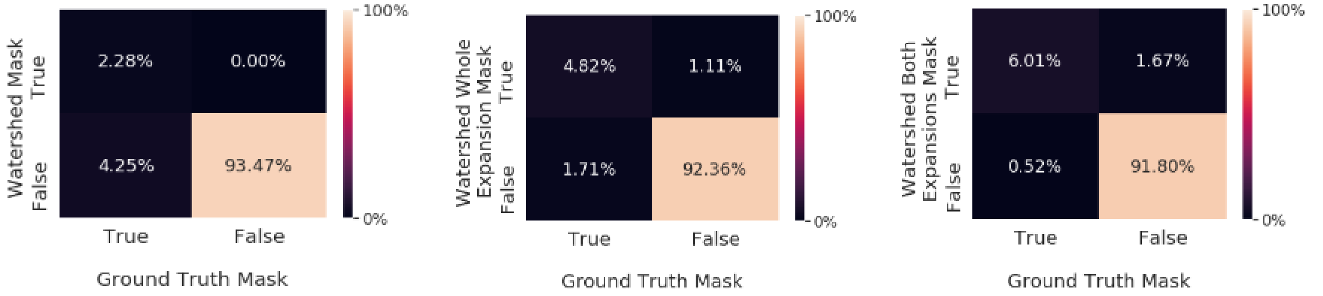


Figure B.2: *The confusion matrices for the original watershed masks, the whole comet expanded masks and the whole comet plus tail expanded masks of image B.*

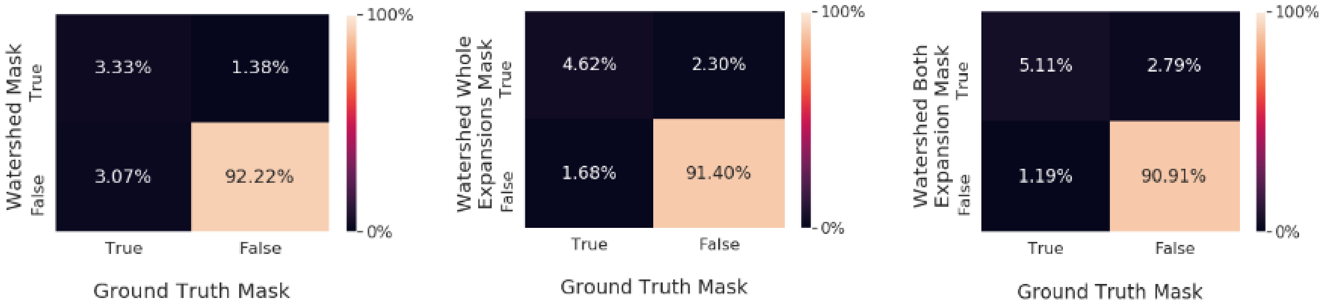


Figure B.3: *The confusion matrices for the original watershed masks, the whole comet expanded masks and the whole comet plus tail expanded masks averaged across all 20 images.*

B.2 Tail Length for Varying Image Contrast

The figures, omitted from section 5.2.5, showing the minimal tail length changes for varying contrast are shown below for exposure values 3 000, 5 000 and 9 000.

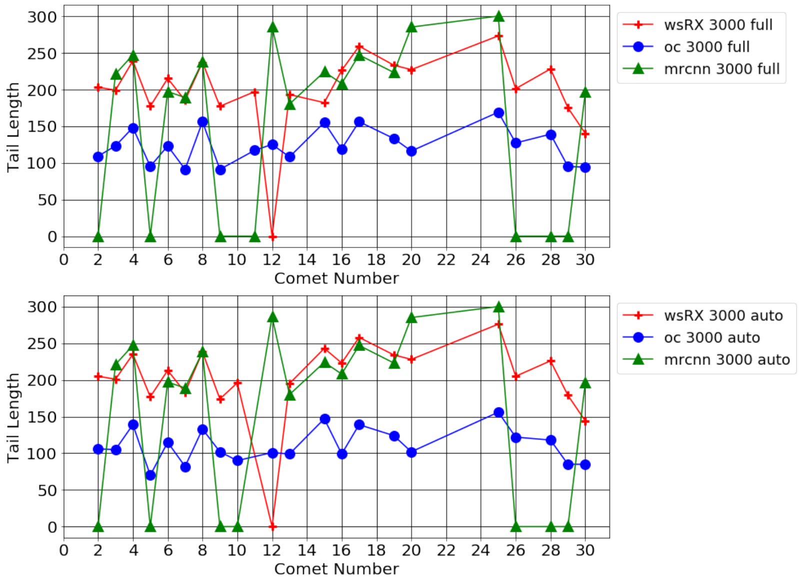


Figure B.4: *Tail length measures for comets detected using the expanded watershed algorithm (wsRX), OpenComet (oc) and the Mask R-CNN (mrcnn) model using the auto and full contrast functions at exposure value 3000.*

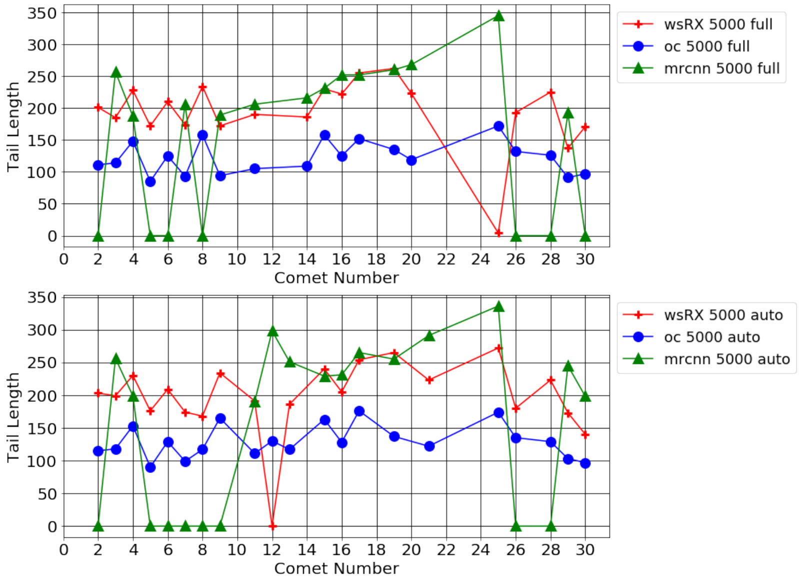


Figure B.5: *Tail length measures for comets detected using the expanded watershed algorithm (wsRX), OpenComet (oc) and the Mask R-CNN (mrcnn) model using the auto and full contrast functions at exposure value 5000.*

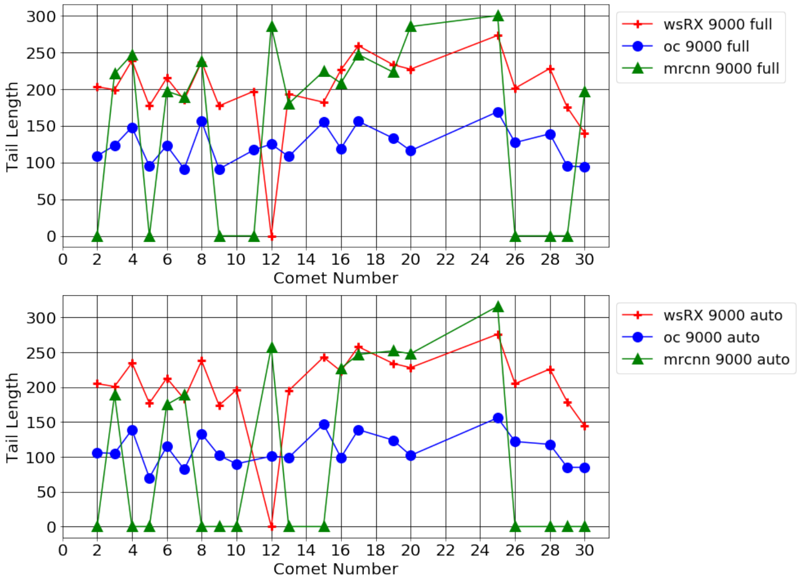


Figure B.6: *Tail length measures for comets detected using the expanded watershed algorithm (wsRX), OpenComet (oc) and the Mask R-CNN (mrcnn) model using the auto and full contrast functions at exposure value 9000.*

B.3 Average Background Measurements

When calculating the background intensity to subtract for each comet, an average was found from the additional expanded region around the comet as described in section 5.2.6. The average background value found for each comet across all analysed images is shown in the table below.

Image	Average	Minimum	Maximum	Number of Comets
1	0	0	1	10
2	0	0	0	10
3	0	0	0	14
4	0	0	0	4
5	0	0	0	10
6	0	0	0	10
7	0	0	0	9
8	0	0	1	11
9	6	0	18	23
10	7	0	22	21
11	6	0	18	15
12	6	0	16	29
13	9	0	18	18
14	3	0	13	27
15	9	0	19	18
16	6	0	11	20
17	1	0	2	12
18	0	0	3	13
19	0	0	0	21
20	0	0	0	12

Table B.1: *The background intensity measurements found for each comet in the 20 images that were analysed.*

B.4 Linear Fit Analysis

Figure B.7 shows the linear fit model. The χ^2 value per degree of freedom is 1.09, indicating the model represents the data well, but the quadratic fit had a marginally better agreement with the data, with a χ^2 value of 1.09 per degree of freedom.

The linear fit parameters are shown in Table B.2.

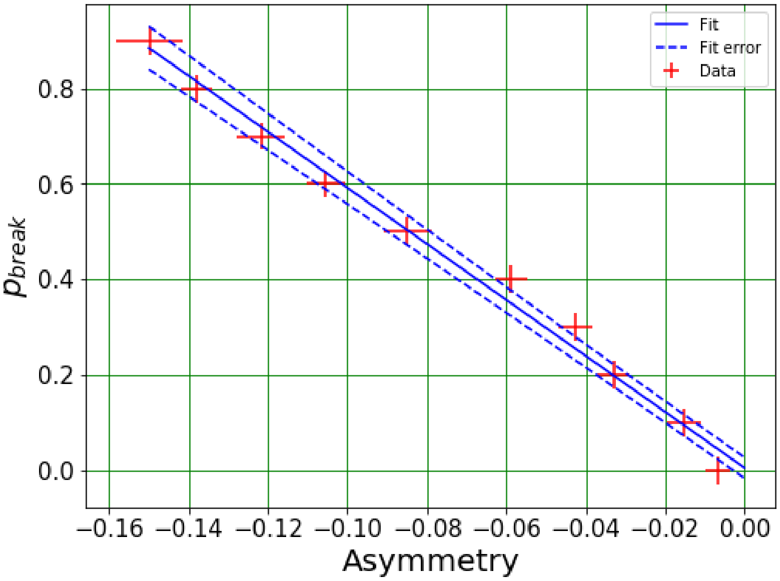


Figure B.7: *The model and its associated error plotted with the asymmetry data used to create it.*

Parameter	Value	Error
a	0.006	0.021
b	-5.857	0.263

Table B.2: *A summary of the parameters found when applying a linear fit to the asymmetry values of the MC data. All values have been given to 3 decimal places.*

Appendix C

Source Code

C.1 Monte Carlo

The source code used to create simulated comet Assay images can be found at https://github.com/Selina9/The-Modelling-and-Analysis-of-Comet-Assay-Data.