Applying a deep neural network-based approach to automating the micronucleus (MN) assay

Researched by Qiellor Haxhiraj (BSc (Hons)), submitted to Swansea University in fulfilment of the requirements for the Degree of MSc in Medical and Health Care Studies by Research.

Swansea University

2020

Swansea University
Prifysgol Abertawe

Summary:

The Micronucleus (MN) Assay is a test mandated for use in genetic toxicology testing by regulatory bodies such as the Food and Drug administration (FDA). An increased quantity of MN is an indication of chromosomal damage which can be characterised into chromosomal breakage (caused by a clastogen) and chromosomal loss (caused by an aneugen). By comparing a dose response, estimates can be made into the potency of the chemical. Historically the cell scoring procedure takes place through the 'gold standard' of manual scoring by light microscopy following staining. However, despite being classed the gold standard, this method is laborious and subjective, with archiving of results not a possibility.

This leads to the need to develop a new technique to streamline the process, whilst still maintaining accuracy. The result is the creation of a ground truth based deep learning algorithm. By using imaging flow cytometry to carry out the MN assay, a ground truth was created, consisting of different cellular types, including MN. By scoring these images manually by eye, a ground truth of images to teach the deep-learning algorithm is created.

By applying a deep neural network, the algorithm uses multiple layers to differentiate information, mimicking the way neurons work in the brain. This approach allows for differentiation between different cellular types based on the ground truth images scored. By assessing more images, the accuracy is further increased. This is advantageous as a MN count is generated directly after processing the imaging flow cytometry file. This streamlines the process completely whilst maintaining accuracy. Also, by using three different laboratory datasets in the production of the ground truth, application was shown to be accurate for cross-laboratory use, a novelty in this research setting.

This allows for the existing ground truth to be used for future MN scoring, allowing for the MN assay to be fully automated.

DECLARATION

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed QIELLOR HAXHIRAJ (candidate)

Date 30/09/20

STATEMENT 1

This thesis is the result of my own investigations, except where otherwise stated. Where correction services have been used, the extent and nature of the correction is clearly marked in a footnote(s).

Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed QIELLOR HAXHIRAJ (candidate)

Date 30/09/20

STATEMENT 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed QIELLOR HAXHIRAJ (candidate)

Date 30/09/20

NB: *Candidates on whose behalf a bar on access has been approved by the University (see Note 7), should use the following version of Statement 2:*

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loans **after expiry of a bar on access approved by the Swansea University.**

Signed QIELLOR HAXHIRAJ (candidate)

Date 30/09/20

Contents

List of Figures

Tables

Abbreviations

CBMN: Cytokinesis-b Micronucleus Assay

Cyto-B: Cytochalasin

LOGEL: Lowest-observed genotoxic effect level

MGMT: O^6-MethylGuanine Methyltransferase

MN: Micronucleus

NOGEL: No-observed genotoxic effect level

O6BG: O^6-BenzylGuanine

OECD: Organisation for Economic Cooperation and Development

Acknowledgments

During this long, challenging and rewarding experience of undergoing this Master's by research, I have been blessed to have been helped and encouraged by some truly great people.

I would first like to thank my mother and father, without their belief and encouragement to always strive to be the best version of yourself, I would not nearly have developed as much as an individual and I feel blessed to be called your son. To my sister also, for being there for me during all times I thank you.

In the laboratory, I would like to express my sincerest gratitude to my supervisor Dr George Johnson, for showing the faith, encouragement and belief in me and my abilities as a Scientist and for helping to instil this life long passion for scientific research which I am sure will follow. Always thank you for being there for me and answering my questions always no matter how silly they may have been. And thank you for broadening my horizons and motivating me to carry out new techniques and new areas of research. You have my utmost, gratitude, and respect as an individual.

I would like to acknowledge Dr Ben Rees, who helped me to develop as an independent scientist and grow into the research, constantly answering my question and helping me to develop. I'd also like to thank Professor Paul Rees for completely changing my way of thinking in research and helping me as I embarked upon Matlab based analysis, thank you for your patience with me and your willingness to ever help me and to allow us to carry out some pretty cool research. I'd also like to thank my secondary supervisor Dr James Cronin for helping me out with my calculations when I was stuck and helping to improve my understanding of the methods required.

I'd like to thank the lab, all the girls and the Tom's for being there for me when times were tough, but also for spending many quality moments together over the course of this project (I don't think I'll listen to any Disney without reminiscing). The summer of 2019 was a truly great one.

And lastly, I'd like to thank my friends outside of the lab and Swansea Uni hockey club for allowing me to switch off from labs too and for being there for me and making my Swansea experience a truly great one. To everyone, thank you.

1. Literature Review

i) Cell cycle.

The cell cycle is split into 4 main stages, G1, S, G2 and Mitosis (Copper, 2000). There is first an increase in cell size, characterised by the G1 stage (Copper, 2000). Following this, the cell synthesises new DNA resulting in the S stage. The cell is then required to prepare for division in the G2 phase and carry out the division in the Mitosis phase. G1, S and G2 are commonly grouped together under the interphase stage. Mitosis itself can be further distinguished into 5 main stages: Prophase, Prometaphase, Metaphase, Anaphase and Telophase. It is during this replication state, where the cell is vulnerable to errors taking place. Each time cellular division occurs, replication errors manifest. This can be highlighted via micronucleus (MN) formation, when damage to chromosomes causes a smaller MN to be formed during Anaphase, though this will be expanded on later (Fenech, 2011).

There are 3 mechanisms responsible for S phase replication occurring without fault (Kunkel, 2009, Ganai and Johansson, 2016). The first is the: nucleotide discrimination of the polymerase activity of the replicative DNA polymerases. The second is the: proofreading excision of mismatched primer nucleotides, and proofreading excision of mis-incorporated primer nucleotides by the 3’ to 5’ exonuclease activity of Pol ε and Pol δ. The third is: post replication mismatch repair (MMR) which works in combination with DNA replication in order to spot, excise and therefore replace any mismatched nucleotides or recently replicated daughter strands which remain (Bui D and Li J, 2019).

ii) Cellular damage.

Despite these mechanisms present for ensuring nucleotide integrity, the nucleotide error rate still lies at roughly 10^{-10}Bui and Li, 2019). This leaves room for mutation and consequently damage to occur. Cells are constantly at threat from sources of damage, be it endogenous sources, such as with a DNA mismatch or reactive oxygen species (ROS) or exogenous such as: x-rays or cigarette smoke (Chatterjee and Walker, 2017). These sources of damage to the cell can cause a great deal of genomic instability, which is one of the hallmarks of cancer (Hanahan and Weinberg, 2011). Moreover, this can cause certain key proteins to mutate, with proteins such as p53, known as the 'guardian of the genome' mutated and losing function in over 50% of human cancers (Lane, 1992 and Barker et al., 1989). Some sources, such as radiation can also lead to damage on a chromosomal level and was amongst the first discoveries to show that physical agents cause damage and therefore alterations to genetic matter (Evans et al., 1977). Moreover, it has been shown that these chromosomal abnormalities are of a direct consequence of damage occurring at the DNA level with breaks in chromosomes taking place due to double strand breaks in the DNA itself and an error in DNA mis-repair leading to chromosome rearrangement (Savage, 1993). This chromosomal damage has been shown to be of major importance in many diseases with cancers being a leading outcome (Roos et al., 2016).

iii) Cancer

The use of genotoxic compounds and their exposure to the population causes chromosomal, DNA and cellular damage to take place as mentioned. This damage can result in a somatic mutation when the damage has taken place in a somatic cell. This can then result in a transformation of the cell resulting in malignancy (Phillips and Arlt, 2009). This theory can be described as the 'Somatic Mutation Theory' and despite the theory being questionable for most cancers, it remains true that exposure to genotoxins is a cause of cancers (Brücher and Jamall, 2016).

Cancer is an ever-growing problem, in 2018 it was thought that around 1.7 million new cases of cancer appeared in the United States, with just over 600,000 of these cases resulting in mortality (National Cancer Institute, 2018). By inducing key mutations, defence mechanisms are stifled and the cancer is allowed to keep on

dividing and, in the later stages, metastasise to other areas of the body, causing complications and in 163.5 per 100,00 individuals on the whole mortality (National Cancer Institute, 2018). The generic signs of cancers are highlighted in two key papers, the Hallmarks of Cancer, and the Hallmarks of Cancer: the next generation (Hanahan and Weinberg, 2000 and Hanahan and Weinberg, 2011). The disease has many diverse and fatal forms, ranging from liquid cancers targeting the blood such as Hodgkin's lymphoma to brain cancers derived from cells producing the fatty covering of the nerves such as an oligodendroglioma.

That so many distinct forms of cancers exist, merely highlights the significance of testing drugs, food, and drinks in order to ensure genotoxins and carcinogens are not present or present in such minute doses that the effect is negligible.

1.1 Genetic toxicology.

i) Genotoxicity overview

Genetic toxicology is the study of chemicals ability to directly damage DNA. By directly damaging DNA and causing lesions, such genotoxins contribute to cell death and mutations which can have a direct link to diseases such as cancers. It is important to differentiate mutagenicity and genotoxicity. Mutagens have the ability to cause damage to DNA in both direct and indirect forms, whereas genotoxins only cause direct damage. All mutagens are therefore genotoxic, but not all genotoxins are mutagenic.

Due to the deleterious effects, highlighted in disease such as cancer, there is the need to have precise and accurate tests, both sensitive and specific for assessing such genotoxicity. There is the requirement to distinguish the mechanisms by which genotoxins or carcinogens may present themselves. Therefore, it is vital to undergo a multitude of tests on a substance which is being tested. Therefore, there is a battery test system in place. This involves the Ames test and the MN assay as two main test systems (OECD, 2013).

The Ames test is used to distinguish point mutations with the main mutation types detected being frame shifts and base substitutions (Mortelmans and Zeiger, 2000). The test focuses on using histidine dependent bacteria strains, where a mutation is required for viable bacteria to grow. This allows for mutant causing chemicals to be distinguished as when added these substances would cause an above normal proportion of viable bacteria to form on the plate. As these bacteria would have undergone mutations to his+ and thus will be able to grow the colonies (Mortelmans and Zeiger, 2000).

There are other tests used in accordance with the Ames test, including the Comet assay which can be used to assess low levels of DNA damage (Olive and Banάth, 2006). The test in focus, however, is the MN assay which is an OECD approved test used assess chromosomal damage (OECD, 2014).

ii) MN assay.

a) Background and Explanation.

The MN test is a standard genotoxic test used to quantify chromosomal damage (OECD, 2014). Chromosome damage includes both chromosome breaks and chromosome loss. Chromosome loss is caused by aneugenic substances. Clastogenic substances cause chromosome breaks. The test is approved by regulatory authorities and is a standard test used in chromosomal damage tests (OECD, 2014 and ICHS2(R1), 2012). It is of vital importance to fully grasp the importance and mechanisms surrounding the MN assay to move onto the techniques used in its analysis.

MN form when entire chromosomes or fragments of the chromosome cannot travel to the spindle during mitosis and lag. They are therefore not part of one of the main nuclei during division and their nuclear content is covered by their own separate nuclear envelope (Fenech, 2000). The phenotypic differences between a normal cell and one carrying a MN and the general ease in differentiating the two forms a large part of the reason for their use in genotoxic studies (Hintzsche et al., 2017).

b) History.

The MN was initially used as a marker of chromosomal damage over 40 years ago (Schmid, 1975 and Heddle, 1973). MN were known to haematologists in dividing cell populations such as those found in the bone marrow. Incidentally, the bone marrow is one of the OECD testing standards for the MN assay (OECD, 2014).

The MN assay took over from the more complex and time-consuming approach of metaphase aberration counting (Natarajan and Obe, 1982). In this technique, chromosomes were studied by spotting and counting aberrations shown during metaphase. Despite the detail, the loss of chromosomes during metaphase due to preparation methods, compounded with the time-consuming nature and complexities led to a simpler method required for chromosomal damage analysis (Fenech, 2000). This method was the MN assay.

c) MN mechanism

As mentioned previously, MN are used to detect either clastogenic substances through chromosomal breaks or aneugenic substances through chromosomal loss (See Fig. 1). This is due to the chromosomal damage disrupting the journey of the entire chromosome to the spindle during mitosis. This allows for a nuclear envelope to form around some of the DNA in the form of chromosomes and fragments at this point. The DNA material unwinds and shares much of the same morphology as a normal nucleus during interphase, the main difference being the far smaller size. It is also possible for nuclear buds and nucleoplasmic bridges to form at a similar stage, however such events are much rarer than MN and can be noted when scoring MN.

There are two ways of carrying out the assay. The test can be carried out both *in vitro* and *in vivo* (with the use of animal tissue, primarily liver cells when using chemicals requiring oxidation). The *in vitro* method is becoming a more popular approach in the scientific community overall, due to the issues raised ethically with the use of animals in scientific experiments. This is also compliant with the 3 R's principle towards animal testing, Replacement, Reduction and Refinement (NC3Rs, 2020). As

it is becoming clearer that the use of animals when carrying out *in vivo* methods is slowly becoming outdated and the systems are not in fact as reliable as previously thought when compared to the working human environment. Therefore, the 3 R's principle is becoming more prominent in research, with an increased emphasis on increasing mechanistic understanding of biological systems in order to replicate this in a more accurate *in vitro* test. It must be noted that 23% of oncology drugs in a late stage clinical development review failed due to cytotoxicity reasons (Jardim et al., 2017). More of this will be touched upon later, the MN work carried out during this project was undertaken in an *in vitro* setting.

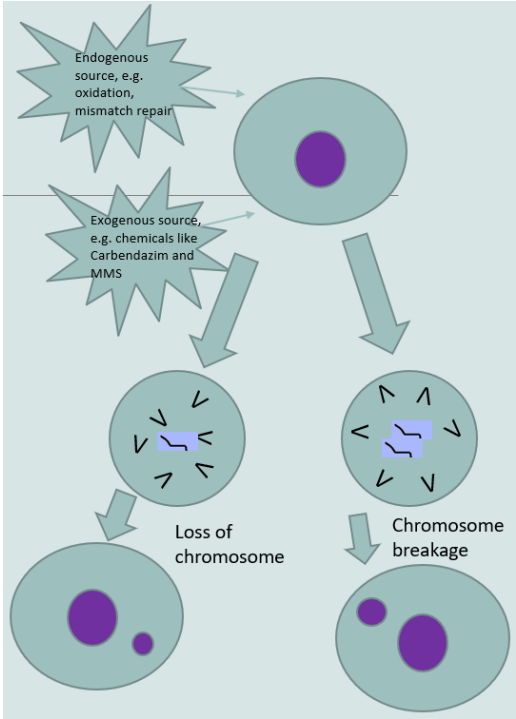


Figure. 1 Schematic showing MN formation by chromosomal breakage and chromosomal loss following endogenous and exogenous damage.

d) Cytocholasin-B

It is important to note there are two main forms of MN assay, whether it is carried out in an *in vitro* or *in vivo* setting. The assay can be carried out using mononucleated cells and thus analysing mononucleated cells containing a MN to determine the MN frequency. The other type is the binucleated MN assay, where Cytocholasin-B (cyto-

B) is added to form binucleated cells, which are then analysed and thus binucleated MN cells are analysed to obtain a MN frequency (See Fig. 2).

Cyto-B is a mycotoxin permeable to the cellular membrane which inhibits cytokinesis without affecting nuclear division by preventing actin filament formation. This leads to cells forming a binucleated shape after undergoing cyto-B treatment. This is advantageous as the formation of a binucleated cell gives a confirmation that nuclear division has taken place, which cannot, as of yet, be identified when cyto-B is not used and mononucleated cells are analysed (Fenech, 1997) (SeeFig. 2). The use of cyto-B results in the cytochalasin-b Micronucleus assay (CBMN). This is, given the confirmation of cellular division, the preferred method used in laboratories worldwide. For the MN percentages to be deemed accurate, 2000 cells as a minimum are required when analysing mononucleated cells after performing the MN assay or 1000 for binucleated cells when carrying out the CBMN assay (Fenech, 2000). By comparing the control percentage of MN to the dosed percentage of MN, it is therefore possible to calculate if chromosomal damage has taken place.

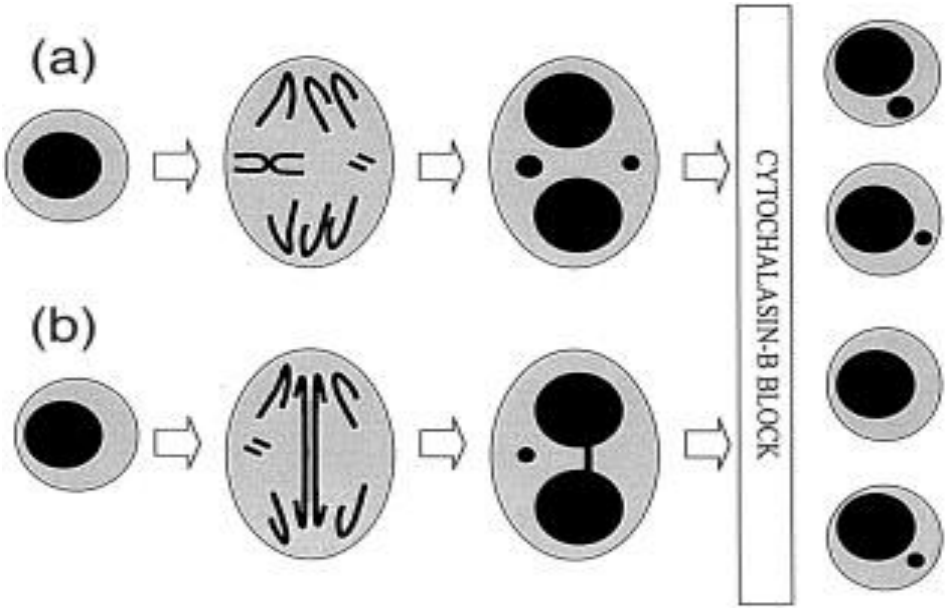


Figure. 2. (a) A MN originating from a lagging whole chromosome and acentric chromosomes fragments at the anaphase stage. (b) Shows the formation of a nucleoplasmic bridge from a dicentric chromosomes, the centromeres are pulled to opposite poles of the cell. Can also see the formation of an MN from the acentric fragment of the chromosome. The role of cyto-B in stopping cells from dividing at the binucleated stage can also be seen. This is for a cell with two pairs of chromosomes (Fenech, 2000).

1.2 Types of genotoxin

i) Clastogens and Aneugens

As previously mentioned, genotoxins causing chromosomal damage are split into two main group. Clastogens causing chromosomal breaks and damage and aneugens causing chromosomal loss to take place, with both leading to DNA damage as a direct result and the capability to be tested using the MN assay. More well-known clastogenic examples include Methyl methane sulfonate (MMS) (See Fig. 1) (Chatterjee and Walker, 2017). There are believed to be three major types of endogenous classes of clastogens (Emerit, 2007). The first class contain lipid peroxidation projects synthesised from arachidonic acid of membrane with aldehyde 4- hydroxynonenal being a leading example (Emerit e al., 1991). The second class

are cytokines, including tumour necrosis factor alpha (Emerit et al., 1995). The last category are uncommon nucleotides, with an example being inosine tri and diphosphate (Auclair et al., 1990). Aneugens are responsible for chromosomal loss, with the initial hypothesis stating that the MN produced by aneugens would be larger due to whole chromosome loss (Yamamoto and Kikuchi, 1980). However, it is a very difficult process to accurately check this due to the differences in chromosome sizes being an issue between species (Rosefort et al., 2004).

Aneugens form MN through chromosomal loss (See Fig. 1). Colcemid is an example of a well-known aneugen with work being carried out for decades on its mode of action (Rudd and Hoar, 1991). Carbendazim is another example of an aneugen which has been studied extensively and used in our laboratory previously (Verma et al., 2017 and Verma et al., 2018). Aneugens are thought to induce different shaped MN with these MN not perfectly circular in shape and thus a cause for concern as many machines, including automated microscopy have had difficulties in scoring these. Both classes of chemical cause serious cellular damage and the OECD have guidelines on the use of such chemicals and the classifications it belongs to (OECD, 2011 and OECD, 2013).

A general rule is applied in the pharmacology industry where under 1.5µg/tablet is considered a safe level of genotoxic substance for the majority of these chemicals. NDMA is among a select group of chemicals under an exception due to these being considered extremely potent and the guidelines are extremely strict on this (EPA, 2017).

ii) DNA Damage and repair

a) MGMT

O^6-methylguanine methyltransferase (MGMT) is a DNA repair protein responsible for the removal of alkyl adducts from the O^6 position of guanine (Estellar *et al.,* 1999). The alkylation of DNA at the O^6 position is formed as a response to either environmental pollutants, tobacco-based carcinogens and anticancer medication (Christmann *et al., 2011)*. O^6MG is however a secondary adduct, accounting for only around 7% of all adducts initially formed by alkylating agents, with N7-methylguanine accounting for 65% of all adducts formed upon initial exposure (Liteplo *et al., 2002)*. However, despite being a less common adduct, O^6MG is highly mutagenic and has the greatest potential to lead to apoptosis (Kaina *et al., 2010*). MGMT repairs O^6MG by shifting the alkyl group to a cysteine residue in its active site (Christmann *et al., 2011).* Following this, the protein becomes inactivated, ubiquitinated and targeted for degradation by the proteasome (Xu-Welliver and Pegg, 2002). Without MGMT present, O^6MG forms point mutations, leads to double strand breaks which trigger apoptosis by due to cellular replication and DNA mismatch repair (MMR) (Ochs and Kaina, 2000).

The most common environmental alkylating agents are the N-ntroso compounds with NDMA being the first N-nitro compound to be found as well as the most prevalent in the diet (Lijinsky, 1999). Thus, MGMT has the possibility to have great potential in the DNA damage/repair pathway in NDMA.

The consequence of alkylation of O^6 leads to cancer progression due to the similarity in conformation to adenine and thus pairing with thymine during replication (Estellar *et al., 1999)*. This causes genomic instability (a hallmark of cancer (Hanahan and Weinberg, 2011)). Moreover, the change in base pairing caused from the alkylated O^6 also leads to dysfunction and increases the chance of mutation, again leading to cancers. MGMT has been found to be manipulated by cancers in the response to anticancer drugs. A major mechanism of cancer resistance to drugs is by enhancing the activity of MGMT which thus counters the effect of DNA-alkylating chemotherapy drugs at the O6MG position (Fan *et al., 2013)*. Therefore, MGMT can play both a positive role in repairing DNA damage at the O^6 position and restoring the cell to a healthy status. But it can also play a negative role by reducing the effectiveness of anti-cancer DNA alkylating chemotherapeutic agents by countering the damage caused by repair (*Christmann et al.*,2011).

b) O^6BG

O^6-Benzylguanine was originally designed by focusing on the biomolecular displacement reaction between the leaving group at the O^6 position of guanine and the MGMT protein (Dolan *et al., 1985),* (Dolan *et al., 1990).* Benzyl groups are used with more ease in biomolecular groups when compared to alkyl groups (Dolan and Pegg, 1997). When adding micromolar concentrations of O^6BG, it was observed that MGMT levels were entirely depleted which led to increased sensitivity to O^6guanine alkylating agents (Dolan *et al., 1990)*. Because of these properties, O^6BG is used greatly to sensitize tumours to lower doses of alkylating agents in a chemotherapy setting. However, due to its role in the depletion of the role of MGMT, it is possible to determine the levels of DNA repair ongoing whilst comparing samples when O6BG is added to samples when it is not. This leads to the determining the effect of DNA repair on nitrosamine compounds and specifically the nature of MGMT repair specificity in tandem with dose. By adding O^6BG to control samples, it is possible to compare and gage levels of endogenous DNA damage taking place through the creation of O^6MG and determine the effect of inhibiting MGMT DNA repair.

c) Nuclear Stains

In order to visualise cells appropriately, nuclear stains are used in the preparation of the cells for visualisation, be it manual microscopy, automated microscopy or flow and imaging flow cytometry. Different stains are used in tandem with different machinery accordingly to optimise the peak intensity levels of the specific stain. A nuclear stain is vital as it helps to differentiate an artefact of a similar size to true nuclear material, which is vital when carrying out genetic toxicology tests such as the MN assay.

1.2.2C i) Draq 5

Deep Red Anthraquinone 5 (DRAQ5™) is a far-red DNA fluorescent dye used to stain nucleic acids and differentiate from debris in live or fixed cells (the latter being of more interest to this project). (BD Pharmingen, 2017). DRAQ™ has a maximum excitation of 598/646nm but can also be used sub optimally with the 488nm laser

(BD Pharmingen, 2017). By staining with Draq5™, it is possible to compare the brightfield and fluorescent images on the IDEAS® program during analysis and differentiate debris by its appearance in the brightfield image but lack of appearance in the fluorescent (DRAQ™ in this case) channel.

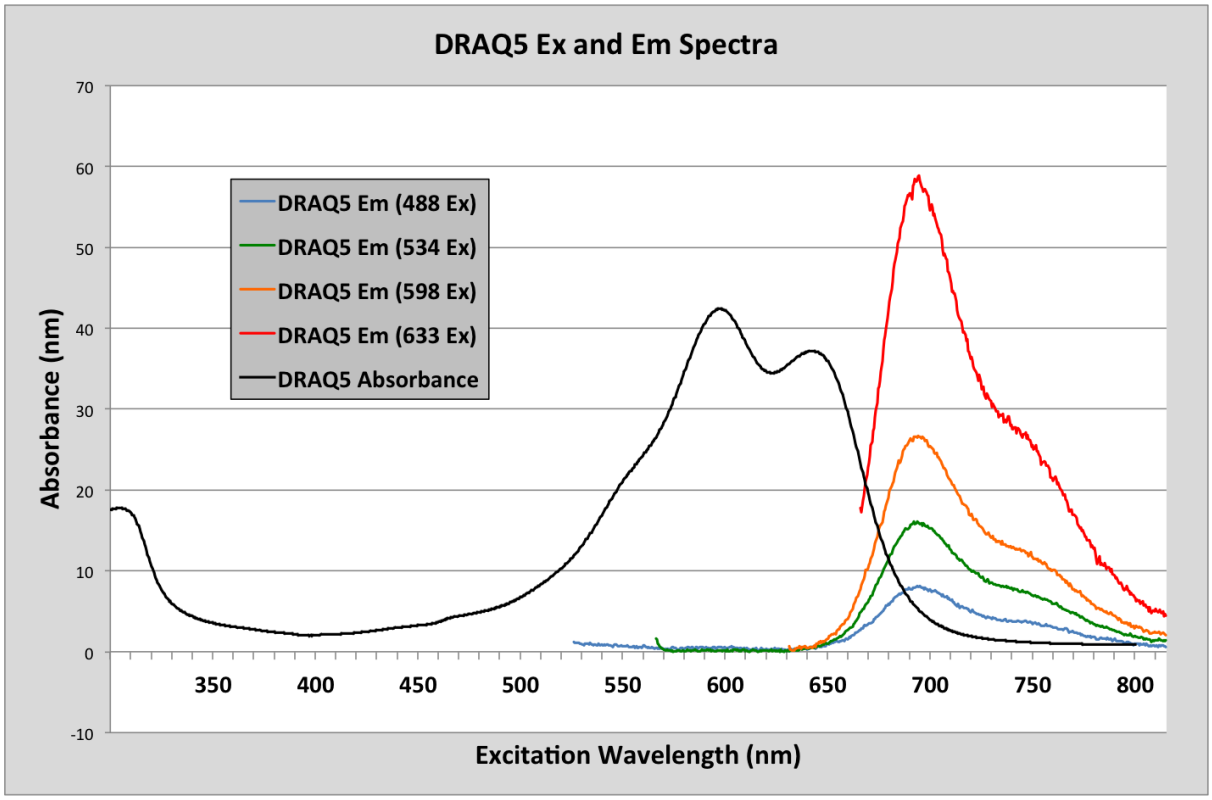


Figure. 3 Emission and excitation spectrum of DRAQ5™ (Biostatus, 2017).

1.2.2C ii) Hoescht

Hoescht 33342 Solution is a fluorescent reagent used in the staining of DNA and nuclei in live or fixed cells (just as DRAQ5™). Hoescht 33342 is advantageous due

to its high specificity for double stranded DNA binding, with a preference for A-T binding (BD Biosciences, 2020). Thus, the dye is extremely useful in labelling double-stranded DNA and in turn the nucleus where the DNA resides. A blue fluorescence is emitted with a maximum emission at 461nm when binding to DNA (BD Biosciences, 2020). The specificity for DNA double-stranded binding allows for ribonuclease treatment to be skipped and non-specific RNA staining is avoided (BD Biosciences, 2020).

d) Cell lines

Human lymphoblastoid cells provide a comparable testing sample to use due to similarities in morphology to primary lymphocytes, therefore they have been used historically to model genotoxic systems (Verma *et al*., 2017). These cell lines can be differentiated into different subgroups, based on levels of cytochrome activity incorporated into the cell line.

1.2.2d i) TK6 cells:

Human lymphoblast, thymidine kinase heterozygote form the more commonly known TK-6 cell line. These are frequently used in genetic toxicology in both industry and academia due to their suitability in the OECD test guidelines for the *in vitro* MN assay (OECD, 2014). TK6 cells are normally employed for chemicals without a need for metabolic activation due to the lack of cytochrome activity present. Cell lines derived from human lymphoblastoid cells are also larger in size than primary lymphocytes, this allows the cell lines to be used on a wider range of machines with reduced magnification capacity which is in turn cheaper for the laboratory and more accessible, allowing for a wider range of research to be carried out.

1.2.2d ii) Metabolically active cells

1.2.2.d.2 a) AHH-1

AHH-1 cell lines are, just as TK-6 cells, derived from human lymphoblastoid cells. Differentiating AHH-1 cells from TK-6 cells is the addition of the cytochrome CYP1A1 expression to a high level (Crofton-Sleigh *et al, 1993)*. This allows this cell line to be used in the assessment of metabolically requiring test chemicals, whereas TK-6 cells cannot. The AHH-1 cell line is a 'parent cell' to the more metabolically competent MCL-5 cell line, which contains a plasmid containing an additional 4 cytochromes. Thus, these two cell lines have great use in tandem with one another due to this similarity. The preparation methods are extremely similar, with the same media, horse serum and glutamine used for both.

1.2.2.d.2 b) MCL-5

Metabolically competent MCL-5 cells are also derived from human lymphoblastoid cells. These cells provide continuous expression of active cytochrome p450 metabolic enzymes which explains why they are frequently employed in the assessment of test chemicals requiring metabolic activation (OECD, 2014). The cell line is a TK derived cell line with the AHH-1 cell line being the parent cell line, expressing only CYP1A1 (Crofton-Sleigh *et al., 1993).* The MCL-5 cell line also contains the cytochromes: CYP1A2, CYP2A6, CYP3A4, CYP2E1 in a plasmid, thus enabling greater metabolic output. This renders the MCL-5s more sensitive to potent metabolising carcinogenic compounds and provides a comparison with AHH-1 cell lines on which cytochromes are used in the metabolic processing of such test chemicals. Therefore, these two cell lines are ideal candidates to use in the assessment of test chemicals, such as NDMA. As mentioned previously, the preparation methods of MCL-5 cells is very similar to AHH-1 cells, with the exception that hygromycin is required to be added to the MCL-5 media in order to maintain the plasmid integrity, the plasmid contains a hygromycin resistance gene, therefore the addition of hygromycin allows for the selection of hydromycin resistant cells . (Aranda *et al., 2014).*

1.3 Micronucleus scoring methods

i) Traditional scoring

Currently, the MN assay is carried out by a variety of different methods and using different equipment, ranging from microscopy to the use of imaging flow cytometers and the scope for robotics. Traditional scoring methods for measuring MN have been based on light microscopy.

In this manual approach, cells are stained with giemsa and manually counted to quantify MN. Giemsa staining is used as the dye attaches well to DNA rich regions and more specifically to the adenine-thymine rich regions. However, different stains can and are used, such as 'Diff Quik' (Lab- Aids, Australia) which is the recommended stain by Michael Fenech in his revolutionary paper on the MN assay (Fenech, 2000). The Diff Quik stain is a version of the Giemsa stain with a major advantage of streamlining the procedure from about 4 minutes to around 15 seconds. When undergoing fluorescent microscopy, acridine orange is the recommended stain to be used (Fenech, 2000).

After the cells have been stained, the slides are examined by use of a light or fluorescent microscope. It is recommended that a magnification of 1000x is used when analysing peripheral human blood cells due to the smaller size of such cells compared to the commonly used immortalised cells (Fenech, 2000). To analyse the cells, both accurately and fairly, various measures are put into place to ensure this. The figure of cells required to be scored increases to 2000 when scoring mononucleated cells. Next, the MN analysed must be between 1/3rd and 1/16th the diameter of the main nucleus and have the same circular/oval shape. Moreover, a code is aligned to each slide so that the dosing applied to each slide is kept hidden to eliminate user bias. These steps maintain the integrity of the results produced and enable the MN assay to be used and approved by many regulatory bodies (OECD, 2014 and ICHS2(R1), 2012).

These are the reasons why manual scoring the cells via manual microscopy is still deemed to be the 'gold standard' (Vermat et al., 2017).

There are therefore key advantages into the manual scoring of MN using light microscopy as with the gold standard. However, despite this method being the 'gold standard', there are evident issues present.

The process is laborious, despite the use of dyes such as the Diff Quik to shorten the process, the procedure still takes much time and is tedious. The scoring can be subjective, despite the coding of the slides, a degree off opinions still exists and this can be vital when dealing with rare events. The subjective scoring can therefore lead to interscorer variability (Doherty et al., 2011). This, coupled with the lack of being able to truly archive results, somewhat limits the manual scoring approach and, though still considered the 'gold standard', makes true scoring difficult to achieve.

ii) Semi-automated microscopy (Metafer).

A new method developed to increase throughput compared to manual microscopy was automated microscopy. In this method, the slides prepared by microscopy can be automatically checked, resulting in a less laborious process of manual scoring, without sacrificing the integrity of true results. The Metafer™ system was a system produced and characterised in regard to MN work, it has been used, reviewed in papers such as Verma et al and compared to manual microscopy as well as flow cytometry approaches (Verma et al., 2017). This semi-automated system, where cells are stained with a fluorescent dye before being scored both in an automated and manual manner, allows for MN to be scored with greater ease and speed (Doherty et al., 2011). The stained slides are loaded onto the scanning platform, images of MN are taken using a 10x lens. These are then checked using a 100x lens in line with the coordinates shown in the display view (Verma et al., 2017). This method, as well as being quicker, was shown to be reliable and produce results in line with traditional MN scoring (Chapman et al., 2014). There is also the tool of storing results for re-evaluation and a dose-based system. This enables a level of inter scorer comparison

to be made which cannot take place by manual scoring and increases confidence in the results seen.

However, as seen in Table.1 there are flaws to the Metafer™ system. There is a lack of cytoplasmic staining, difficulty in differentiating MN overlapping the parent nuclei and the necessity to update the classifier setting conducting MN when scoring different cell lines (See Table.1). Moreover, there is a need to manually validate the images when scoring, which slows down the process and reduces the automation. Changing the classifier setting leads to the system producing an under estimation of MN frequency which is a limiting factor of its use (Verma et al., 2017). However, there is the potential to overcome this with the addition of a visual detection step (Decordier et al., 2009). Moreover, it can take further time to optimise setting for different cell lines and morphologies, which reduces the advantage of the systems speed. The cause of the underestimation is due to cells with novel nuclear morphology's not being identified or large MN being misclassified as nuclei (Verma et al., 2017). Lastly, there is concern about the lack of visualising the cell membrane, this limits the use of this technology and is a major reservation of industry into the use of automated microscopy to determine MN frequency.

MN scoring approaches	Scoring Platforms	Advantages	Disadvantages
Image analysis	Manual microscopy (light microscopy)	❖ Suitable for dose response and mode of action analysis ❖ Simple, economical and adaptable ❖ Suitable for MN scoring in the presence or the absence of cyto-B ❖ Stained slides can be stored for a long time and can be re-analysed ❖ Suitable for assessing bi-, tri- and poly-nucleated cells	❖ Interoperation variation can result in subjective MN scoring ❖ Slow, tedious and time-consuming ❖ Lack multiplexing abilities ❖ Total number of cells scored manually is limited which reduces the overall statistical power
	Metafer™ (fluorescent microscopy)	❖ Semi-automated platform ❖ High content for higher statistical precision ❖ Suitable for dose response and mode of action analysis for most substances ❖ Images of nuclei and MN can be stored for re-validation	❖ Classifier settings have to be optimised for different cell lines and chemicals that induce MN via varied mechanisms ❖ Lack of cytoplasmic staining, detection of small MN and manual validation of the images
Flow cytometry	MicroFlow®	❖ Fully automated platform to score MN objectively ❖ Suitable for dose response ❖ High content and high throughput ❖ Permits cell cycle analysis 10,000 events scored in 1–2 min	❖ Cell lysis is required prior to MN scoring ❖ Misleading MN cannot be re-validated from same sample ❖ Overestimation and underestimation of MN are both possible and require expert analysis ❖ Lack of MOA analysis with TK6 cells

Table. 1 An overview of the advantages and disadvantages of MN analysis using the manual scoring, Metafer™ semi-automated fluorescent microscopy and the MicroFlow® flow cytometry approaches to the MN assay (Verma et al., 2017).

iii) Automated scoring (Microflow).

Due to the need to still manually 'check' the cells being scored using semi-automated microscopy, and the laborious time associated with this resulting in this being a limiting factor, new systems were developed to increase the throughput of the assay. This led to the use of systemssuch as the MicroFlow® flow cytometric approach. This approach eliminates the laborious and time consuming approaches previously used and provides a high throughput, which is vital. Moreover, the use of nuclear stains like ethidium monoazide (EMA) allows for apoptotic bodies and necrotic cells to be differentiated from MN, this is often a challenge when manually scoring. Due to the lack of visibility of the cells, as there is no camera attached to the flow cytometry, as is the case with imaging flow cytometry which will be discussed shortly, it is ever important to use nuclear stains to ensure that MN scored are true MN as much as possible. The ability to score 10,000 cells in a minute is a true advantage and streamlines the scoring process, this is 15x quicker than manual scoring methods at least (Verma *et al.,* 2018). This improves the laborious and tedious methods of manual and automated microscopy with visual scoring no longer an issue.

However, the lack of visualisation is a major disadvantage of this system. 'Double checking' cannot take place and thus some confidence in results decrease. Also, not being able to store samples for a long period of time, in comparison to manual scoring and automated microscopy techniques where the slides can be stored for months, hinders this method considerably (Fenech, 2013). The lack of considerable storage of slides reduces the confidence in the results also by the lack of a 'double check' mechanism being in place by not having an archive of results. Another major disadvantage is the lack of differentiation between bi, tri and multinucleated cells with MN and cells with multiple nuclei (Verma et al., 2017). Furthermore, lysis of the cells occurs, leading to an overestimation of the MN count (Fenech et al., 2013).

Due to the lysis breaking up more parts of the cell, more artifacts are produced, which are of a similar size and shape as MN and are thus sometimes miscounted as MN. Lysis of the cells is not recommended by Michael Fenech, as this can lead to an excess of debris which is difficult to differentiate from MN (Fenech, 2000). Moreover, there is also room for underestimations to take place on MN count, this combination of both MN overestimation and underestimation limits the use of the MicroFlow® and coupled with being unable to visualise the cells limit the use of flow cytometry in carrying out the MN assay significantly.

iv) Imaging flow cytometry (Image stream and Flow Sight).

Following the MicroFlow™, there was a need for an automated system for the assessment of MN to truly bring the approach into the 21st century. However, the major issue with the MicroFlow™ was the need to lyse the cell as well as the lack of visualising the cell. The result was the imaging flow cytometer: FlowSight® (Amnis, part of EMD Millipore). The imaging flow cytometer combines both the automated aspect of flow cytometry with the imaging of manual microscopy. The machine functions as a normal flow cytometer does, forward scatter with side scatter is available. With the base flow cytometric foundations, there is an additional bonus whereby each cell is captured as an image, this allows for each individual cell to be clicked on and analysed, adding an extra degree of confirmation. The tool is extremely powerful, magnifications range from 20x-60x depending on which model is used. The FlowSight® has a magnification of 20x, this is useful and allows for the comparison to microscopy to be made. However, this is not always a strong enough magnification when focusing on smaller events, such as primary lymphocytes. There is however, the Image Stream x Mark II® (Amnis, part of EMD Millipore) which is the more powerful version of the FlowSight®, this allows for magnifications of 40x to be achieved and ‘add ons’ can be applied to achieve a magnification of 60x which increases the capability of the machine and allows for crisper, cleaner images of events in focus. The greater magnification proves useful in the assessment of smaller cells such as primary lymphocytes, which cannot be visualised correctly on the FlowSight® due to being too small.

Amnis boasts that the Image Stream x Mark II® has a high throughput with the ability to process thousands of cells per second. It is ‘intuitive’, ‘adaptive’ and ‘boundless’ (Amnis imaging flow cytometer brochure, 2016). The camera in the Image Stream x Mark II® comes in varying pixel sizes, coming in at 0.1, 0.25 and 1μm². The cells are lit up by means of a brightfield (BF) light-emitting diode (LED) side scatter laser and fluorescence is provided by one or more lasers. The emitted photons are collected by a ‘high numerical aperture objective lens’ (Rodrigues, 2018). The photons pass through a spectral decomposition element which allows for a specific range of wavelength, 400-800nm in this case, to be separated (Rodrigues, 2018). As they are separated, the charge-coupled camera (CCD) takes up to 10 fluorescent images simultaneously at different parts of the camera for each cell. In combination with the two brightfield images produced, 12 images are taken per cell which therefore allows for high detail images of the cell to be obtained (Rodrigues, 2018).

Moreover, the machine arrives with 12 lasers, equipped for a variety of dyes and this fluorescent data is gathered alongside both brightfield and darkfield images, allowing for a high content, high integrity analysis to be carried out. The experiment is carried out without the need to lyse the cells, this being a key advantage over conventional flow cytometric approaches (Verma et al., 2018). The lack of cell lysis also is coherent with OECD guidelines where it is recommended that for MN scoring, the cells should have an intact cytoplasmic membrane (Fenech et al., 2013). The scoring metrics remain the same as per manual microscopy techniques. 1000 cells are scored for binucleated cells and 2000 cells for mononucleated cells.

The initial applications to the MN assay were to manually analyse the images taken using the imaging flow cytometer manually. This can be carried out in a similar manner to manual microscopy scoring; however, the scoring does not need to take place on a microscope but on the IDEAS® program on a computer.

The technique had been compared to manual microscopy in radiation dosimetry with the results found comparable (Rodrigues et al., 2014). This led to the approach being used when comparing chemical dosage with the results again found to be comparable (Haxhiraj et al., 2018 and Verma et al., 2018). This is vital due to manual microscopy still being considered the ‘gold standard’ (Verma et al., 2017).

Experiments were carried out on primary lymphocytes extracted from the blood using the FlowSight® at a magnification of 20x. However, this proved not to be a strong enough magnification (Haxhiraj *et al*., 2018).

The advantages of this technique are being able to access 20,000 cells within minutes, archiving the images produced and being able to further analyse these using the IDEAS® program to play a large factor in the current use of imaging flow cytometry in MN analysis. Moreover, the manual scoring on a laptop can be carried out in a less strenuous manner which can maintain that scorer subjectivity as scoring levels do not waiver and therefore consistent scoring is more likely to be achieved. Moreover, 12 channels provide a multitude of biomarkers and dyes to be used, which can help to further differentiate MN. The masks and templates present in IDEAS® also allow for a greater use of tools to query subjective MN. Lastly, the presence of a Brightfield image allows for the cytoplasmic membrane to remain intact and can allow for a composite image, containing the cytoplasmic integrity of a brightfield image, with the fluorescent identification ease of a DNA label to increase confidence in the users scoring. Specific images can be marked and rescored by other scorers and thus achieve a moderated standard.

There are major disadvantages with the method currently, however. The images produced of the cells still must be manually scored, a laborious and tedious process. The imaging flow cytometers are also expensive, especially the **ImageStreamX,** which is required when analysing primary lymphocytes. This limits the quantity of laboratories with this equipment and with the service charge also being expensive, the method is limited to few laboratories. It is far more economically viable to buy a microscope and the giemsa stain required to carry out the gold standard which allows for a more widespread use. Moreover, as manual scoring is still being carried out with the imaging flow cytometer, expertise is still required in the scoring stage. In order to make the expenditure of an imaging flow cytometer even more worthwhile, full automation is required for the MN assay.

The manual scoring using the IDEAS® program allows for the next step in the process of complete automation to take place and is a key contributor to the next stage in the process: producing a deep learning algorithm to automate the MN assay fully.

1.4 Analytical tools

i) Use of IDEAS® program

a) Background and explanation

IDEAS® is the program which comes linked with the imaging flow cytometry platforms provided by Amnis. When the cells are processed by the imaging flow cytometer, the Inspire software (the computer program which comes as standard with the imaging flow cytometer) is automatically loaded up. A general gating system can be produced here and then the viable cellular data can be transferred to a memory stick and loaded onto the IDEAS® program. The program comes with unique features, enabling for the manual gating of the desired region and allows for each 'dot' on the scatter plot to be clicked upon and visualised. Each image obtained also has an individual number attached to it, this image can be double clicked and copied to another region. This allows for cells containing MN to be visualised separately and the number attached allows for each image to be archived and checked again at a later date by a more experienced scorer if needed in order to ensure that the scoring integrity is kept. By being able to save different populations, it is possible to group and save rare phenotypes (MN, nuclear buds (N-buds) and nucleoplasmic bridges etc).

b) Template and Tools.

The IDEAS® program has specific features and tools which allows for analysis to be undertaken more quickly. By applying specific features, the user can manipulate pixels which can allow for the differentiation of cells to be carried out with more ease. These features and mask can vary from simple metrics, such as 'Spot intensity' which allows for circular groups of pixels to be identified with more ease, to a

combination of a variety of masks and features which in turn eliminate much of the cells and allow for much more specific phenotypes to be identified. This is the basis of some of the work which has been undertaken by Rodrigues (Rodrigues, 2018). By creating highly specific templates using complex masks and features in IDEAS®, it is possible to obtain specific phenotypes, be it MN, N-buds, Nucleoplasmic bridges or just binucleated cells. By combining features and masks together, it is possible to narrow down the quantity of cells deemed a specific phenotype. This allows for a pool of cells to be formed, with these cells being checked to confirm that the phenotype is what was thought. This speeds up the time taken to make up the 1000 binucleated cells which are needed, but the manner is not perfected and this is why much work is being undertaken on producing an algorithm which is both quick and is comparable to the 'gold standard' of manual microscopy (Verma et al., 2017).

c) previous use and comparison (Rodrigues paper).

The combination of masks and features in tandem was used by Rodrigues to create an algorithm which aligned with criteria set out by Fenech previously (Rodrigues, 2018, Fenech et al., 2003, Fenech, 2007). This algorithm, set out for scoring the CBMN assay, reduces the scoring time required in manual scoring, this increase in scoring rate allows the assay to not only be performed quicker, but with more statistical integrity also (Rodrigues, 2018). In 2018, Rodrigues, showed that the calibration curve produced using his algorithm had similarities to others produced in literature and therefore showed some promise (Rodrigues et al, 2016, Rodrigues, 2018). The conclusion generated by this was that this method could produce radiation dose estimates to within +/- 0.5Gy of the actual dose, this is appropriate for triage radiation biodosimetry (Rodrigues et al., 2016). Moreover, the four chemicals used in this study, two aneugens and two clastogens showed a significant increase in MN at all but the lowest two doses of Colchicine (0.005 and 0.01μg/ml and the lowest dose of VS (0.005μg/ml) (Rodrigues, 2018).

However, it must be noted that the base MN level in the Rodrigues study was found to be 0.19% (Rodrigues, 2018). This figure is considerably lower than the historical MN levels in literature of between 0.32%-1.38% when using more traditional methods, such as microscopy and standardised flow cytometry (Lovell et al., 2018).

The lower background MN rate is an area of concern and despite the advantages seen and produced by the algorithmic method applied here, ‘caution must be taken when attempting to draw conclusions based on comparisons between the experimental results presented here and published literature’ (Rodrigues, 2018).

d) Next steps

This latest attempt at increasing the throughput of the MN assay whilst retaining the integrity and accuracy of the results led us to the algorithmic approach we have undertaken. The next challenge was to keep the throughput high, but to produce a comparable result to the ‘gold standard’ of manual microscopy scoring (Verma et al., 2017).

1.5 Algorithm theory.

i) Background and Origins

In order to automate the assay, an algorithm would have to be created to allow for artificial intelligence (AI) to be used to significantly reduce the laborious nature of the test, whilst keeping throughput and accuracy at a level comparable to the ‘gold standard’. Two main methods branch out from this, machine learning and deep learning.

ii) Machine learning

Machine learning, in regard to MN assay developmental, use revolves around writing a script which incorporates ‘structural rules’ into the script. These set of rules can then be used in the application of the MN assay. For the assay, a rule would be to search for a region with a diameter between $1/16^{th}$ and $1/3^{rd}$ of the main nucleus in the cell as per Fenech’s guidelines (Fenech, 2000). Moreover, an aspect ratio closer to 1 would signify a circular shaped MN. By combining these factors and more, as

was carried out by Rodrigues, the resulting MN can be analysed, and a dose response carried out.

Rodrigues carried this process out in the IDEAS® software, by manually inserting these guidelines of a MN by use of masks and features as specified earlier. Writing a machine learning algorithm is a way to streamline this process without the time-consuming approach of having to create and modify many different masks and features. However, due to the ability to save templates on the IDEAS® program, the process is relatively streamlined for further use.

The issue with the Rodrigues approach to automating the MN assay, and the Achilles heel of machine learning in this scenario, is that it is still very difficult to classify MN without manually analysing the MN individually. Due to certain MN not be perfectly circular, with a perfect circle being difficult to be capture when viewing a 3D object in 2D, accuracy levels are not always as high as needed.

The result is that images are captured of near perfect MN in phenotype, The resulting cohort of MN produced are specific to the point of being over specific. This is advantageous in that no false positives are included; however, this is not a true reflection of the total MN in a sample and therefore cannot be reflective of the DNA damage taking place in the cell. This reduces the applicability of a dose response and does not forward the use of the MN test. This could be a reason as to why the background level shown by Rodrigues in his automation attempts were so low, as not all MN adhere perfectly to the guidelines instated.

iii) Deep learning

a) Background

Deep learning is seen by many as the evolution of machine learning, the next step in artificial intelligence. Despite seeming similar however, deep learning focuses on a different approach to solving problems and issues.

With deep learning, fewer initial rules or structures are required. This makes the approach favourable when dealing with more complex issues, whereby one common theme does not necessarily apply to all the subjects. By attempting to mimic how neurones act in the brain, the deep learning algorithm comprises of a neural network. Where neurones in the brain link to one another and form connections, the neural network is split into different layers. Each layer communicates to another layer and passes its verdict to the next layer and so on. Thus, the more layers, the greater the computational demands of the network but also the added integrity of the results produced.

In order to know what to look out for, the neural network is trained by a 'ground truth' of images**.** In much the same way that we learn to classify a ping pong ball and an American football as both being variations of a ball. We can distinguish the difference between these two variations, not only having experience with viewing both a ping-pong ball and American football, but by viewing the different types of balls in between, a tennis ball, golf ball etc. By having this bank of mental images, we are able to identify the differences in identifying the different types of ball, much like the more 'correct' images the network can train with, the greater the accuracy.

Much in the same way, by producing images by which the network can learn from, the more images the better, the network can run these images through, layer by layer, to come to a verdict on the identity of each specific image. The network therefore improves, through the more images it scores and can also provide an accuracy level which is likened to how one would manually score the same image.

b) Ground truth

The ground truth is what provides the neural network with the initial data by which it can make decisions on unseen data. As such, it is important that the training data and

validating data not be mixed, as this leads to an unrealistic and unreliable accuracy rating being shown and the network has not been tested with new data.

In order to apply deep learning to the MN assay, images are needed of the varying cellular phenotypes. As such, mononucleates with and without MN, binucleates with and without MN, trinucleates with and without MN, quadranucleates with and without MN are all needed to be identified and incorporated into the teaching ground truth in order to differentiate between these different phenotypes and therefore identify MN and the resulting mononucleate or binucleate which is needed in tandem with the MN count in order to produce a %of Mn and thus a dose response.

With a system needed to transcend the manual scoring of images which takes place during imaging flow cytometry scoring, an algorithm was still deemed the correct approach to the problem. The solution was different to the approach taken by Rodrigues. Whilst automating the manual scoring of images on the IDEAS® software seemed to be the answer in theory, the reality was that there were still many MN not scored in the data.

The ‘ground truth’ of images which could be fed to the algorithm was the next step, with previous proof of concept experiments showing the use of imaging flow cytometry for manual scoring of cells, which could be applied in this case leading to a ‘ground truth’ (Verma et al., 2018).

A ‘ground truth’ of images are a set of images of a specific phenotype. These images are checked in order to ensure the integrity of the image and this is a crucial step. By forming a ‘ground truth’ of images you know to be of a phenotype, the goal is to teach an algorithm this and let it adapt and develop the more data sets it can be process. Much like how we learn new information, by identifying the correct features when taught by a more experienced scorer and discarding other images, so the more you do it, the better you get. Thus, the more data the algorithm has at its disposal to analyse, the better it will be.

iv) Analysis and tools

a) Transfer Learning

i) ResNet Neural Network Use

Therefore, by ensuing that the images scored in the ground truth were accurate, we hypothesised that a smaller set of ground truth images could be used successfully to repurpose the ResNet50 neural network successfully.

By using an already established network, one could hypothesise that there was the potential to more accurately assess cellular images and place them into the correct corresponding category if the network had been trained on other images previously. As a result, this would lead to network familiarity with distinguishing images into differing categories (Warden, 2017).. As this principle is transfer learning, teaching the network 9 new classes would be achievable using a reduced bank of ground truth images was my hypothesis.

ii) 3 channel approach

The resulting approach was the use of 3 channels in the ResNet 50 model. Brightfield, Darkfield and Fluorescent channels were chosen. The Brightfield and Fluorescent channels are used in literature in IDEAS® imaging analysis to allow to distinguish artefacts from nuclear material for fluorescent stains and for the cytoplasmic integrity which can be shown in the brightfield channel, a significant limitation of the Metafer® semi-automated microscopy method (Verma *et al*., 2017, Verma *et al*, 2018). The Resnet model had limitations in repurposing the images used to initially define this network, which limited its use and led to the DeepFlow Network being developed and used going forward.

iii) Limitations

This 3-channel approach was taken forward, using ResNet50 as the network of choice and repurposing this network to score for cellular categories instead. However, the accuracy produced was not high enough to maintain the integrity the MN assay requires. Repurposing the network and adopting a transfer learning approach did not come to fruition, quite possibly due to the cellular categories being vastly different images to the images the network was originally trained on. This would lead to the network considering the cellular images to be more likened to one another and therefore not fully recognising each individual class. This possibly led to two sub groups of ground truth images being used by the network in assessing cellular images, one being the original images used to train the ResNEt50 network and the second being the cellular images used to create the cellular based ground truth in the IDEAS® program. This thus reduced the specificity of the network and therefore the integrity of the results. A scoring system which is not specific enough leads to false positives, which undermines the accuracy of the test.

Therefore, after the master template created in IDEAS® proved to be too specific, the ResNet neural network proved to be not specific enough. Just as the nursery rhyme goes, one automation attempt had been too ‘hot’, one had been too ‘cold’, the case was on to find the one which was ‘just right’.

iv) DeepFlow

To increase the network accuracy, the ResNet50 neural network repurposed approach was abandoned in favour of a newer, MN automation specific, neural network, DeepFlow. This network was created specifically to be trained on the ground truth generated. This was then tested with and without augmentation also. Augmenting the data can help to produce greater volumes of the rarer image categories by applying different measures (rotations, filters etc) to the images and thus creating multiple images from one. This can be extremely useful when only a small quantity of training images exists. However, did not prove useful in this scenario, and did not increase the accuracy.

b) Adobe Bridge®

Adobe Bridge® is a software tool commonly used by photographers to organise files and allows for renaming, assigning colour labels and star labels which allow for the files to be grouped together and analysed with like images. However, it was also found to be a very useful tool for grouping different cellular phenotypes in an efficient and accurate manner.

The starring and colour coding system for grouping images allows for the different cellular morphologies to be separated with great ease. The starring system can be used to designate how many nuclei are present in the image. Be it 1 for a mononucleate, 2 for a binucleate, 3 for a trinucleate and 4 for a quadranucleates. When the cellular morphology is ambiguous or looks to be dead, the image can be labelled with a 5. Moreover, the colouring labels allows to further differentiate these images of cells in these categories into if a MN is present or not. If a MN is present, then the image can also be colour labelled with one of 5 colours. When analysing the images post scoring, the images can be differentiated by colour and/or by star rating, allowing for an easy export f the data. Moreover, there is the opportunity to zoom in and out of the images, which allows for further inspection of a cell when scoring is taking place.

c) MATLAB®

To compensate for the differences in intensities and settings between imaging flow cytometers in different laboratories, whereby slight fluctuations in intensities and frequencies from various factors can make a difference, MATLAB® was used to normalise the images produced by the imaging flow cytometers. The output of the normalisation is 8bit images which can then be analysed in order to form the ground truth. Moreover, for the images to be shown in Adobe Bridge®, the images have to be a one channel image. This is different to the images input into MATLAB® whereby the images are 2 channels with the images being in 2 different layers, brightfield and nuclear. Therefore, before the cells can be analysed to form a ground truth population on Adobe Bridge®, the file containing the images needs to be converted to a single channel image using a code in MATLAB®.

i) General

MATLAB® is a coding computing tool, whereby users can run their code and form algorithms for use in both machine learning and deep learning. The simplicity of MATLAB ® compared to other coding programs provides a great advantage. MATLAB ® does not require a computational language for the processing, such as python or java, and therefore makes the use of it easier for the user. By learning the rules and abbreviations required for the process of MATLAB ®, an understanding of the tasks can be undertaken.

Moreover, MATLAB ® has many different toolboxes, including both machine and deep learning toolboxes, targeted to help individuals in the analysis and automation of data.

The 'Statistics and Machine Learning Toolbox' helps users apply functions and apps in the analysis and automation of data **(**MATLAB ST, 2020**)**. Moreover, statistics and plots are available for exploring the data and analysis. Visualisation and regression add-ons included allow the user to perform tasks more seamlessly with confidence in both 2D and 3D colourful graphs including scatter plots. This helps greatly in the ease of analysis whilst maintaining accuracy.

The 'Deep Learning Toolbox' helps users to apply and analyse different levels of neural networks. It can allow more basic users to use shallow neural networks, with not much depth, but can provide proof of concept data for the application of neural networks into new areas of research (MATLAB DL, 2020**)**. The application of deep learning neural networks is a more complex and computational heavy approach, which MATLAB helps to simplify when compared to other programs, without sacrificing the capability of the program to produce quick and accurate results. This toolbox provides pretrained models and applications as well as providing

convolutional neural networks, the type of network used in this analysis to automate the MN assay. This toolbox includes applications including the vital ‘Train Deep Learning Network to Classify New Images’. This is a vital tool in the deep learning field, allowing for a convolutional neural network to be taught for the specific choice of the user, in this case the identification of MN and cellular morphologies required to achieve an accurate dose response.

A convolutional Neural Newtwork (CNN) is made up of a series of layers, where each layer has a specific function. Once the image has gone through one layer, it connects to the next layer. The idea is to replicate how neurones work in the brain, however in the brain the neurones are attached to one another and to multiple other neurones. In this system, the layers proceed only to the next layer and the one previous. It is normally in the first layer, which is normally an image-Input-Layer which denotes the properties of the image which can be processed.

The next layers in the network are normally pooling, rectified linear units and repeating blocks of convolutional layers. These are the core layers of building a convolutional neural network which can help to confirm filter weights, although these may be changed during the training of the network. This can be used during the training of this network to give the rarer MN phenotypes greater weight as there are less images to train with (MATLAB DL, 2020)

The repeated blocks of convolutional layer acts to achieve a non-linear aspect to the network, allowing for an approximation to be made of non-linear functions which help to trace image pixels to the pattern of the image. Pooling layers allow for a downsample of the data as the network is flowing. However, caution must be taken with downsampling when using a more complex network with more layers as downsampling may take place too early and this can lead to the loss of important information from the image.

Moreover, this tool provides the user with the ability to adopt ‘transfer learning’. This can help by taking a pretrained network for use as an opening step for a new assignment, by updating a pre-existing network as opposed to creating a ne wone from scratch, the user saves time and convenience. The smaller quantity of training figure produced can then be transferred using their acquired features.

A conventional way to avoid learning and validating the same piece of data is to split the original data, it is commonly carried out at around a 3:1 ration of training images to validating images. This ensures that enough images are used to accurately train a network but allows for enough images for the validation to be accurate also and provides a wide enough range of images (depending on the overall quantity of images used).

Before the images can be trained, the code specifying the images has to include the properties of the images in order for these images to be correctly identified by the network. An image in the incorrect setting will not appear on the network.

ii) Validation and error rate

As the network is running to create a network, several features can be shown on the graph that is being produced. An accuracy and error rate are two of the main features that can be identified. By allowing the network more time to run at the beginning, the neural network has more opportunities to assess the images and form an assessment on what the likely category the image will belong in. At the beginning this error rate will be greater and the accuracy rate lower due to the network only having one opportunity to analyse the images and to learn from these. However, the accuracy rate should then increase and in tandem the error rate should decrease.

A low accuracy rate can lead to two main conclusions. The first is that the training data is not of a high enough quality and the network is becoming confused as it is not being taught with a great enough detail. In this scenario, the network itself is not the issue but the training images. The issue can be seen more specifically if the issue is specified to only one or a couple of subgroups or if it the entire data set. This can be determined by analysing the confusion matrix (more detail will be provided of this shortly). It can be shown on the confusion matrix the accuracy of each individual subgroup and where the accuracy may be lacking. By analysing the images prdocued by certain sub-groups of the confusion matrix, the user is able to identify if the training data is up to a sufficient quality as these images will be manually assessed

by eye and are open to counter-checking by other users, ensuring the integrity of the training data.

The second issue may be that the training data itself is of a good standard but the issue stems from the network itself. This could be caused by too many pooling layers which may have taken vital information from the images and caused confusion. If there is an issue with the network, it would most likely be expected that each subgroup would show a lower accuracy rate and this thus resulting in a decreased overall accuracy rate.

Lastly, the issue could an error with the network in combination with a lower quality or quantity of training images. In this case, each subsection would have to be identified and rectified MATLAB DL, 2020).

iii) Epochs and Batch/Mini batch

When training the network, it is important to consider the few factors based on the length of training. To maximise accuracy, it is not always beneficial to allow for the network to train for the greatest possible time-period to achieve the maximum accuracy. This may seem counter intuitive initially, however, after training for too long a period and more specifically too many epochs, the error rate can increase. This increase takes place as when training for an extended period of epochs, it is possible for the network to focus on specific features which are not true markers of the particular subgroup due to overtraining, akin for looking for patterns when they are not there in more popular culture. An epoch is defined as a complete training cycle on an entire data size.

The batch size refers to the quantity of samples which will be put through the system at a time. The network runs through the number of images designated in the batch size and trains the network. It then goes through the second batch of images until it has gone through all the images in a sample, which is designated as an epoch. As an example, a batch size of 100 with an image bank of 1000 would go through 10 iterations of images, each hundred, before training the network. Were the batch size 50, this would increase to 20 iterations.

There are both pros and cons to using greater or smaller batch sizes, with a balance needed. A smaller batch size requires less memory and thus less computational power is needed and a greater spectrum of computers can undertake the analysis. Moreover, when using mini-batch sizes, the network tends to train at a quicker rate due to the weights being updated after each iteration.

The disadvantage is that the smaller the batch, the lower the accuracy shown of the gradient accuracy, with many more peaks and troughs shown due to each training step taking place after a smaller quantity of samples and therefore more fluctuation caused.

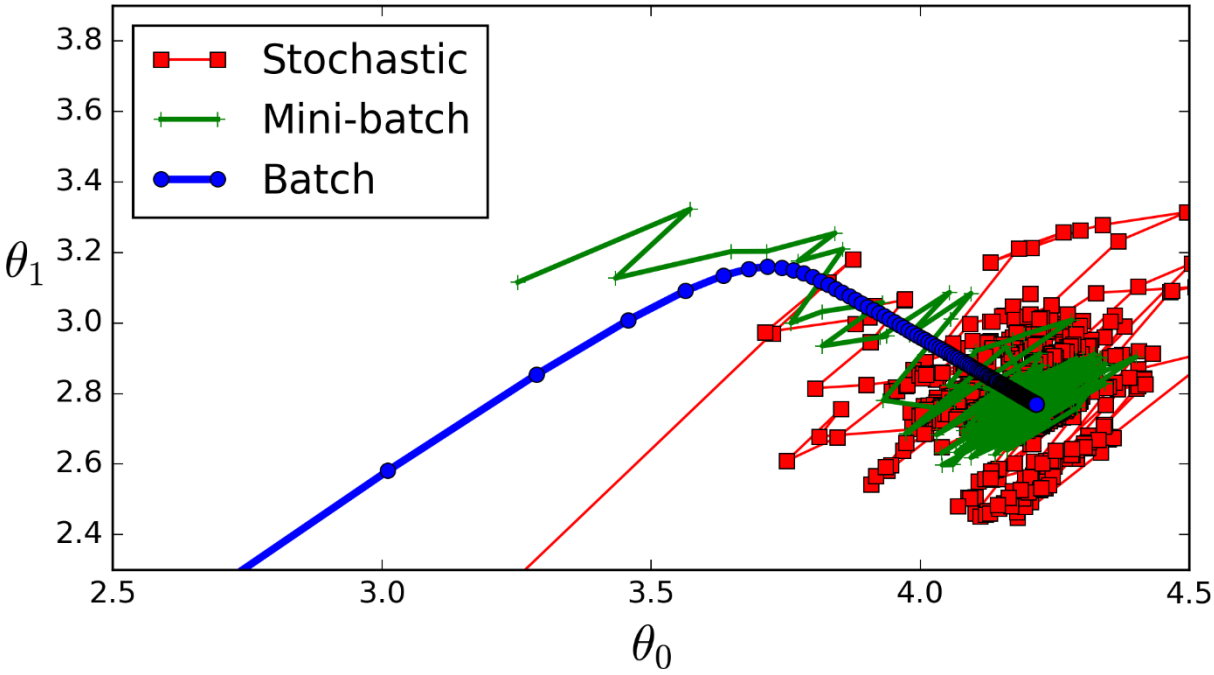


Figure. 4 Schematic showing the difference between, stochastic, batch and mini-batch. 4 Far greater fluctuation seen in the mini-batch graph as opposed to the batch (Cross Validated, 2020).

So, a balance must be struck when using batches and mini batches. Ideally, a greater batch size would be used, if computational power and time were not limiting factors in the analysis, as this leads to more images being sampled at a time and thus greater accuracy for each iteration. However, this is not always possible and thus the balance needs to be struck between computational power, time and accuracy.

The validation is then plotted once per epoch and the change in accuracy as the epoch quantity increases. The quantity of epochs to let the network run for can be pre-set before once the optimal number is known. In order to recognise the optimal

number of epochs needed to run the network, a free run must be carried out and the number of epochs denoted before the error rate begins to increase.

v) Confusion matrix

A confusion matrix is a table produced which summarises the performance of a classification algorithm. By producing a table detailing the accuracies of each sub class, the user can identify which classes are performing well and which class requires improvement.

By producing an overall accuracy and error rate, it is possible to compare each class to the average accuracy and thus determine which class is performing better/worse than the average. It is important to take into account the quantity of images in each class, if one class makes up the majority of the images and accuracy, it could be possible that the overall accuracy is not as good as suggested as only one sub class may be performing and thus inflating accuracy levels.

By showing the user the accuracy and error rates of each sub class, as well as a total accuracy and error rate, it is possible to distinguish where each sub class may be scoring incorrectly and this can help the user to identify issues and to implement ideas to improve the total accuracy.

Thus, by displaying the data in a table displaying the accuracy rates of each class, the user can easily identify areas of strength and weakness and look to rectify these for greater future accuracy (MATLAB DL, 2020).

1.6 Aims and Objectives

i) Different Laboratories

The combination of the different laboratories helped to contribute to the ground truth and carrying out the dose response analysis. The Cardiff and Cambridge formed the ground bank of cellular images which formed the ground truth. Moreover, a Cardiff dataset was also used in forming a dose response. A separate dataset was used from the Imaging flow cytometer at the Newcastle Laboratory, and this was used in a dataset and compared with the Cardiff dataset to carry out a dose response analysis of the non cyto-B MN assay. Lastly, the GSK dataset was used in calculating a cyto-B MN dose response comparison between manual scoring and the deep learning automated method developed and analysed in this project. Moreover, this GSK manually scored dataset has the potential to be used in future to create a larger bank of ground truth images. The use of these different laboratories allows for reproducibility to be demonstrated across data formed across different laboratories.

ii) Aims

Therefore, the aim of the project was to develop a method to automate the laborious and time-consuming nature of the traditional MN assay (manual scoring), without compromising the accuracy shown traditionally in what is the gold standard.

By using and applying a deep learning neural network approach to this issue, images can be manually scored by a user and divided into groups to form a 'ground truth' of images. This ground truth, is then used to teach the neural network the parameters of the cellular images to assess in order to carry out the scoring of the assay. Having this form of ground truth removes any user subjectivity to the results. Moreover, using different laboratories in the creation of this ground truth, allows for the method to be an inter-laboratory application, capable of making the assay high throughput and accurate.

Developing a system whereby a high throughput is developed, whilst maintaining the gold standard accuracy, would streamline the assay and allow for greater use yet of this and transform its use into a truly 21st century approach.

2.0 Materials and Methods

i) Chemicals

Carbendazim (Cas no. 10605-21-7), purchased from Sigma-Aldrich. The working concentrations, 0.00, 0.40, 0.60, 0.80, 1.00, 1.20 and 1.60 (µg/ml) were selected based on the data produced by Verma et al., 2017. Not all concentrations were in specific laboratories, where only three doses were used in addition to a control.

ii) DNA staining

DRAQ5™ DNA (Cat. No. 564902 supplied from BD Biosciences) was used to label nuclei and MN for the Cardiff and GSK laboratories . Samples incubation time for DRAQ5 was a minimum of 20 minutes.

Hoeschst 33342 (Cas No. 87576-97-1, supplied from Sigma-Aldrich) was used to label nuclei and MN for the Cambridge laboratory.

These datasets were used in the creation of the ground truth.

These staining events were previously carried out and the images of the cells used and re-purposed in order to create the ground truth of the different cellular morphologies.

iii) Cell lines and treatment

Human lymphoblastoid TK-6, AHH-1 and MCL-5 cells were obtained from American Type Culture Collection (ATCC), Manassas, VA, USA. AHH-1 and MCL-5 cells have a doubling time of 22-24 hours. The cells were cultured in RPMI 1640 media (Gibco, Paisley, UK), supplemented with 1% Glutamine (for MCL-5 cells specifically 40 µg/mL Hydromycin) and 10% heat inactivated horse serum (Gibco, Paisley, UK). Cells were seeded at $1x10^5$ cells/ml in $25cm^2$ flask (Fisher brand), incubated at 37°C in a humidified atmosphere of 5% (v/v) $C0_2$ and established into subcultures once confluence was reached. When carrying out the experiment, cells were seeded at $2x10^5$ cells/ml in $25cm^2$ for 1.5-2 cell- cycles in the presence of genotoxic agent with no recovery.

Tk6 cells have a doubling time of 13-15 hours (Lorge *et al*., 2016). The cells were cultured in RPMI 1640 media (Gibco, Paisley, UK), supplemented with 1% Pen/Strep (100 U/mL Penicillin and 100 µg/mL Streptomycin) and 10% heat inactivated horse serum (Gibco, Paisley, UK). Cells were seeded at $2x10^5$ cells in $25cm^2$ flask (Fisher brand), incubated at 37˚C for 1.5-2 cell- cycles.

The *in vitro* MN assay was used to assess MN formation in the TK6, AHH-1 And MCL-5 cells with following treatment with Carbendazim, or in the case of the MCL-5 and AHH-1 cells; a water and methanol control. Cells were treated, and then incubated for 1.5-2 cell cycles. After incubation, the cells were centrifuged at 200xg for 10 minutes in preparation for harvesting. The supernatant was then removed, and the pellet re-suspended in 10mL of phosphate-buffered saline (PBS) (Gibco®).

iv) Data acquisition on the imaging flow cytometers and IDEAS analysis®

After the removal of the test chemical, cells were washed with PBS and fixed with BD FACS™ lysis solution (CAS- 349202) ratio FACS lys:dH2O 1:10 to allow for membrane permeabilization and cell fixation. Two ml of the fixative was added to each pellet and incubated at room temperature for precisely 12 minutes. The cells were then spun at 200xg for 10 minutes and the staining process was carried out. During this staining process, the nuclei and MN are stained with 0.05mM DRAQ5™ (CAS-564902, BD Biosciences) and incubated at room temperature prior to acquiring the images with the FlowSight® and ImageStreamX-MkII®.

The 488nm laser, found as part of FlowSight®, and the 405nm, used in the ImageStreamX-MkII®, are both equipped with at least three lasers and were used to excite DRAQ5/Hoescht 33342 stained cells. Hoescht 33342 was used for the Cambridge laboratory, Draq5 was used for the: Cardiff, Newcastle and GSK laboratories. This led to images being captured automatically with the use of INSPIRE® 3.0 software. In order to acquire these images, 80µl of the stained cell suspension was inserted into the FlowSight® or ImageStreamX-MkII® and the resulting DRAQ5/Hoeschst 33342 cellular images were processed, along with brightfield images to enable later analysis.

Aspect ratio and area features were used from the brightfield images to gate out any debris or cells which had died from the population as they would be irrelevant in this analysis. Altogether, 20,000 single cells were captured per dose per replicate.

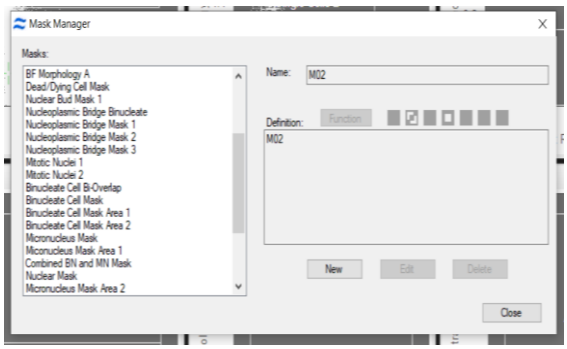
MN frequency was obtained by scoring 2000 Mononucleated cells per dose for the non-cytochalasin-B MN assay and 1000 Binucleated cells for the cytochalasin-B MN assay where possible using FlowSight imaging flow cytometer. The data was then saved as a raw image file and analysed in IDEAS® version 6.2 using either manual scoring of images or the creation of templates. For analysis to take place on IDEAS®, the raw image files (rif) are converted to compensated image files (cif) and then to data analysis files (daf).

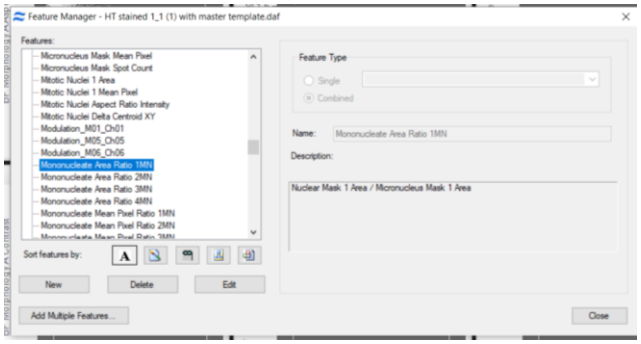
Once on IDEAS®, manual scoring of previously scored data was checked and 'pulled out' to form a population of a specific phenotype. The 'master template' was formed using masks and features on IDEAS® and was updated as required for each different data set. This template was used to help to 'pull out' different populations with more ease and increase the 'ground truth' pool.

By brainstorming the specific morphologies of different cell types, it was possible to use masks and features on the IDEAS® software to differentiate between these cell types and therefore form multiple templates within the master template for 'pulling out' the different cellular populations.

A

B


Mask Manager
Masks:
BF Morphology A
Dead/Dying Cell Mask
Nuclear Bud Mask 1
Nucleoplasmic Bridge Binucleate
Nucleoplasmic Bridge Mask 1
Nucleoplasmic Bridge Mask 2
Nucleoplasmic Bridge Mask 3
Mitotic Nuclei 1
Mitotic Nuclei 2
Binucleate Cell Bi-Overlap
Binucleate Cell Mask
Binucleate Cell Mask Area 1
Binucleate Cell Mask Area 2
Micronucleus Mask
Micronucleus Mask Area 1
Combined BN and MN Mask
Nuclear Mask
Micronucleus Mask Area 2
Name:
M02
Definition:
Function
M02
New
Edit
Delete
Close


Feature Manager - HT stained 1_1 (1) with master template.daf
Features:
Micronucleus Mask Mean Pixel
Micronucleus Mask Spot Count
Mitotic Nuclei 1 Area
Mitotic Nuclei 1 Mean Pixel
Mitotic Nuclei Aspect Ratio Intensity
Mitotic Nuclei Delta Centroid XY
Modulation_M01_Ch01
Modulation_M05_Ch05
Modulation_M06_Ch06
Mononucleate Area Ratio 1MN
Mononucleate Area Ratio 2MN
Mononucleate Area Ratio 3MN
Mononucleate Area Ratio 4MN
Mononucleate Mean Pixel Ratio 1MN
Mononucleate Mean Pixel Ratio 2MN
Feature Type
Single
Combined
Name:
Mononucleate Area Ratio 1MN
Description:
Nuclear Mask 1 Area / Micronucleus Mask 1 Area
Sort features by:
New
Delete
Edit
Add Multiple Features...
Close

C

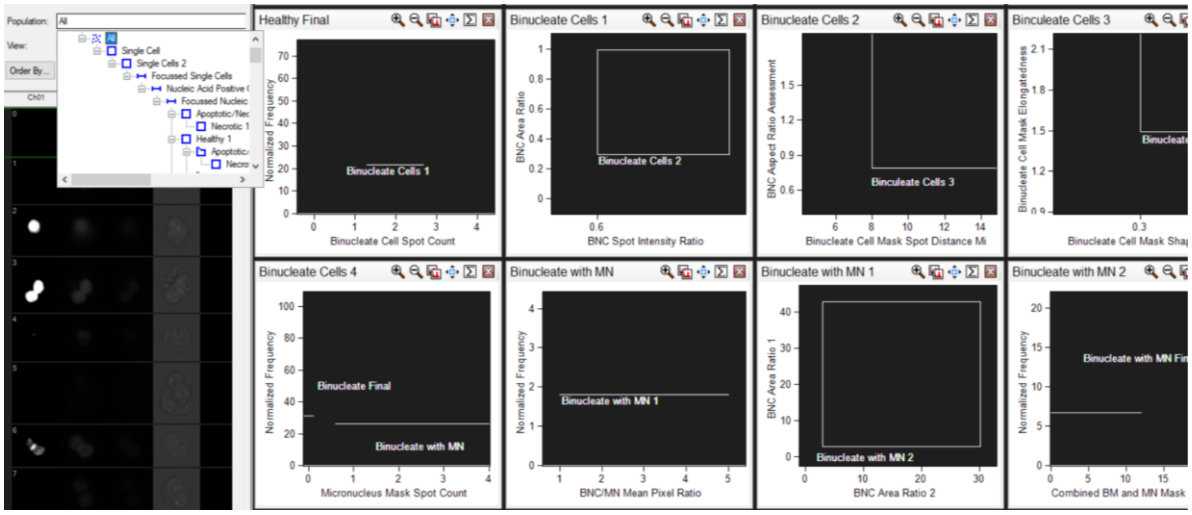


Figure. 5 An overview of some of the masks and template used in the IDEAS® software for the master template creation in an attempt to automate the MN assay through this method.
A) Shows a portion of the variety of masks used within the analysis of the master template.
B) Shows a portion of the variety of features used within the analysis of the master template.
C) An example of how specific populations of cells can be differentiated from one another using the characteristics of a particular phenotype, in this case how binucleated with MN cells would be differentiated from binucleated cells.

v) IDEAS® based Ground truth

This ground truth was generated in two different ways. One way was to score images of the cells manually as has previously been undertaken (Verma et al., 2018). The other method was to manually score a selection of cells via IDEAS® and to confirm the cellular phenotypes. This was undertaken until a master template was formed on the IDEAS® software. A master template was formed by a variety of masks and features which are found in IDEAS®. These masks and templates helped to differentiate cellular phenotypes and allow for different populations to be 'pulled out'. Like much in the same way that Rodrigues attempted to automate the MN assay, we used an approach which is likened to an extent (Rodrigues, 2018). However, each data set of the 'pulled out' populations was analysed manually to ensure that they belong in the phenotype assigned. These formed part of the 'ground truth'. The advantage of forming a ground truth via this method as opposed to

carrying out the entire experiment as such is that an underestimation of MN which may occur is not problematic as there are more data sets from which to analyse more MN.

This master template was tweaked to work from different data sets due to different laboratories providing data and using different lasers, thus subtle changes will be needed to the masks and features to normalise the different datasets. This is as different laboratories have different channels opened and corresponding to the frequencies; thus the 'nuclear channel' may not be in the same channel across the different laboratories and the subsequent mask would have to be edited to ensure that a nuclear mask would be encompassing the nuclear channel for the specific dataset. This includes adapting the channels set out in the masks based on the laboratory due to different channels used (The fluorescence channel may be channel 5 in some laboratories and then channel 11 in others etc). Different nuclear dyes used appear in different channels during imaging flow cytometry analysis and thus the same master template cannot be used for these differing data sets.

The aim was that the deep-learning algorithm will replicate the results produced in historical data. If this occurs, then this project could help to lead a significant advance in increasing the throughput of the MN assay and revolutionise the assay truly.

vi) IDEAS® analysis

In an effort to reduce the laborious and time-consuming approach of manually scoring MN on the IDEAS® software, a master template was created to produce a level of automation in MN analysis. This approach was coined using the masks and features tools available in IDEAS®. By producing a template which could identify the features of a MN, the other cells would be gated out and a MN % obtained, allowing for MN analysis and thus insights into chromosomal damage. This approach was based on the Rodrigues approach (Rodrigues, 2019).

vii) Training the network

To train the network, two sets of ground truth images were first analysed using the Adobe® Bridge platform. For the images to be in the correct format in Adobe®Bridge, they must be in an 8-bit tiff format. To generate this tif, the IDEAS® program is opened and a raw image file (.rif) generated by the inspire

software is opened. This automatically produces a compressed image file (.cif) and a data analysis file (.daf). The .CIF file directory was copied and applied to a MatLab® code. This MatLab® file is used to convert a cif file to the individual tif file containing the cellular images:

Final_script_3_channel_from_cif_to_tif.m

This script is opened on MatLab® and the cif directory is pasted onto the relevant line for validations. The result of this is the creation of tifs. These tifs then need to be converted to a 1 channel tiff in order to be visualised using the Adobe®Bridge platform, as they are in 2 or 3 channels at the minute.

The:

'bridge_script_1_channel_for_bridge_just_one_input_file.m'

script allows for tif 1 channel conversion. The images are saved in an updated one channel folder and ready to be analysed.

viii) Bridge analysis

The images in a single channel tiff format, obtained from the cif file conversion were opened on Adobe®Bridge which is a file manager where images can be marked with 5 different colours and 5 different stars. The images were scored according to the following criteria:

A single * denotes a mononucleated cell.

Two stars ** denote a binucleated cell.

Three stars *** denotes a trinucleated cell.

Four stars **** denotes a quadranucleated cell.

5 stars, denotes a cell which is either dead (apoptotic or necrotic) or unscorable, this category is known as 'others'.

Colours are also used to annotate these images. By pressing a number 6, a yellow colour is denoted. This is used to signify a MN, so that cells with MN are split by how many main nuclei are present in the cell also. A mononucleated cell with a MN would be signified by a single star * and the colour yellow. When analysing the cells, the user can distinguish between all cell classes and those in the class with and without a MN, as well as being able to view the total MN number, regardless of the number of main nuclei in the cell.

2.1 Running the network

i) Creation of .cif files in IDEAS

In order to run the network, the raw image file (.rif), generated using the INSPIRE software from the imaging flow cytometer, is opened in IDEAS®. When the user opens the .rif file in IDEAS®, a compressed image file (.cif) and a data analysis file (.daf) are automatically generated. The .daf file is the file which allows the user to view the data on IDEAS®. It is using this .daf file, where the user makes a note of which channels are used for: brightfield, darkfield and DNA fluorescence. It is important to make a note of this as the channel order differs based on the imaging flow cytometer machine and the DNA stain used. It is this .cif file which is needed when running the script to determine a dose response.

ii) Generating tif files for Deep Learning

1. Open MATLAB and open the script ‘final_script_3_channel_from_cif_to_tif.m’

2. Make all the bioformat files are in that directory together with the image padding script

3. In the script make sure to change the channel numbers to get the right images in

4. Change the cif file name/location in the script and choose a directory to store the images

NOTE: – in all the work carried out so far, the tiff was ordered BF, DNA, DF

iii) Test images on a previously trained network

1. Open MATLAB and open the script ‘explore_output_v2.m’

2. Alter the network model parameter file in the script, it is a .mat file which currently looks like this ‘DeepFlow25-Feb-2020-17-47-23-Epoch-100.mat’

3. Change the directory for the files to test i.e. imageFolder_Validate ='…….’

a) Stats

When carrying out a dose response, it was integral to carry out analysis to check if the data was normally distributed or not in accordance with the methods highlighted in Johnson *et al.*, 2014. The initial test used to assess the normality of the data is a Shapiro-Wilk test, where a P value of >0.05 equated to the data being normalised. If this was <0.05, then a Bartlett’s test can be carried out to assess normality in the data, with a P value of >0.05 showing the data being normalised. The log of the data and square root can also be obtained for use in assessing normality of the data if required.

To carry out a one-sided Anova, a Dunnett’s test in this case, the data must be normally distributed. This is as the one-sided Anova calculates variation between dosed samples and the control and thus relies on data to be normally distributed in order to calculate this. If the P value is <0.05, then the dose is of significance

compared to the control, if the P value is >0.05, then the dose is not significant compared to the control.

3.0 Results

The original methodology for the creation of the ground truth was to use the IDEAS® program, which allows for the differentiation of cellular categories by the use of masks and features as displayed in Figure. 5, Figure. 6 and Figure. 7.

Limitations with proceeding in this method led to the ground truth creation taking place by using the Adobe®Bridge approach for manually placing different images of cells into different categories, Table. 2 formed the basis of the categories to use and images to be analysed manually. The result of this manual scoring of the different categories of cells and their quantities is displayed in Figure. 9, Figure. 10 and Figure. 11. A further dataset which could be used in future work shown in Figure. 12.

The Cambridge ground truth was the first ground truth to be generated using this approach. Fig. 9 shows the distribution of the Cambridge network with 9178 cells scored in total. This is a far smaller figure than is used in literature for neural network training, where millions of images are routinely used to accurately train networks. The golden rule has historically been, to use 1000 images per categories, and in many cases, 1000 categories are used, yielding the million images used which was eluded to earlier (Warden, 2017).

As shown in Table. 3, where over 99% of cellular images are being assessed into the others category, the 3-channel: brightfield, fluorescent, darkfield approach could not be used going forward as the network was being confused and not allowing for

useful information to be obtained as a result. This led to the 3-channel approach to be used, consisting of, a singular brightfield channel with two fluorescent channels. This was then further reformed to a two-channel approach, consisting of a single brightfield and fluorescent channel in use, with the increase in accuracy displayed form Table. 4 to Table. 5. This 2-channel approach was carried forward, following determination of the most accurate networks derived from Table. 5.

Network 3 therefore proved to be the more accurate network across all these categories, which on first glance makes it seem like the Cardiff dataset is the more accurate of the two (since both were validated on the Cambridge ground truth). However, on reflection, the two networks were trained using differing levels of epochs, with Network 3 using the optimal 20 epochs and Network 8 using the sub-optimal 30 epochs (Table.4). Given the presence of over-training, it would be expected that a larger disparity would have been shown between the two networks than a 0.4% difference in overall accuracy (Table.4). The MN accuracy differences are both under 10%, with an 8.4% difference in binucleated MN cell accuracy and a 1.9% difference in mononucleated MN accuracy (Table.4).

Network A displaying the highest accuracy for Mononucleated MN and thus what would be used for carrying out the non cyto-B MN assay. Network D displayed the highest accuracy levels for Binucleated cells with MN and thus this network would be used for the determination of a cyto-b MN assay.

A dose response was therefore carried out, using both the cyto-b MN assay and the assessment of Binucleated cells and the non cyto-b MN assay and the assessment of Mononucleated cells and this is shown in Figure. 14, Figure. 15 and Figure. 16. A manual assessment was carried out using the cyto-b MN assay and this was compared to chosen networks deemed to have the highest accuracy as a comparison, including Network D and compared in Figure. 14. The manual scoring of mononucleated cells was not carried out, instead a dose response was generated using the more accurate neural networks as was carried out for the Binucleated cells (See Figure. 15). Moreover, a comparison of the Cardiff and Newcastle Carbendazim data-sets was carried out and compared to the historical background rate of MN in non cyto-B MN assays in literature (See Figure. 16). Figure. 17 Shows a proof-of-

concept analysis, using a Tk-6 cell derived neural network to assess MCL-5 and AHH-1 cells.

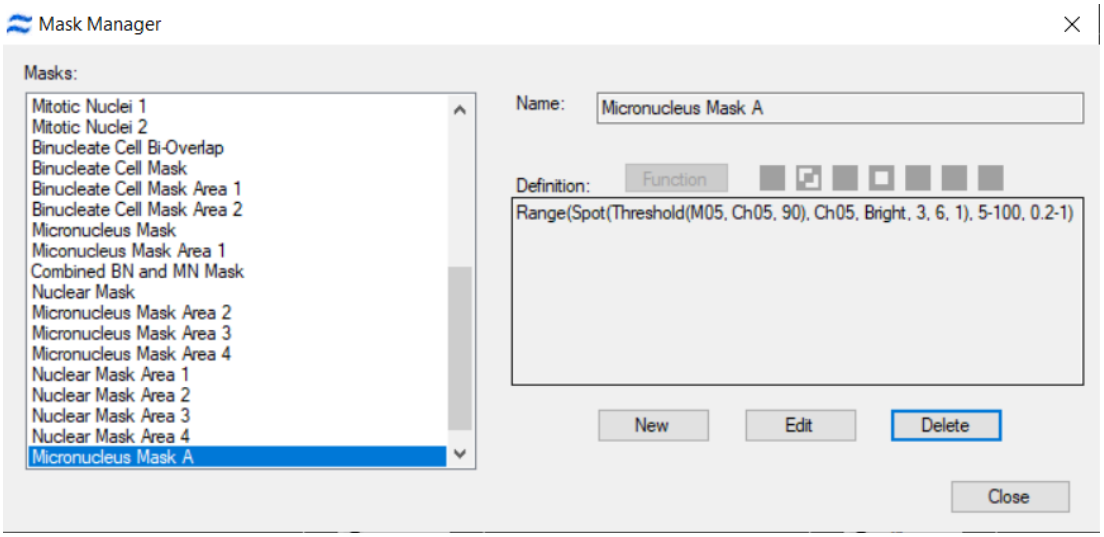


Fig. 6 An example of the components of a mask used in the IDEAS® attempted automation of the ground truth. The channels set for the mask must be manually adjusted for each differing laboratory. Example of a mask used in the master template development.

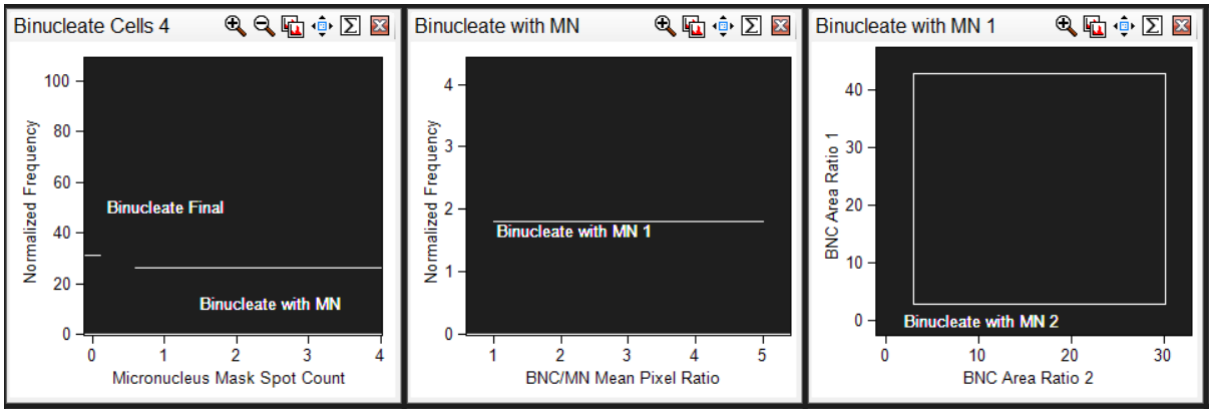


Fig. 7 Progression of part of the master template, detailing the Binucleate MN template. The specificity becomes greater with each graph, as each graph gates off more and more MN based on Fenech's description of MN's characteristics (Fenech, 2000).

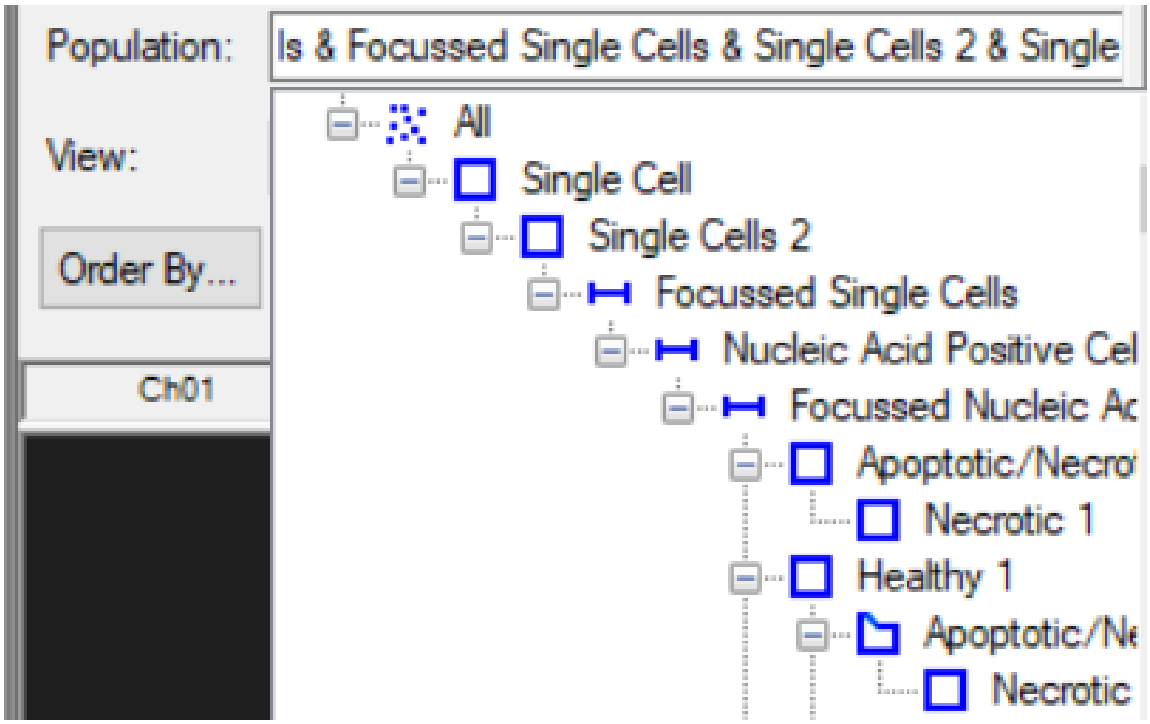


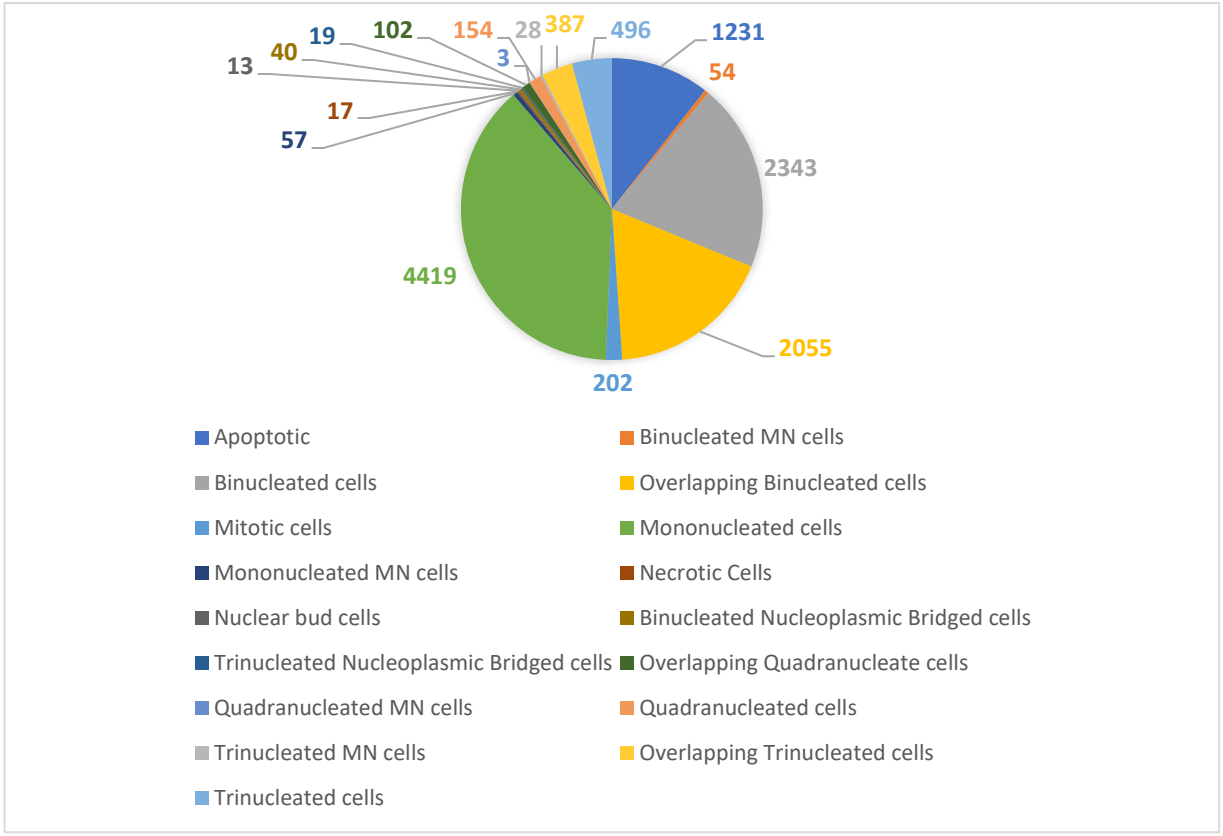
Fig. 8 Example of the flow found differentiating cellular groups in IDEAS®. This allows the user to identify which population they wish to view. A feature useful when calculating the MN% using IDEAS® and helpful in initial ground truth formation

i) Ground truth grouping

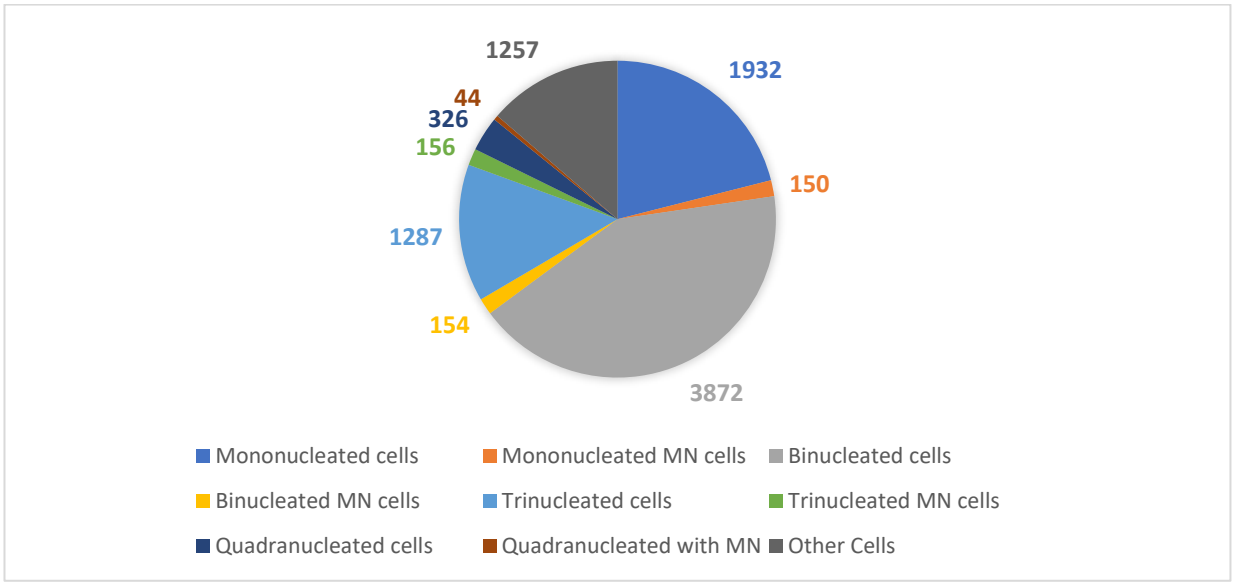
Cellular Phenotype	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Sum Total
Apoptotsis	289	477		495					1261
Binucleates	909	1381		53					2343
Binucleates with MN	5	11	8	1	7	10	4	9	55
Binucleate-Overlapping	651	558		846					2055
Mitotic	139	41		22					202
Mononucleates	1222	1689		2025					4936
Mononucleates with MN	5	19	5	6	8	4	3	6	56
Necrotic	8	20		13					41
Nuclear buds	3	1	1	2	4	1	0	0	12
Nucleoplasmic Bridge Binucleate	11	0	6	3	11	5	4	0	40
Nucleoplasmic Bridge Trinucleate	2	1	2	2	4	5	1	2	19
Nucleoplasmic Bridge Quadranucleate	0	0	0	0	0	0	0	0	0
Quadranucleate	30	25	15	18	17	9	16	24	154
Quadranucleate with MN	2	0	0	0	0	1	0	0	3
Quadranucleate Overlapping	22	12	7	18	11	3	12	17	102
Trinucleate	147	142		207					496
Trinucleate wih MN	6	5	2	3	3	1	6	3	29
Trinucleate Overlapping	56	71	32	69	12	19	59	69	387
									12191

Table. 2 The phenotypic breakdown of the initial ground truth created using the IDEAS® program in the original automation attempt, using the Cambridge data set with Hoescht 33342 as the DNA stain. This formed the basis of the ground truth image bank before analysis using Adobe®Bridge.

A


1231
54
2343
2055
202
4419
57
17
13
40
19
102
154
28
387
496
3
Apoptotic
Binucleated MN cells
Binucleated cells
Overlapping Binucleated cells
Mitotic cells
Mononucleated cells
Mononucleated MN cells
Necrotic Cells
Nuclear bud cells
Binucleated Nucleoplasmic Bridged cells
Trinucleated Nucleoplasmic Bridged cells
Overlapping Quadranucleate cells
Quadranucleated MN cells
Quadranucleated cells
Trinucleated MN cells
Overlapping Trinucleated cells
Trinucleated cells

B


1257
1932
44
326
156
150
1287
154
3872
Mononucleated cells
Mononucleated MN cells
Binucleated cells
Binucleated MN cells
Trinucleated cells
Trinucleated MN cells
Quadranucleated cells
Quadranucleated with MN
Other Cells

C

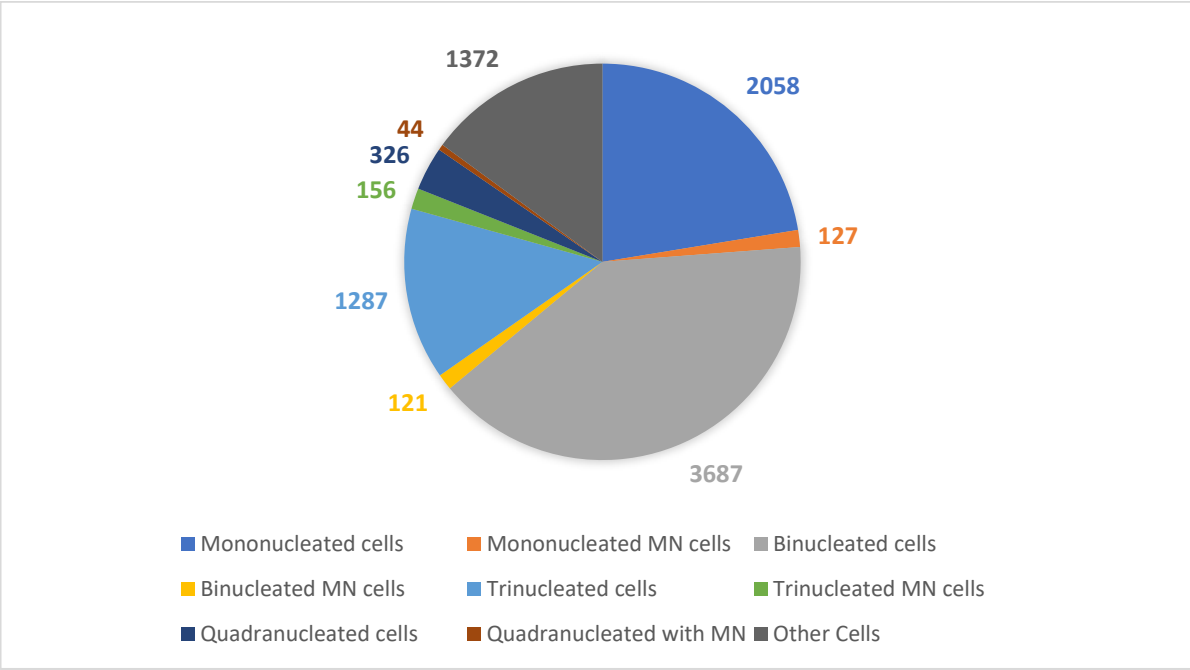


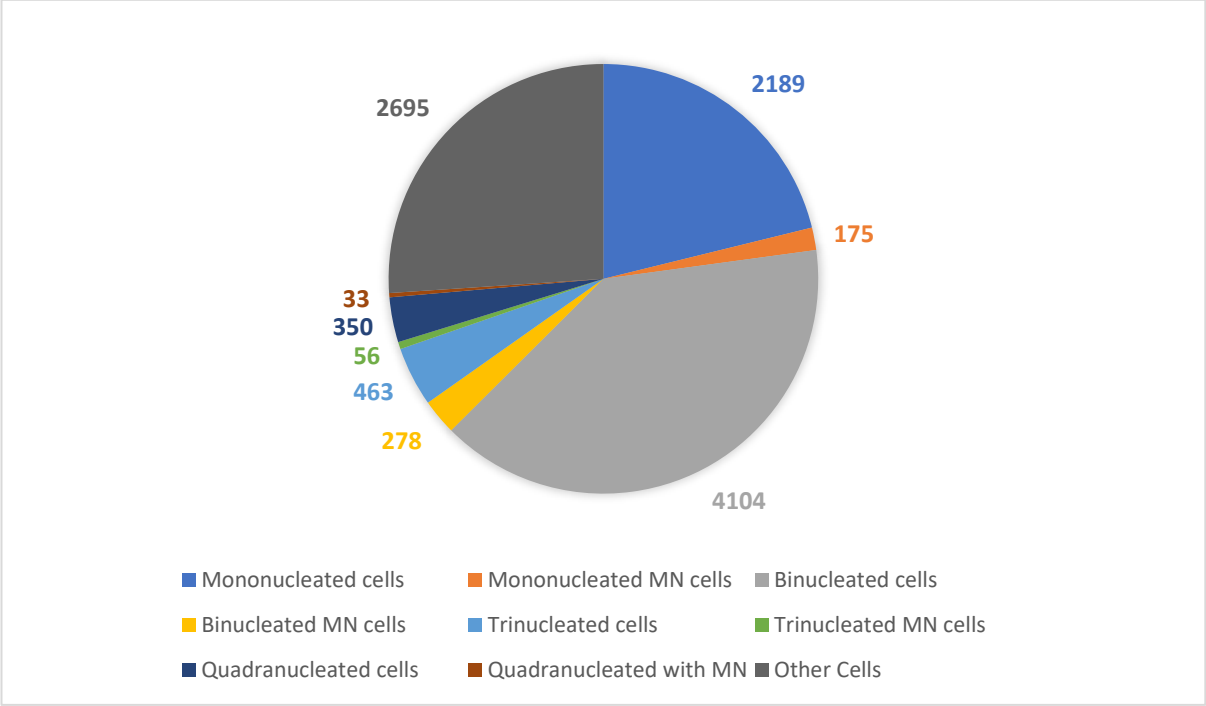
Figure. 9 Distribution of manually scored cellular images from the Cambridge data set. Cellular images manually analysed using AdobeBridge® software based on cellular phenotype. Data represented as %'s in the Pie Chart, however, depicting quantities of cellular images. 9178 cells scored in total in B+C. These cellular splits were used in the creation of the neural networks, which were used to asses dose responses in the MN assay.

A) The original distribution of the Cambridge data set, after initial IDEAS® grouping, featuring a wider range of phenotypes to which the cells were attributed to.

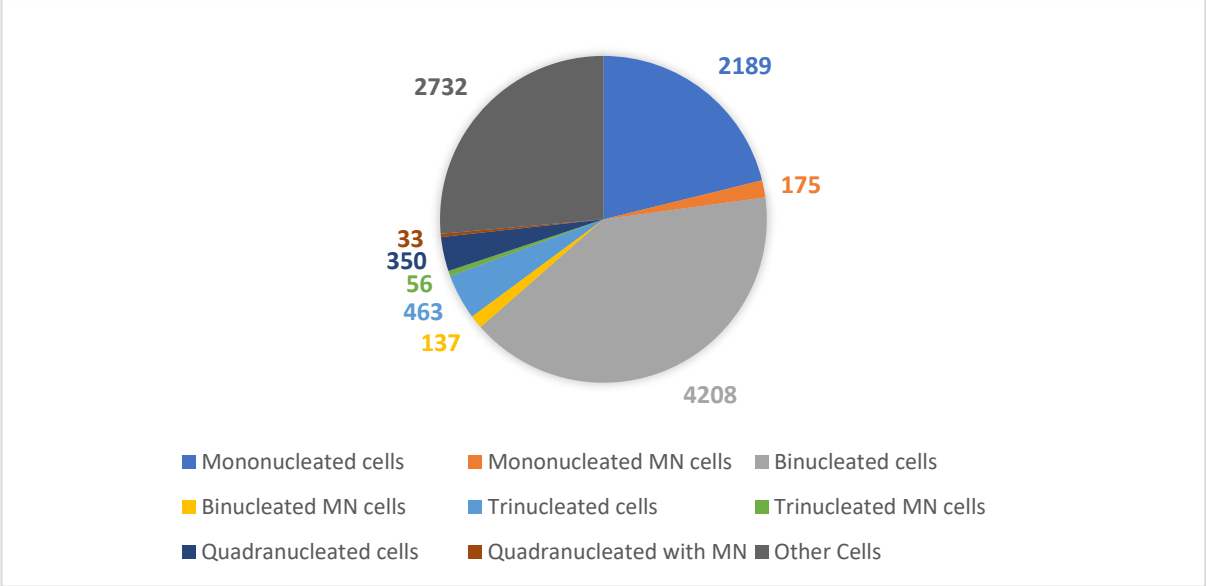
B) Distribution of the cells following initial analysis on AdobeBridge® and regrouping. 'Others' category added and increased accuracy.

C) Distribution of the cells following re-analysis of all MN validated from the network using MatLab® and the confusion matrices to pinpoint potential mis-classifications. Most accurate and latest ground truth.

A


2189
175
4104
278
463
56
350
33
2695
Mononucleated cells
Mononucleated MN cells
Binucleated cells
Binucleated MN cells
Trinucleated cells
Trinucleated MN cells
Quadranucleated cells
Quadranucleated with MN
Other Cells

B



C

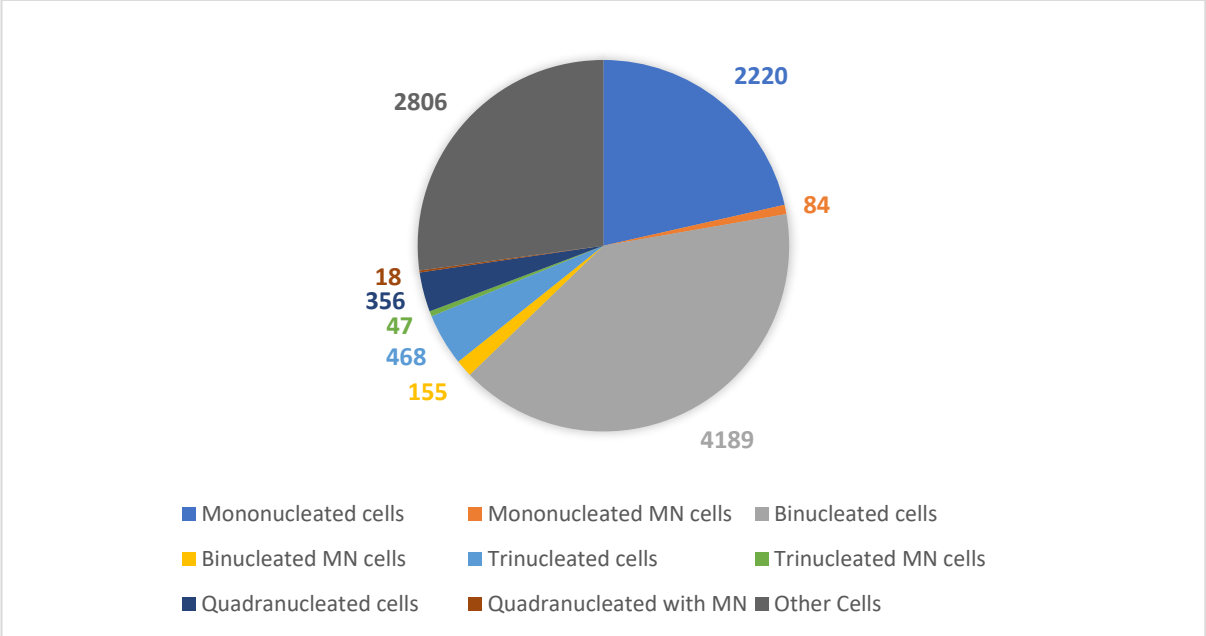


Figure. 10 Distribution of manually scored cellular images from the Cardiff data set. Cellular images manually analysed using AdobeBridge® software based on cellular phenotype. Data represented as %'s in the Pie Chart, however, depicting quantities of cellular images 10,343 scored in total. These

cellular splits were used in the creation of the neural networks, which were used to asses dose responses in the MN assay.

A) The original distribution of the Cardiff data set
B) Distribution of the cells following analysis on AdobeBridge® and regrouping.
C) Distribution of the cells following analysis of all MN produced from the network using MatLab®. Most accurate ground truth

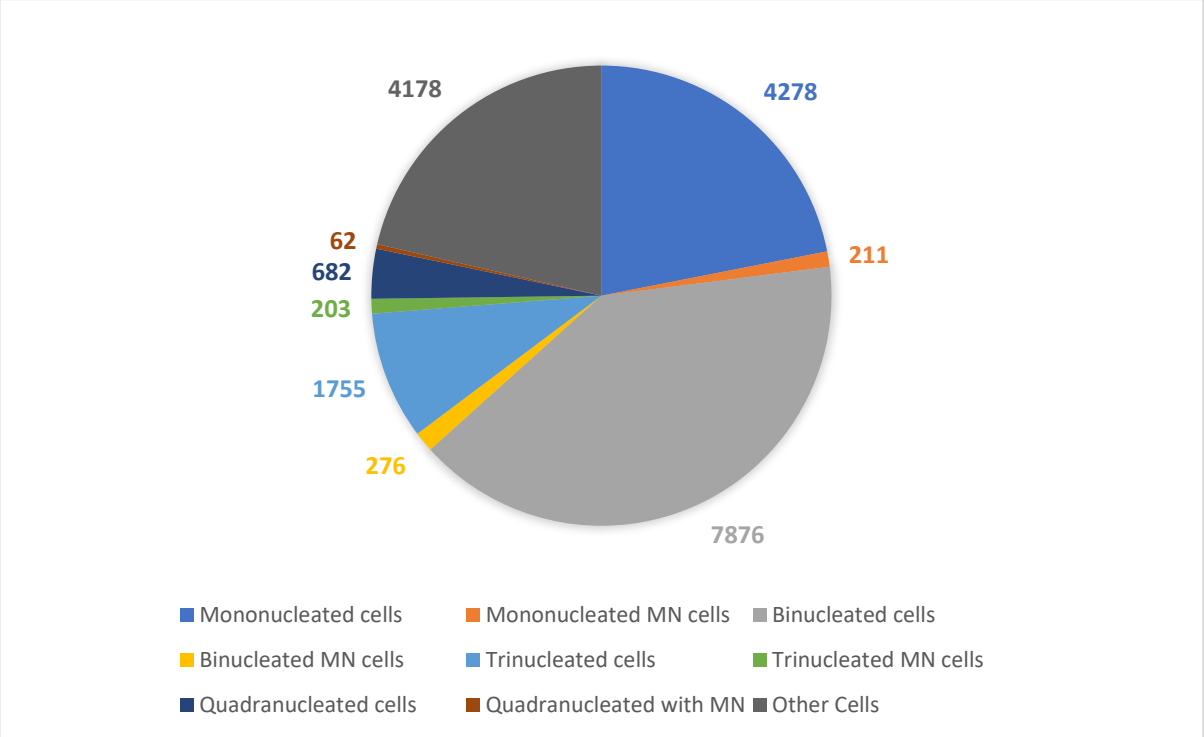


Figure. 11 Combination of the distribution of manually scored cellular images from the Cambridge and Cardiff data sets combined. These data-sets combined are used as another dataset in the training and validation of the networks. Cellular images manually analysed using AdobeBridge® software based on cellular phenotype. Data represented as %'s in in Pie Chart, however, depicting quantities of cellular images. 19,521 cells analysed in total. These cellular splits were used in the creation of the neural networks, which were used to asses dose responses in the MN assay.

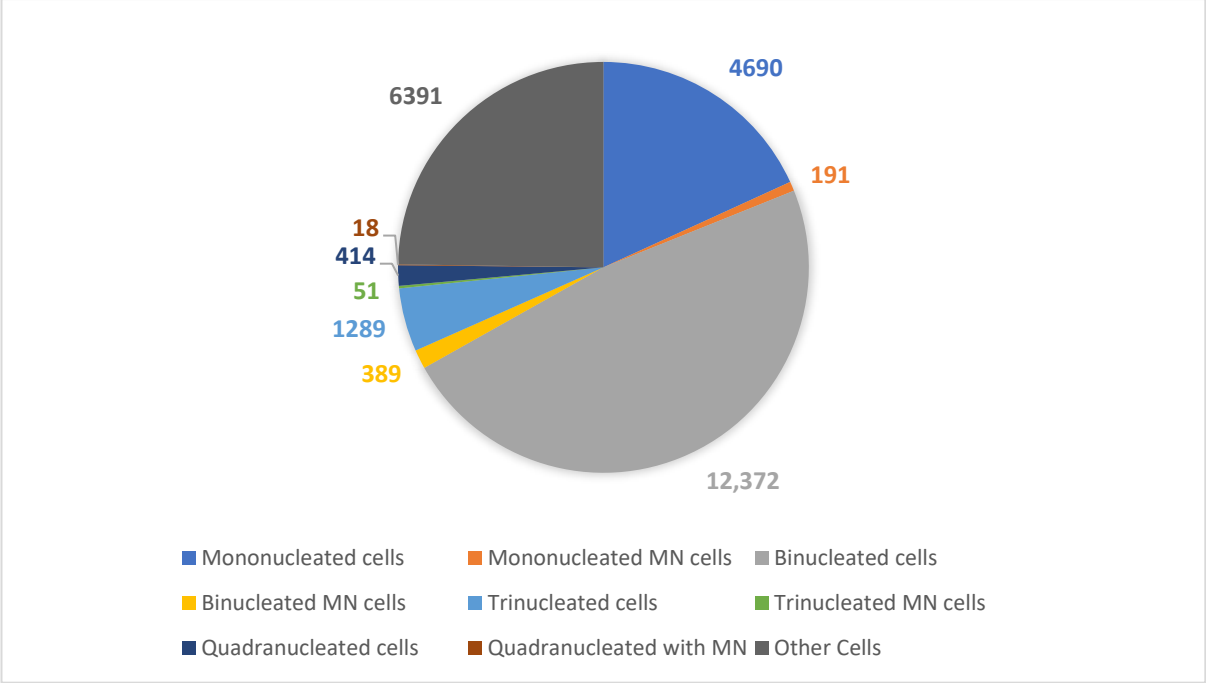


Figure. 12 Distribution of manually scored cellular images from GSK data set. Cellular images manually analysed using AdobeBridge® software based on cellular phenotype. Data represented as %'s in the Pie Chart, however, depicting quantities of cellular images. 25,805 cells analysed in total. This data set was not used in the creation of a neural network and could be helpful in future work as provides scope for a 3rd laboratory data-set to be used in tandem with the Cambridge and Cardiff ground truths.

Dose (µg/ml)	Scored as Others	Others (%)
0	29,900	99.67
0.8	29,817	99.39
1.2	29,801	99.34
1.6	29,814	99.08

Table. 3 Table showing a great proportion of cells were analysed as others when using the 3-channel neural network approach: Brightfield, Darkfield, DNA. N=3.

ii) Neural Network Tables

Network Short-hand	Network full name	Dataset Trained on	Dataset validated on	Binucleated Cell Accuracy	**Binucleated Cell MN Accuracy**	Mononucleated Cell accuracy	**Mononucleated Cell MN Accuracy**
Network **9**	TCambVCamb	Cambridge	Cambridge	94.1	59.5	92.3	72.4
Network **2**	TCardiffVCardiff	Cardiff	Cardiff	97.7	86.5	95.5	71.4
Network **5**	TCardVCamb	Cardiff	Cambridge	99	81.9	95.9	67.9
Network **6**	TCambVCard	Cambridge	Cardiff	77.5	15.3	95.4	63.2
Network **1**	TCardVCard	Cardiff	Cardiff	98.4	71.8	96.9	59.8
Network **3**	TCardVCamb	Cardiff	Cambridge	82.9	52.6	96.6	53.3
Network **8**	TCambVCamb	Cambridge	Cambridge	70.4	44.2	77.8	51.4
Network **7**	TCambVCard	Cambridge	Cardiff	93.8	69.4	93.5	51.2
Network **10**	TCambridge VCambridge	Cambridge	Cambridge	93.9	63.6	94.6	45.7
Network **4**	TCardVCamb	Cardiff	Cambridge	75.3	34.7	86.6	38.6

Figure Legend: Dark grey fill = Networ ks post Cardiff ground truth update

Light grey fill = Following initial 2 channel network accuracy

No Fill = original ground truth datasets used

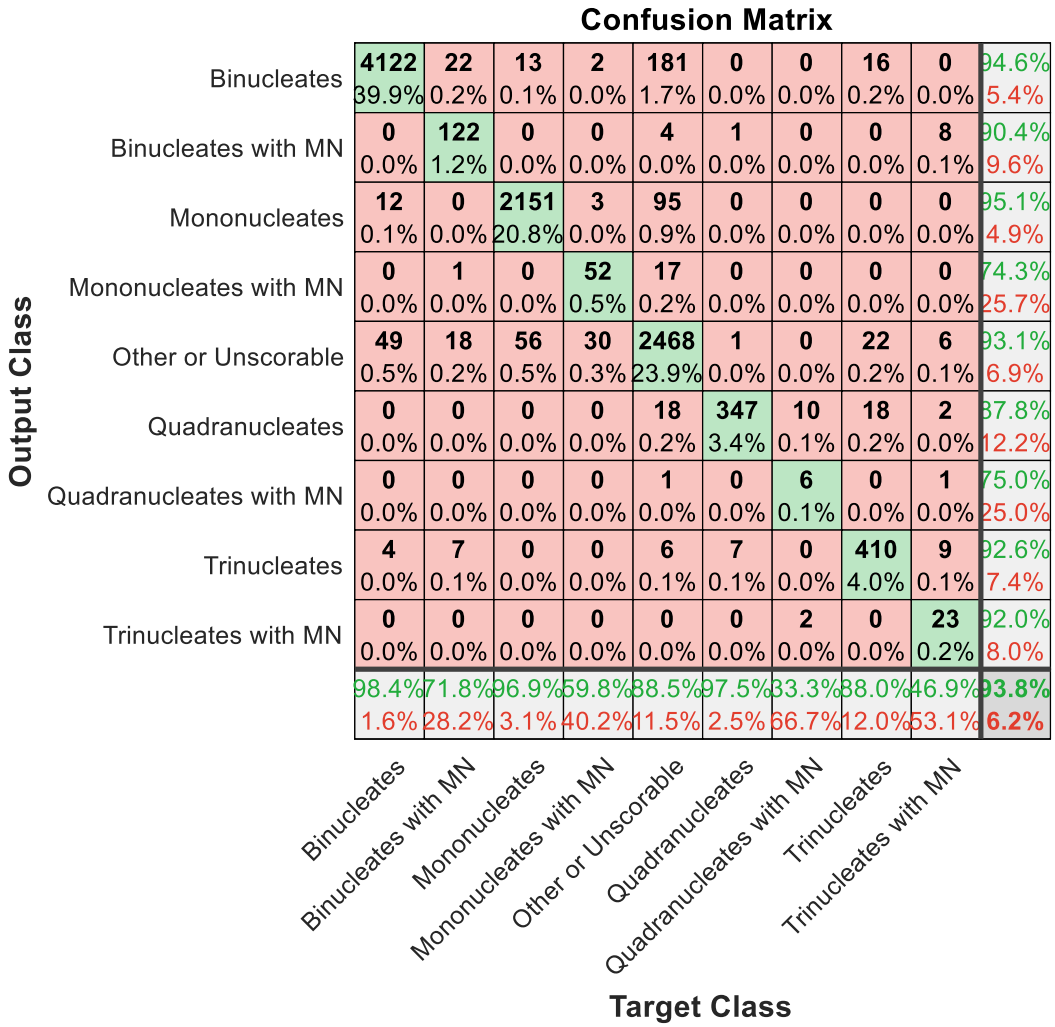
Table.4 Table showing neural network accuracies at different stages of the ground truth update. Binucleated Cell MN Accuracy' and 'Mononucleated Cell MN Accuracy' columns are in bold due to being the most important factors when deciding the 'best' network to use for a dose response. 'Binucleated Cell MN Accuracy' determines the network for use in a cyto-B dose dependent MN assay. 'Mononucleated Cell MN Accuracy' determines the network for use in the non cyto-B dose dependent MN assay. The table is sorted by Mononucleated MN cell accuracy, from most accurate to least.

Network Short-hand	Network full name	Dataset Trained on	Dataset validated on	Binucleated Cell Accuracy	**Binucleated Cell MN Accuracy**	Mononucleated Cell accuracy	**Mononucleated Cell MN Accuracy**
Network A	TCamb&CardVCamb	Cambridge &Cardiff	Cambridge	95.9	**75.3**	96.8	**77.3**
Network F	TCambVCamb	Cambridge	Cambridge	95.7	**59.5**	95.2	**75.6**
Network D	TCardVCamb	Cardiff	Cambridge	98.4	**89**	97.4	**72.6**
Network G	TCambVCambCard	Cambridge	Cambridge &Cardiff	95.3	**62.8**	98	**72.4**
Network I	TCardVCambCard	Cardiff	Cambridge &Cardiff	96.3	**72.8**	91.4	**68.7**
Network B	TCambCardVCard	Cambridge &Cardiff	Cardiff	96.4	**63.8**	96.3	**67.3**
Network E	TCamb_VCard	Cambridge	Cardiff	94.6	**56.2**	91.1	**62.2**
Network H	TCambCardVCambCard	Cambridge &Cardiff	Cambridge &Cardiff	96.9	**84.5**	95.4	**53.6**
Network C	TCard_VCard	Cardiff	Cardiff	98.6	**76.8**	97.7	**45.2**

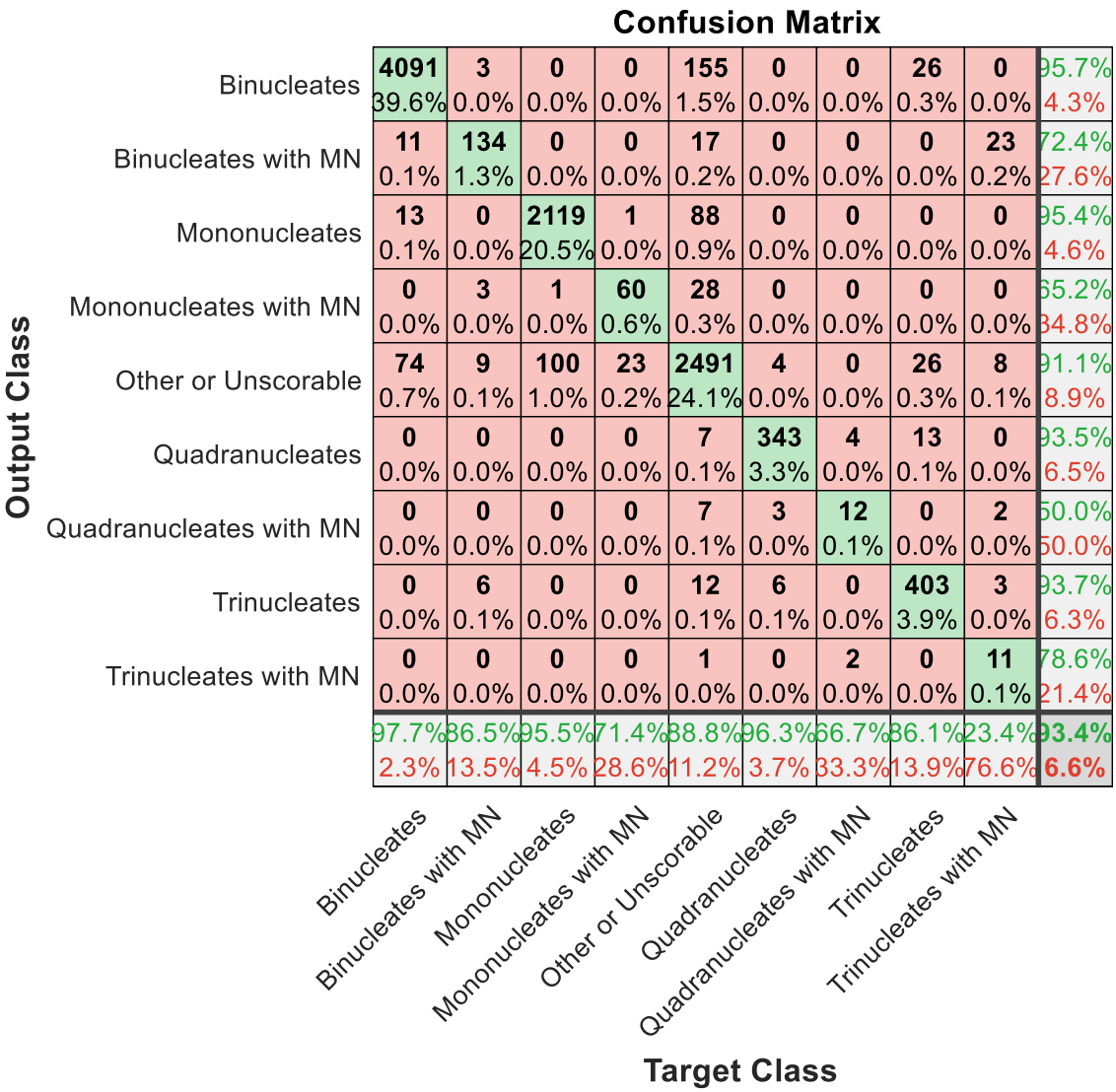
Figure Legend: Gold = Best accuracy, Silver = 2nd best accuracy, Bronze = 3rd best accuracy.

Table. 5 Table showing accuracy levels of the neural networks following training and validation on the latest updated Cardiff and Cambridge ground truths. ‘Binucleated Cell MN Accuracy’ and ‘Mononucleated Cell MN Accuracy’ columns are in bold due to being the most important factors when deciding the ‘best’ network to use for a dose response. ‘Binucleated Cell MN Accuracy’ determines the network for use in a cyto-B dose dependent MN assay. ‘Mononucleated Cell MN Accuracy’ determines the network for use in the non cyto-B dose dependent MN assay. The table is sorted by Mononucleated MN cell accuracy, from most accurate to least.

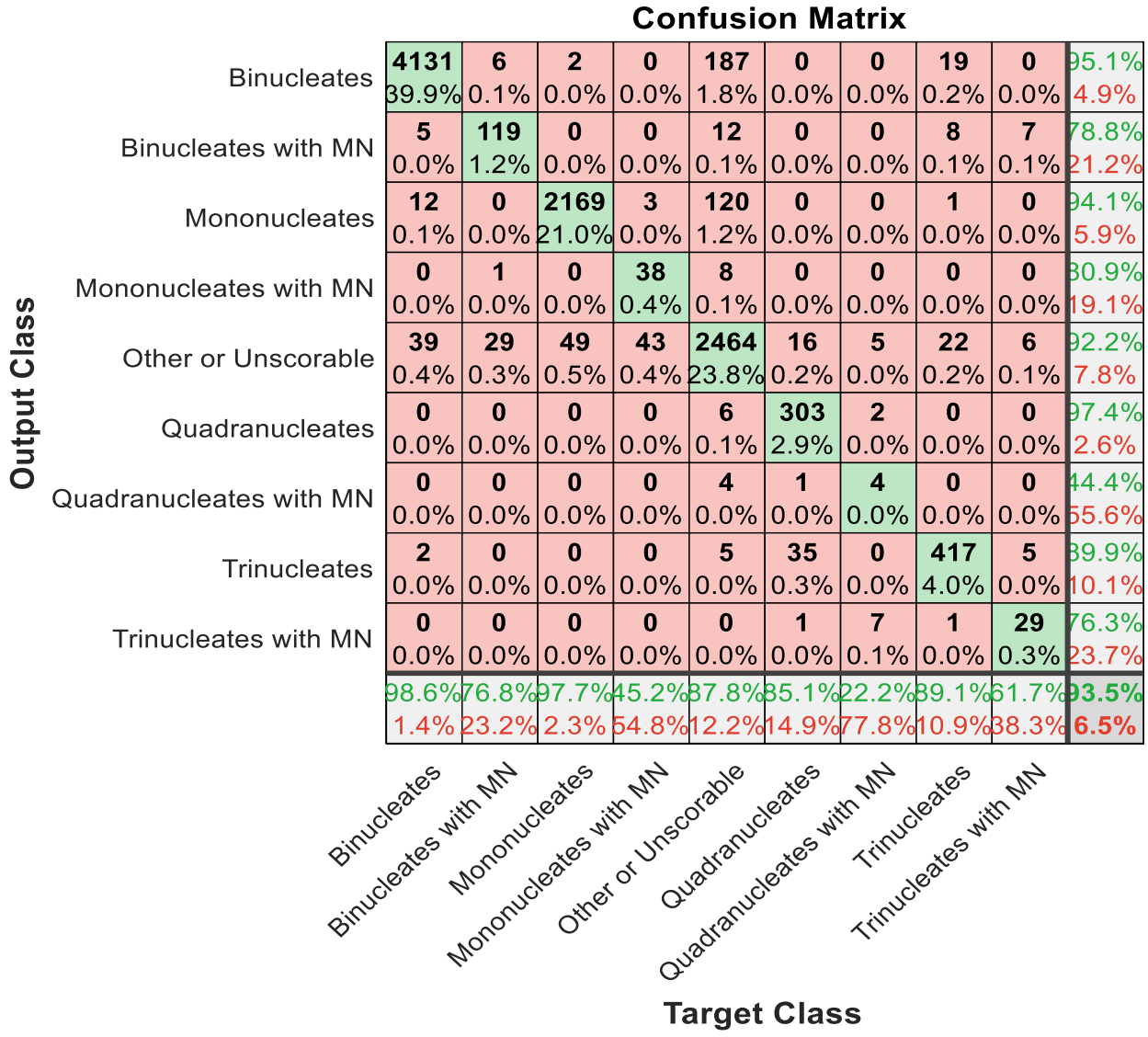
iii) Confusion matrices and Network training



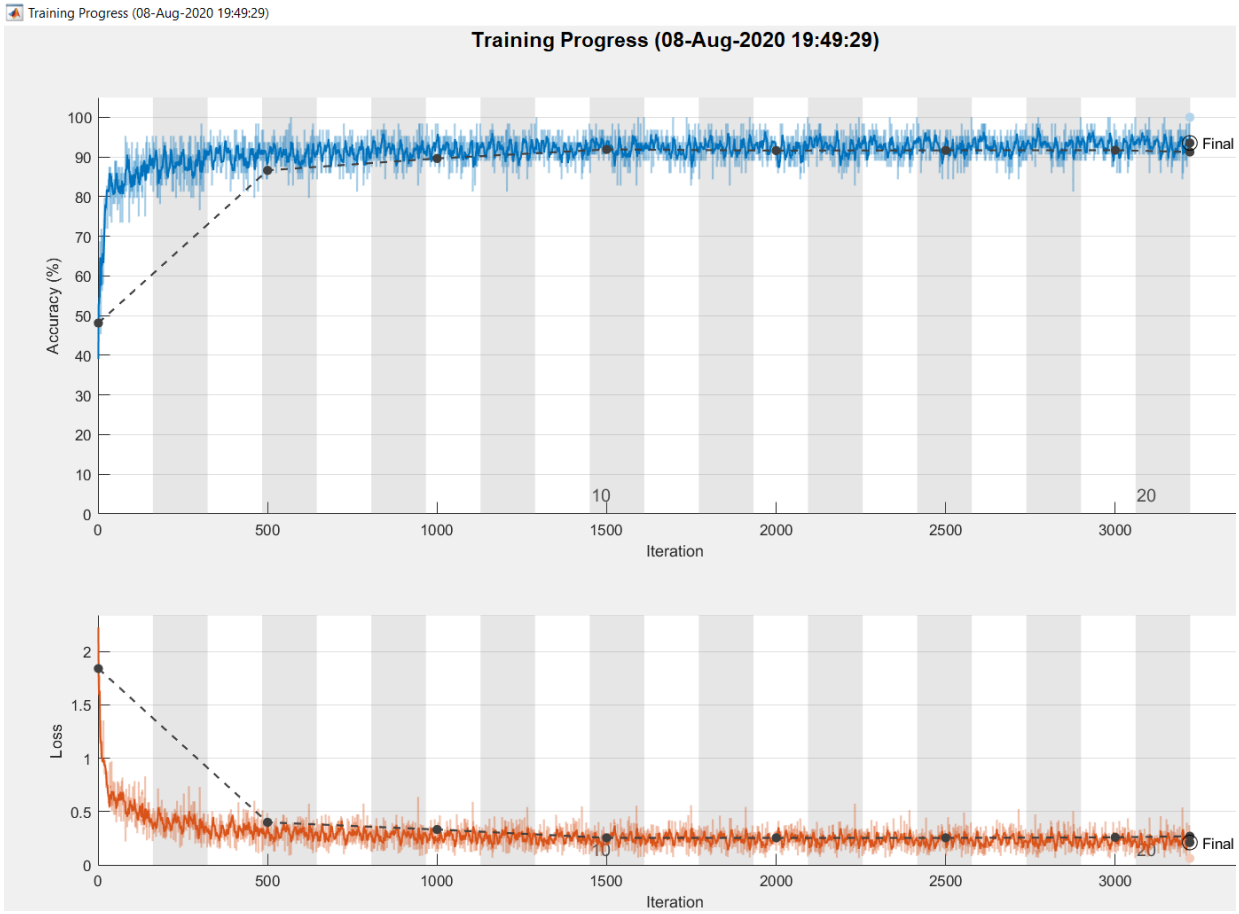
B


Confusion Matrix
Output Class
Target Class
Binucleates
Binucleates with MN
Mononucleates
Mononucleates with MN
Other or Unscorable
Quadranucleates
Quadranucleates with MN
Trinucleates
Trinucleates with MN
4091 39.6% | 3 0.0% | 0 0.0% | 0 0.0% | 155 1.5% | 0 0.0% | 0 0.0% | 26 0.3% | 0 0.0% | 95.7% 4.3%
11 0.1% | 134 1.3% | 0 0.0% | 0 0.0% | 17 0.2% | 0 0.0% | 0 0.0% | 0 0.0% | 23 0.2% | 72.4% 27.6%
13 0.1% | 0 0.0% | 2119 20.5% | 1 0.0% | 88 0.9% | 0 0.0% | 0 0.0% | 0 0.0% | 0 0.0% | 95.4% 4.6%
0 0.0% | 3 0.0% | 1 0.0% | 60 0.6% | 28 0.3% | 0 0.0% | 0 0.0% | 0 0.0% | 0 0.0% | 65.2% 34.8%
74 0.7% | 9 0.1% | 100 1.0% | 23 0.2% | 2491 24.1% | 4 0.0% | 0 0.0% | 26 0.3% | 8 0.1% | 91.1% 8.9%
0 0.0% | 0 0.0% | 0 0.0% | 0 0.0% | 7 0.1% | 343 3.3% | 4 0.0% | 13 0.1% | 0 0.0% | 93.5% 6.5%
0 0.0% | 0 0.0% | 0 0.0% | 0 0.0% | 7 0.1% | 3 0.0% | 12 0.1% | 0 0.0% | 2 0.0% | 50.0% 50.0%
0 0.0% | 6 0.1% | 0 0.0% | 0 0.0% | 12 0.1% | 6 0.1% | 0 0.0% | 403 3.9% | 3 0.0% | 93.7% 6.3%
0 0.0% | 0 0.0% | 0 0.0% | 0 0.0% | 1 0.0% | 0 0.0% | 2 0.0% | 0 0.0% | 11 0.1% | 78.6% 21.4%
97.7% 2.3% | 86.5% 13.5% | 95.5% 4.5% | 71.4% 28.6% | 88.8% 11.2% | 96.3% 3.7% | 66.7% 33.3% | 86.1% 13.9% | 23.4% 76.6% | 93.4% 6.6%

C I


Confusion Matrix
Output Class
Target Class
Binucleates
Binucleates with MN
Mononucleates
Mononucleates with MN
Other or Unscorable
Quadranucleates
Quadranucleates with MN
Trinucleates
Trinucleates with MN
4131 39.9% 6 0.1% 2 0.0% 0 0.0% 187 1.8% 0 0.0% 0 0.0% 19 0.2% 0 0.0% 95.1% 4.9%
5 0.0% 119 1.2% 0 0.0% 0 0.0% 12 0.1% 0 0.0% 0 0.0% 8 0.1% 7 0.1% 78.8% 21.2%
12 0.1% 0 0.0% 2169 21.0% 3 0.0% 120 1.2% 0 0.0% 0 0.0% 1 0.0% 0 0.0% 94.1% 5.9%
0 0.0% 1 0.0% 0 0.0% 38 0.4% 8 0.1% 0 0.0% 0 0.0% 0 0.0% 0 0.0% 80.9% 19.1%
39 0.4% 29 0.3% 49 0.5% 43 0.4% 2464 23.8% 16 0.2% 5 0.0% 22 0.2% 6 0.1% 92.2% 7.8%
0 0.0% 0 0.0% 0 0.0% 0 0.0% 6 0.1% 303 2.9% 2 0.0% 0 0.0% 0 0.0% 97.4% 2.6%
0 0.0% 0 0.0% 0 0.0% 0 0.0% 4 0.0% 1 0.0% 4 0.0% 0 0.0% 0 0.0% 44.4% 55.6%
2 0.0% 0 0.0% 0 0.0% 0 0.0% 5 0.0% 35 0.3% 0 0.0% 417 4.0% 5 0.0% 89.9% 10.1%
0 0.0% 0 0.0% 0 0.0% 0 0.0% 0 0.0% 1 0.0% 7 0.1% 1 0.0% 29 0.3% 76.3% 23.7%
98.6% 1.4% 76.8% 23.2% 97.7% 2.3% 45.2% 54.8% 87.8% 12.2% 85.1% 14.9% 22.2% 77.8% 89.1% 10.9% 61.7% 38.3% 93.5% 6.5%

II



III

Results	
Validation accuracy:	93.53%
Training finished:	Reached final iteration
Training Time	
Start time:	08-Aug-2020 19:49:29
Elapsed time:	42 min 44 sec
Training Cycle	
Epoch:	20 of 20
Iteration:	3220 of 3220
Iterations per epoch:	161
Maximum iterations:	3220
Validation	
Frequency:	500 iterations
Other Information	
Hardware resource:	Multiple GPUs
Learning rate schedule:	Piecewise
Learning rate:	0.001

Figure. 13 Confusion matrices produced post neural network creation using MatLab®. Training on the 'Cardiff' dataset and validation also on the 'Cardiff' data set. Overall accuracies shown as well as accuracies per individual subgroups.

A) Neural network formed after using the first updated ground truth and using a 3-channel approach of Brightfield, Fluorescence, Fluorescence, Network 1 produced this confusion matrix.

B) Neural network formed after using the first updated ground truth and the 2-channel approach of: Brightfield and Fluorescence. Network 2 produced this confusion matrix.

C) i) Neural network formed post Cardiff and Cambridge ground truth updates and the 2-channel approach of: Brightfield and fluorescence (1, 11). Network C produced this network
ii) Figure showing the training development of Network C up to 20 epochs. Accuracy and error rate are both shown.

iii) The results section after completing a network run, showing the accuracy rate of the network produced after completing 20 epoch cycles, other key statistics are also shown, such as iterations taken and time elapsed.

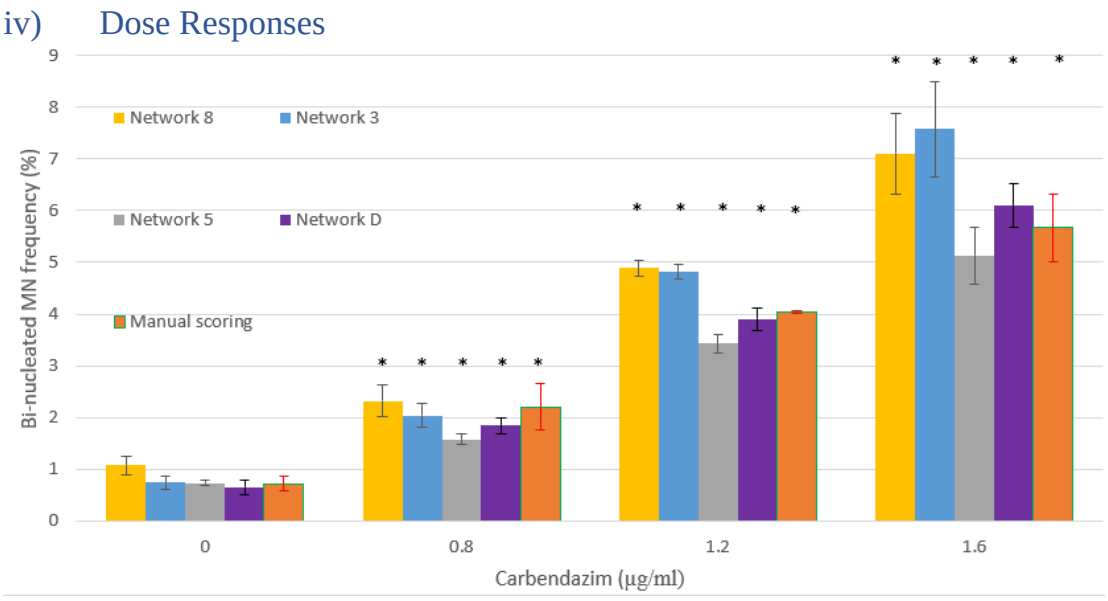


Figure. 14 Comparison of cyto-B MN dose response assay treating with Carbendazim and assessing using 4 triplicates from the neural network complex and one manually scored triplicate collected by the ImageStreamX-MkII®. Network 8, 3, 5 and D were used. N=3, mean =+/-StDev *Denotes a significant dose dependent increase ($P<0.05$)

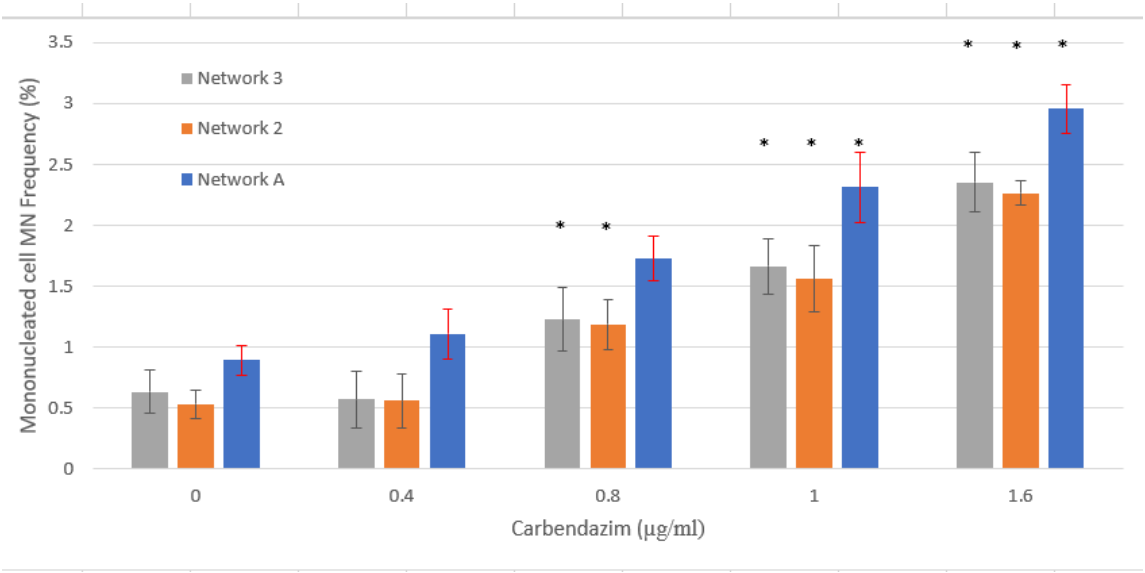


Figure. 15. Development of non cyto-b MN dose response assay treating with Carbendazim. The Newcastle data set was assessed for neural network accuracy development. Networks 2, 3 and A were used in analysis for comparison of the different ground truth stages. N=3, Mean = +/-- STDev. *Denotes a significant dose dependent increase (P<0.05)

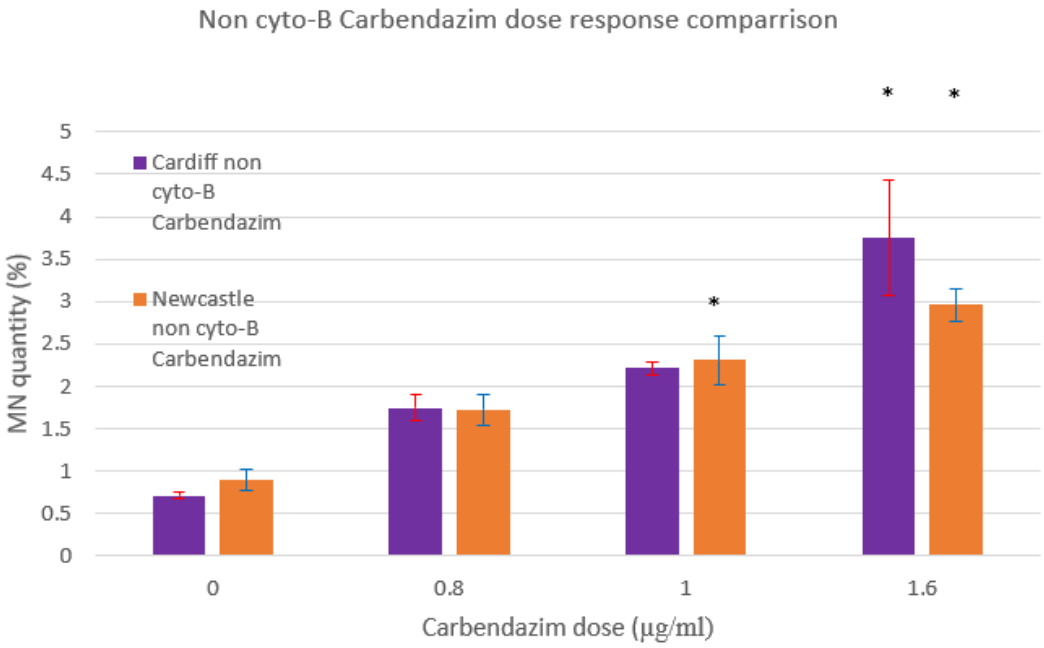


Figure. 16 Comparison of Cardiff and Newcastle non cyto-B Carbendazim dose responses using the automated neural network A. N=3[y], Mean = =+/- StError Increase seen from control in both laboratories at the top dose and similarity shown between datasets. *Denotes a significant dose dependent increase ($P<0.05$)

[y]N=3 for every dose excluding the Cardiff control and 1.0µg/ml doses.

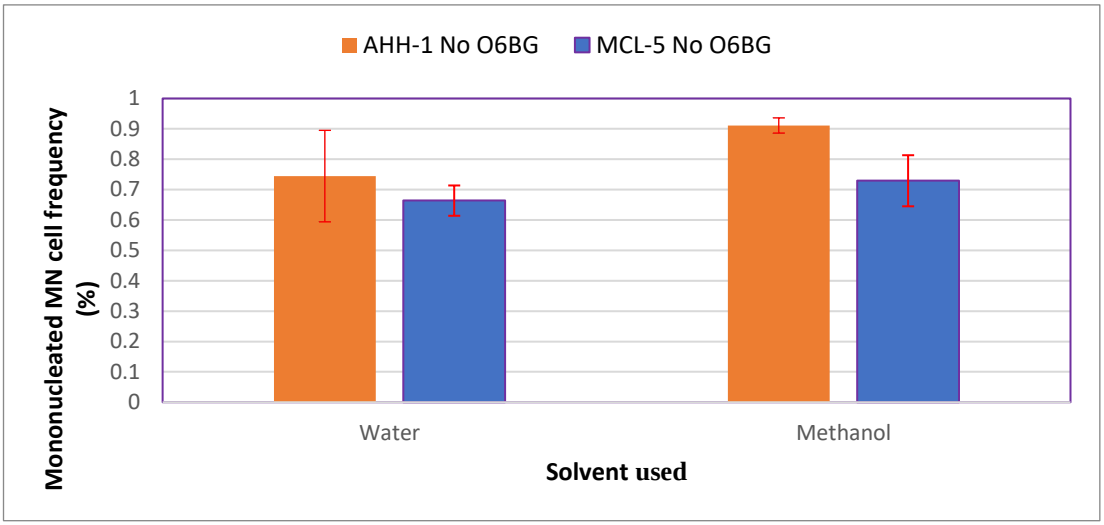


Fig. 17 Comparison of background MN rates in two different cell lines, MCL-5 and AHH-1 cells, using a TK6 derived ground truth neural network, to compare the ability of the ground truth to assess other similar cell lines. Neural Network A. N=3. Mean = +/- StDev

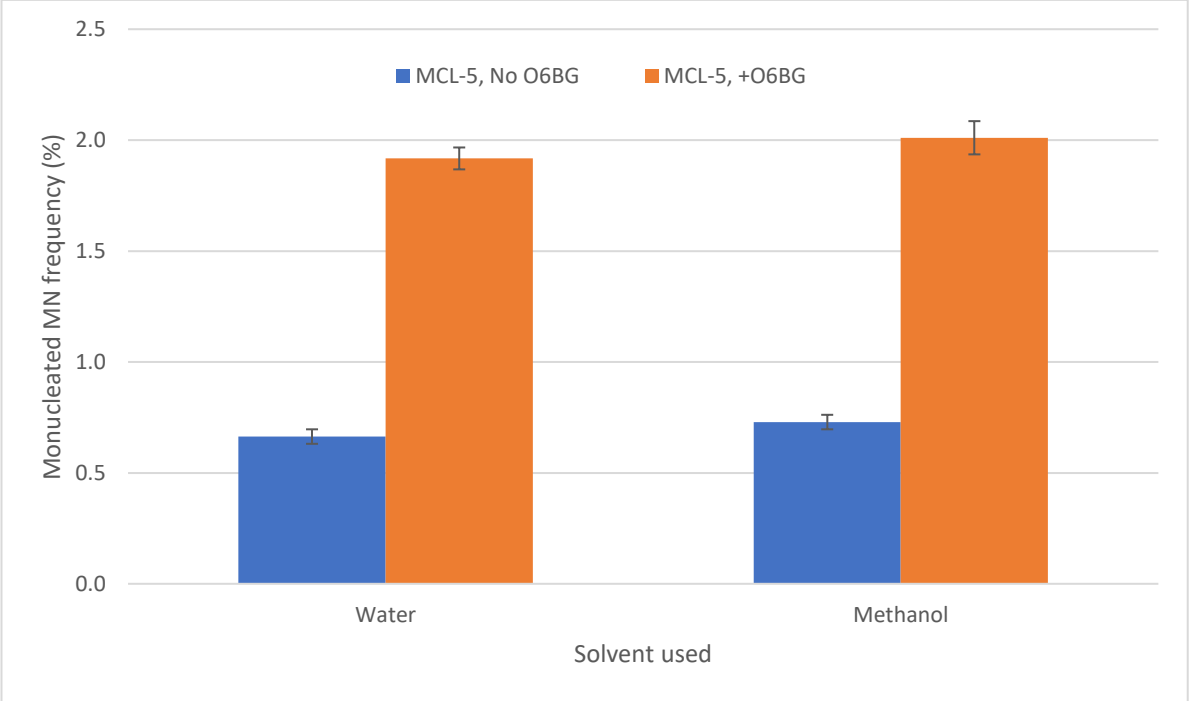


Figure. 18. Comparison of background MN levels in MCL-5 cells with and without O6BG in a solvent control sample. Assessment using the neural Network automation method to determine capability of using a TK-6 cell trained neural network on other cell types with and without O^6-BG. Analysed using Neural Network A. N=2 (Outliers omitted). Mean = +/- StError.

v) Flow of Work Undertaken

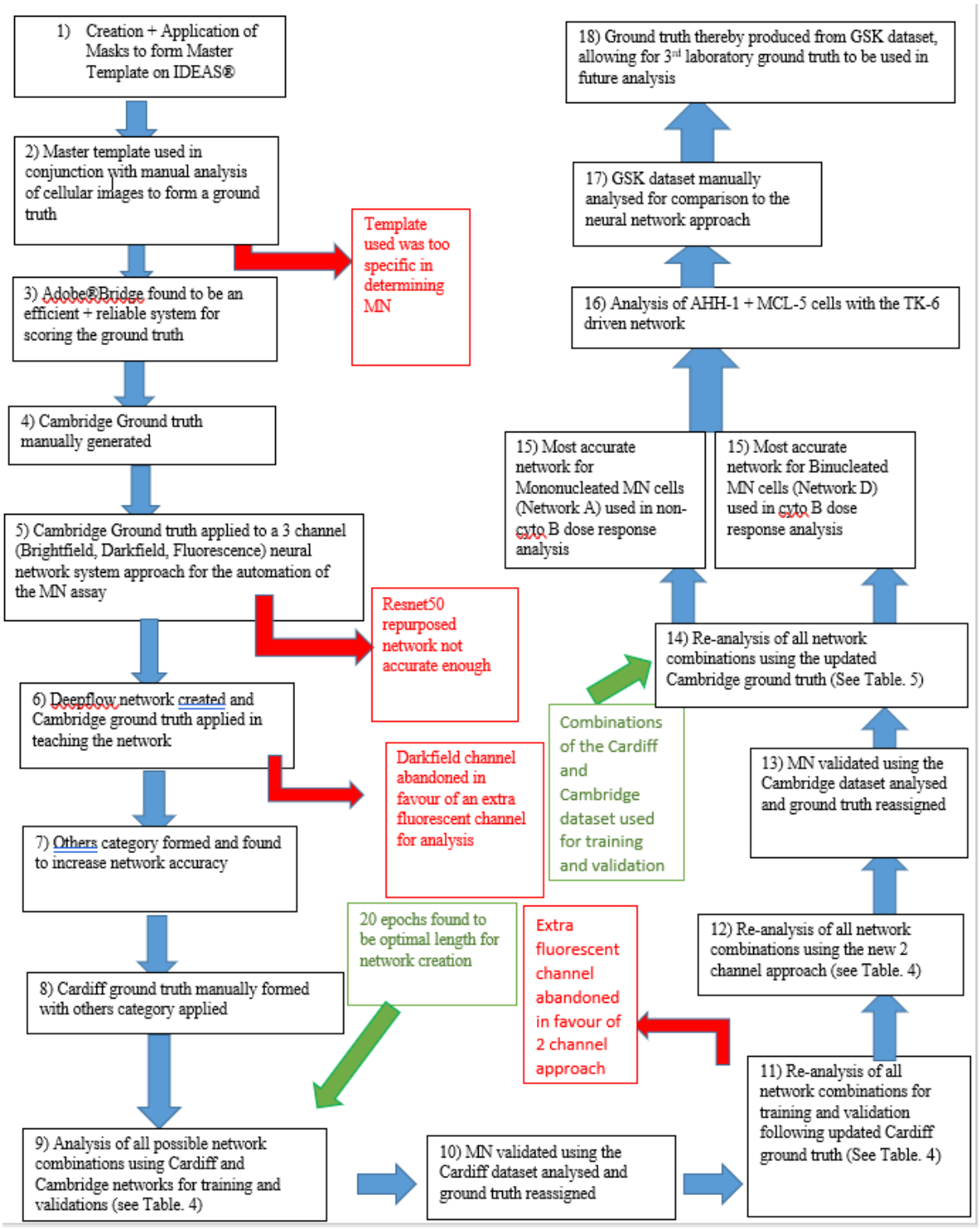


Figure Legend: Blue arrow denotes continuation onto the next step in development. Green arrow denotes a new step in the process added in. Red arrow denotes a step abandoned in order to improve accuracy levels.

Figure. 19. Flow-chart showcasing the development of the systems used to automate the MN assay, from the IDEAS automation attempts to the final 2-channel approach used in the dose response analysis.

4.0 Discussion

i) IDEAS® Template formation

The IDEAS® program is the recommended program working in tandem with the imaging flow cytometry equipment and the INSPIRE software which, linked to the imaging flow cytometer, allows for the conversion of the raw data (RIF) into an electronic format and provides initial gating analysis.

Due to the ease of access of the software, an initial attempt was carried out to fully automate the MN assay using a combination of masks and features within the IDEAS® program to allow the user to apply the template and determine the quantity of MN found. This hypothesis was further cemented by literature in the field, including Rodrigues' attempts to automate the MN assay using the ImageStreamX-MkII®flow cytometer (Rodrigues, 2018). However, Rodrigues made adjustments to the guideline outlined by Fenech in his guidelines for a MN with the MN size set to between 0.89-13.3μm² as opposed to the 0.40-11.1μm² range which would adhere to the guidelines of MN having areas between $1/256^{th}$- $1/9^{th}$ of the main nuclei (Fenech, 2000 and Rodrigues, 2018). This is acknowledged by Rodrigues, who cites artifact prevention as the reason for the alteration. By adding extensive templates and masks to determine healthy from alive cells (as shown in Fig. 3), it was our aim to eliminate extensive changes to the MN guidelines. Moreover, when comparing to Rodrigues' method, a background MN rate of 0.19% was observed by Rodrigues, which he once again acknowledges to be slightly lower than the historical range of 0.32%-1.38% (Lovell *et al.*, 2018 and Rodrigues, 2018). In previous research in our laboratory, a background MN range of around 1% was consistently found (Verma *et al*., 2018). Moreover, in my own analysis (in Fig. 20, a background MN% of no less than 0.53%

was determined. This is a threefold increase on the finding from Rodrigues and therefore reduces confidence in the significance of the dose response due to this far lower background MN figure).

By brainstorming the features which specify cellular phenotypes, it was possible to form the masks and features necessary to separate these into different classes (as shown in Fig. 3). (Fig. 4 shows the pathway which is used to determine a cells phenotype) and by using the graphs in IDEAS® to determine the cut off points, it is possible to view a select population and the cells considered to be in that specific population. In IDEAS® the user is able to then view each individual cells plot on a graph and therefore, the minima and maxima regions of a specific are can be extended or shortened if it can be visually analysed that a number of 'correct' cells are falling outside of the region (an advantage of imaging flow cytometry over conventional cytometry). This was used constantly to ensure that the correct features and thresholds were incorporated into the analysis. The brightfield and fluorescence channels were used in tandem as has been traditional in imaging flow cytometry MN analysis (Verma *et al*, 2018). The fluorescent channel allows artefacts to be distinguished so a false positive is obtained and the brightfield channel allows for the integrity of the cytoplasm to be checked so as to confirm if the cell is in a healthy state Verma *et al*, 2018).

The resulting 'master template' produced did accurately gate MN as was the aim; however, the gating was found to be too specific. Only 'perfect' MN were picked up by the strict guidelines the masks and features created in the template. The result was that a low level of non-MN were found to be produced as a result of the mask, as was the initial aim. However, by enforcing strict guidelines into the gating of the MN, not enough MN were gated due to not all MN having a 'normal' MN morphology. Some MN do not adhere to these specific guidelines and therefore are not gated and not deemed MN. This was reported by Verma *et* al, whereby aneugens and clastogens can have a different effect on the resulting MN formed (Verma *et al.,* 2018). Moreover, there was a need to update the masks based on which channels the different labs were using (this differs based on which nuclear stain was used due to different frequencies exciting different stains). This proved to be time-consuming and reduced the effectiveness of automating the system. However, it would be

possible to create variations of the master template for differing nuclear stains for long term efficiency.

This led to an under-estimation of the MN count which in turn affects the use of such a system in a dose response system. One would expect a dose response to be shown, but the levels would not adhere to historical data and therefore the comparison potential is limited. This limits the use in applying the technique to unknown doses and severely impacts the use of this approach going forward. Therefore, these limitations added more difficulty to the automation attempt on the IDEAS program and can be attributed to part of the reason the automation attempt on IDEAS® did not work in the manner required to automate this assay truly whilst maintaining accuracy levels consistent with the gold standard of manual microscopy (Fenech, 2000).

ii) Ground truth creation

Due to these limitations, the new approach was taken outside of IDEAS® and the idea of deep learning came to the fore, whereby the data collected in the IDEAS® approach could be adapted to form a 'ground truth' which would teach the network what a MN should look like by, as opposed to the 'machine learning based' approach carried out in the IDEAS® program which carried too much specificity.

By forming the ground truth by manual scoring on IDEAS®, it was possible to generate a ground truth to teach the neural network which was as accurate as possible. The accuracy was vital, as any incorrect images distorted the accuracy of the neural network. Adobe®Bridge, a tool commonly used in photography analysis, proved to be a useful tool in the creation of the ground truth for the neural network. It allowed for ease in scoring and allowed to differentiate MN scored cells from non-MN containing cells of the same nucleic quantity. However, despite the network scoring images in an initial 3 channel manner, the images were manually assessed using a singular channel, whereby the fluorescent image was found to be the most accurate channel for the individual scoring of cellular images. By only being able to

look at the fluorescent images in the ground truth creation process, it was possible for cellular categories to be more difficult to be differentiate, especially with the large quantity of cellular types originally used in Table.1 and Fig. 5.

It is clear, by looking at Fig. 5 that 1000 images are not reached for categories other than: Apoptotic (1231), Binucleated (2343), Overlapping binucleated (2055) and Mononucleated (4419). This shows that out of the 17 total categories used in Fig. 5a, only 4 contained images over the historical threshold, less than 25% of the total categories. This can be attributed to the great difficulty in generating 1000 images of the rarer phenotypes such as MN or Nucleoplasmic bridges to an even greater extent. The background levels of MN in literature are 0.32-1.38% (Lovell *et al.*, 2018). This would require 100,000 control dosed cells to be manually scored to give around 1000 MN which is a laborious and time-consuming approach to solve an already laborious and time-consuming method. The highest dosed cells could also be analysed; however, this would require carrying out RPD studies and carrying out dilutions in order to dose the cells and increases the labour-intensive nature of the automation process. This would increase the experiment time and would be further increased due to the vast amounts of categories used and number of categories containing sub optimal levels of images.

iii) Alteration of categories, 'Others' added

It was decided to introduce a category called 'others' which would replace the dead and unscorable events. By adding this category, it was possible to condense the images into more specific categories, by removing areas of ambiguity, thereby allowing for less confusion to take place between categories due to a variety of reasons. In the first instance, as there are less categories, there are less opportunities for the cellular images to be placed into the incorrect category, as we reduced the categories from 17 to 9 as seen in Fig. 5. Moreover, the categories which were merged in Fig. 5b and onwards, were similar morphologically in Fig. 5a. This can be shown by the bi/tri/quadranucleated overlapping cells were grouped together with their nucleic counterparts of the same quantity. This allowed for less confusion by placing these in the same category. It must be noted, that due to the images being

manually assessed in the fluorescent channel, the cytoplasm could not be visualised in the same way and therefore more difficulty was seen in differentiating apoptotic cells from necrotic in Fig. 5a. Adding the 'others' category also ensured that 4/9 categories now contained more than 1000 cells in Fig. 5b. This is an improvement from 23.53% to 44.44% in categories containing more than 1000 cells, almost doubling. Lastly, categories such as nuclear buds and MN are very similar morphologically, that having these in two different categories provides the neural network with an impossible task in differentiating the two, which would then result in the confusion of the network when attempting to distinguish between the two categories. The new grouping of categories therefore allowed the network to better predict the category a cell should fall in to and this categorical approach was chosen going forward due to the increased confidence in the number of images in each category and the stark differences between the different categories, which was not the case previously.

iv) Optimal epoch count

The epoch count is one of the most important variables to consider when dealing with neural network analysis. By increasing or decreasing the epoch number, the network can be allowed longer to train and learn from the images. However, a balance must be met, as the error rate can also increase if left to run for an extended period, which can harm accuracy levels.

This was shown in Fig. 12a) which was run for 30 epochs and produced only a 74.1% overall accuracy. More importantly than the overall accuracy, are the accuracy levels for the mono and binucleated cells with MN categories. A 44.4% accuracy is shown for binucleated cells with MN and a 51.4% accuracy for mononucleated cells with MN. When this is compared to Fig. 5b), a lower overall accuracy is seen, 74.1%<83.5%, lower binucleated cell accuracy, 70.4%<94.1% and lower binucleated MN cell accuracy, 44.2%<59.5%. The binucleated and binucleated MN percentage accuracy is important in assessing neural network accuracy due to the analysis of this phenotype when undertaking the cytochalasin-B MN assay.

Given the rarity of MN cells when carrying out a dose response, it is vital that the binucleated MN cell accuracy is as high as possible, a more important factor than overall accuracy and binucleated cell accuracy. Given the importance of this category, a 15.3% increase in accuracy in the 20-epoch sample is a significant value but could be improved further. When assessing mononucleated cell accuracy, there is an increase from 77.8% to 92.3% and an increase from 51.4% to 72.4%. An increase of 21% is shown when comparing the 30-epoch network 8 to the 20-epoch network 9, whereby the two networks have both been trained and validated on the same Cambridge dataset, using identical ground truth populations. Identical channels have been used in both cases: brightfield, fluorescence, fluorescence. Therefore, the increase in accuracy between network 8 and network 9 can be solely attributed to the reduction in epoch frequency from 30 to 20. Given this increase in accuracy shown from Fig. 5a to Fig. 5b, 20 epochs were chosen going forward to be the optimal epoch frequency. Fig. 5dii) shows the accuracy and loss rates per epoch/iteration level. A reduction is not shown in the accuracy level up to the 20-epoch level or an increase in loss ratio. When the accuracy level starts to decrease and the loss rate increase, it is indicative of over training the network, whereby the network over focuses on a specific, non-important, area than what it was intended for. Therefore, future networks were trained to 20 epochs for optimal accuracy and loss levels.

v) Cardiff ground truth generation

The Cardiff ground truth was generated in much the same way as the Cambridge ground truth. However, since the others category was already found to be

advantageous previously, the cellular category was adopted immediately when creating the ground truth for this dataset. When viewing the cellular splits of the Cardiff ground truth in Fig. 19a, 3/9 categories, 33.33%, contained more than 1000 cellular images. All 3 of these categories totalled over 2000 cellular images in fact. Interestingly, less trinucleated cells were generated in the Cardiff ground truth when compared to the Cambridge ground truth, 1287 compared to 463 when comparing Fig. 5b to Fig. 19a. The Cardiff ground truth was once again drawn from the Adobe®Bridge software package, with the star rating system once again being used successfully. An important note is that the Cardiff laboratory uses Draq5 in the nuclear staining of cells, whereas Cambridge use Hoescht 33342. Indeed, it is impressive that these two nuclear stains can both be used in tandem on two different datasets and generate high levels of accuracy.

As there were now two ground truth data sets in the shape of Cardiff and Cambridge, it was now possible to analyse combinations via training on one set and validating on the next, this will be explained in greater detail. But one hypothesised that the difference in nuclear stains used may be a leading factor in some of the variation shown depending on which dataset combination is used for training and for validation.

vi) Darkfield channel abandoned

Following the creation of ground truth from two different laboratories, the network was created in order to determine the most accurate combination between training and validations sets. However, before this could be undertaken, the 3-channel approach of: brightfield, darkfield and fluorescence was revised.

Each individual image the neural network uses, is split into 3 images, containing a: brightfield, fluorescent and darkfield image. It is by using the training data, consisting off all 3 images in 1 (although only the fluorescent channel was assessed in manually scoring the ground truth into categories, the extra layers were used to

add information and therefore increase accuracy). The others rate was greatly reduced when the darkfield channel was abandoned in favour of another fluorescent layer. This can be shown when comparing the others percentage in the final GSK dose responses (35.49% of cells as others), a great reduction from the over 99% shown in Table. 2.

`The decision to add an extra fluorescence layer instead of an extra brightfield layer was made after assessing the accuracy levels produced by the network using the: brightfield, brightfield, fluorescence combination.

The added fluorescence accuracy is not surprising for a few different reasons. The images which were assessed for the ground truth creation were fluorescent images and thus adding more weight to these would be thought likely to result in an increased level of accuracy. Moreover, adding an extra fluorescent channel makes the 3 layered image, 2/3rd /66.67% fluorescent (Verma *et al*., 2018, Haxhiraj *et al*., 2018). This is similar to the weighting of the fluorescent channel in IDEAS® when creating a composite image, created using the brightfield channel and fluorescent channel, whereby the fluorescent channel is set to around 70% intensity in order to be able to distinguish nuclear material from artefact but still maintain the cytoplasmic integrity that the brightfield channel offers (Verma *et al*., 2018).

Therefore, similarities can be made in the creation of a doubly weighted fluorescent channel using the 3-layered neural network image approach. By abandoning the darkfield approach and forming tiff images in the new brightfield, fluorescent, fluorescent manner, the neural network was used to train and validate on the ground truth images, in preparation for dose response prediction.

vii) Network Training and Validation Analysis

The original Cardiff and Cambridge ground truth datasets were used to originally create neural networks. The two most accurate networks produced using the initial ground truth classifications were:

-Network 3 (See Fig. 20), training using the Cardiff data set and validating using the Cambridge dataset.

-Network 8 (See Fig. 22), training and validating using the Cambridge dataset.

Network 3 produced the highest overall accuracy: 74.5% compared to 74.1%, the higher binucleated cell accuracy: 82.9% compared to 70.4%, the higher binucleated MN cell accuracy: 52.6% compared to 44.2%, the higher mononucleated cell accuracy: 96.6% compared to 77.8% and the higher mononucleated MN cell accuracy: 53.3% compared to 51.4% (Table.4, Fig. 20, Fig. 22).

Given the unusually small difference, especially in the mononucleated MN cell accuracy, it was determined that the Cardiff ground truth may have some outliers present in the ground truth creation, which were distorting the network accuracy. This was decided as the probable cause of the lower than expected accuracy shown, over issues such as epoch number and differences in nuclear intensities. For epoch length, there was a clear reduction in epoch count at 30 epochs was shown in comparison to 20 epochs, as can be shown by the in Table.4, when viewing the comparisons between Network 8 and 9. For differences in nuclear intensities, by comparing Network 3 and Network 9 (both validated on the Cambridge but trained on different data sets), a higher differential was shown in the more generic categories, such as binucleated cell accuracy (12.5% (Table. 4)) and mononucleated cell accuracy (18.8% (Fig. 20, Fig. 22)) when compared to the MN containing images (8.4% and 1.9% (Table.4) most likely caused due to the misplacing of MN cellular images having a greater effect on the MN category due to the rarity of this phenotype and thus each individual image carrying greater weight in relation to the category. Thus, to ensure the ground truth integrity was as high as possible, analysis was undertaken of all possible MN cellular images and ensuring their presence in the correct category.

a) Cardiff Validated MN analysis

To further improve the accuracy of the neural networks, the ground truth population from the Cardiff dataset was analysed and any outlier images were identified and moved into the correct category. This was carried out in MatLab® by using the following script in Matlab®:

'deploy_DeepFlow_3_channel_new_pheno_v1.m'

This is the script used in producing a neural network, such as those shown in Table.4 and Table.5. The script contained the specific line of code:

Idx=find(imdsValidation.Labels=='Mononucleates with MN' & YPred=='Mononucleates with MN')

This line of code showed images which had been, in this particular case, placed into the 'Mononucleates with MN' class for training, but had also been validated by the network as 'Mononucleates with MN'.

This line of code can therefore be used to check the cellular images have been placed into the correct categories. The images were produced in windows containing 36 images each, with a number assigned to each individual cellular image from 1-x (the last cellular image applying to that category). The numbers relating to the individual cellular images were then shown in the command window of MatLab®, where the numbers 1-x corresponded to the full-length tiff number for that specific image. When analysing this data, cellular images which were not deemed to be of the

cellular phenotype to which they were originally assigned, were updated and placed into the correct category. By then applying the results obtained, the cellular images in the incorrect file could be dragged and dropped' into the correct folder. The example shown here was 'Mononucleates with MN', but all the cellular categories were applied, and the images manually assessed for any discrepancies.

By applying this method of quality control on the ever important ground truth class, the ground truth was refined and the updated ground truth used to train and validate the Cardiff and Cambridge ground truths once again, resulting in more accurate networks being produced. The updated ground truth can be shown in Fig. 9.

b) Re-analysis of Network combinations

By having a Cardiff and Cambridge data set ground truth available, it was possible to assess network accuracy in 4 iterations: 1) Training with Cardiff, validating with Cardiff, Network 1 (Fig. 13), 2) training with Cardiff, validating with Cambridge, Network 4 (Fig. 20), training with Cambridge, validating with Cardiff, Network 6 (Fig. 21) and training with Cambridge, validating with Cambridge, Network 9 (Fig. 22).

Each network combination was assessed, and accuracies compared to find the network most suitable for use in the automated MN dose response. Network 1 (Fig. 13) showed the greatest overall accuracy at 93.8%, in contrary to Network 4 (Fig. 20) which only showed a 73.7% accuracy, a great 20.1% difference. Network 1 (Fig. 13) also showed the highest binucleated cell accuracy, with a respectable 98.4%, on the

contrary to Network 4 (Fig. 20) which showed only 75.3%, a large 23.1% difference. It was also shown that Network 1 (Fig. 13) had the highest binucleated MN cell accuracy at 71.8%, which was 56.5% greater than Network 6 (Fig. 21) which only had a 15.6% accuracy level. Thus, if one were to carry out the cytochalasin-B MN assay, Network 1 is clearly the network of choice. However, despite Network 1 being the most accurate and showing great specificity in identifying binucleated cells, however, with only an accuracy of 71.8% (Table. 4) for binucleated MN cells, the levels were not quite high enough to maintain the integrity of the MN assay and as such improvements to the network were made, which will be expanded on shortly.

When assessing the mononucleated cell accuracy for potential non-cytochalasin-B use, Network 1 (Fig. 13) was again found to have the highest accuracy level for mononucleated cells at 96.9%. However, unlike with the binucleated cells, even the worst network for predicting mononucleated cells Network 4 (Fig. 20), still maintained an accuracy of 86.6%, a difference of just 10.3%, far smaller than the 23.1% difference observed in binucleated cell accuracies (Fig. 9, Fig. 20). When comparing mononucleated MN accuracies, Network 9 (Fig. 22) shows the highest accuracy with 72.4%, whilst Network 4 (Fig. 20) shows the worst at 38.6%, a difference of 33.8%. Moreover, just as with the binucleated and binucleated MN cells, a large difference is seen between the differences in accuracies with and without MN present. The ‘Mononucleated MN’ highest accuracy level of 72.4% (Fig. 22) is still not optimal accuracy levels for use in the MN assay.

It is also somewhat unsurprising that the best accuracy levels for MN at this initial stage, were both obtained via training and validating on the same datasets, due to the same nuclear stains being used in both. It was interesting to note how the Cambridge network gave the highest accuracy levels for mononucleated MN cells (Fig. 22), whereas the Cardiff dataset provided a greater accuracy level to binucleated MN cells (Fig. 13). After this analysis, I was interested on ways to improve the network accuracy, focusing predominately on improving MN accuracy levels, due to the cellular accuracies levels being far greater and at above 90% for mononucleated and binucleated cells (Fig. 13).

The difference to the original dataset can be viewed by the differences in accuracies shown when compared to the updated ground truth dataset. The highest overall

accuracy in the updated network is 93.8%, an increase of 19.3% when comparing the updated Network 1 to the older Network 3 (Table.4). Indeed, a 15.5% increase is viewed in the binucleated cell category, a 19.2% increase in binucleated MN cell accuracy. Moreover, this trend is reduced in the mononucleated cell accuracy using the same networks, where only a 0.3% increase is shown and only a 0.5% accuracy in mononucleated MN cell accuracy. Therefore, an improvement was made by updating the ground truth, especially in the binucleated cell category. This improvement suggests that the original ground truth did contain some outlier images in the incorrect category which reduced the accuracy levels.

However, a cause for concern in the analysis of the neural network accuracies was the sub-par performance shown when cross-network validation was carried out. The different intensities of the lasers and slightly different conditions across different laboratories can provide an explanation for lower accuracies, despite such attempts to reduce this factor by normalisation. Network 4 which is trained on Cardiff and validated on Cambridge and Network 6 which is vice versa, showed worryingly low levels of accuracy in the binucleated cell MN category (Table.4). Accuracy levels of 34.7% were shown by Network 4, with Network 6 showing 15.3% (Table.4). This accuracy level is a further reduction of 17.9% lower than the best network using the original ground truth populations (Table.4). This is problematic for the technique, since cross validation is a sign of reproducibility between laboratories and would need to be improved going forward if this technique is to be adapted on a larger scale.

Cross laboratory validation allows for a single network to be applied to a host of laboratories, forgoing the need to create a ground truth for each new laboratory. As well as showing that cross laboratory reproducibility could be obtained using differing nuclear stains also. Since the Cardiff ground truth was updated, the fault must fall with the Cambridge ground truth or a fault associated with the images format, with the 3-channel approach being reviewed.

viii) 2 channel approach

Following the inconsistencies shown in the 3-channel approach, where the cross-validation training networks were not shown to be as accurate as required, a 2-channel approach was taken going forward. This two-channel approach consisted of a fluorescent and brightfield channel each being used in equal proportions, with the extra fluorescent channel being sacrificed.

In order for this approach to be carried out, the '**two_channel_tiff_reader'** was employed in order to create a new batch of tiff images in 2 layers, as opposed to the 3 layers used previously. This was carried out in much the same manner as the initial 3 channel tiffs, a somewhat laborious process. Care was taken to ensure that the correct channels are used, due to different channels corresponding to the brightfield and fluorescent channels used in the Cardiff and Cambridge networks.

Once the tiffs had been created for the 2-channel approach, training and validation was undertaken using all permutations of training and validation, including cross-laboratory valuation and the results evaluated.

a) Re-analysis of all network combinations using 2 channels

The ‘DeepFlow’ neural network was again used and the accuracies evaluated. Network 5 had the highest overall accuracy at 93.7%, highest Binucleated cell accuracy at 99% and highest Mononucleated cell accuracy at 95.9%(Fig. 20) (Table. 4). Network 2 showed the highest accuracy level for binucleated MN cells at 86.5% and Mononucleated MN cell accuracy at 71.4% (Fig. 13).

It is important to note, that Network 5 had the greatest accuracy shown in 3/5 of the main categories, despite being a cross-validation network, which shows the great increase the 2-channel approach had on accuracy levels. The success of the cross-validation approach widens the scope for potential wider scale use of this approach. When compared to the previous 3-channel approach, using the same training and validations datasets (training on Cardiff and validating on Cambridge), a 20% increase was shown in overall accuracy, a 23.7% increase in binucleated cell accuracy, a 66.6% increase in Binucleated MN cell accuracy, a 9.3% increase in Mononucleated cell accuracy and a 29.3% increase in mononucleated MN cell accuracy (Fig. 10) (Table.4). Therefore, a substantial increase was shown in both mononucleated and binucleated cell accuracy, with a focus on the binucleated cell increase.

Altogether, the 2-channel approach increased the accuracy of the rarer phenotypes, with a 14.7% increase shown on the previous best binucleated MN cell accuracy (Network 2 compared to Network 1, see Table.4). Some of the other major categories did take a slight decline (when comparing the most accurate network used in each system), overall accuracy decreasing by 0.1%, mononucleated cell accuracy decreasing by 1% (Network 5 compared to Network 1, see Table.4). However, these minute detractions are more than accounted for by the great increase in the binucleated MN cell category.

However, the mononucleated MN cell accuracy was not shown to improve in the 2-channel approach, a decrease of 1% shown when compared to the previous network creations (Network 2 compared to Network 9, see Table.4). Interestingly, the Network 2 accuracy of 72.4% was the highest mononucleated MN cell accuracy shown on datasets validated on the Cardiff network, a testament to the increasing

accuracy of the 2-channel approach and the updated ground truth. However, throughout the previous networks, it had been the dataset validated on the Cambridge network which has shown the highest mononucleated accuracies. This brought up the idea of updating the Cambridge ground truth also in much the same way the Cardiff ground truth had been updated. This was supported again due to the ranking of accuracies in this 2-channel approach, whereby the networks were ranked by totalling the accuracies of mono/binucleated cells with and without MN and showed:

Network 2>Network 5>Network 7>Network 10.

It is unsurprising that Network 2 is trained and validated on the Cardiff dataset, Network 5 is trained using the Cardiff dataset and Network 7 is validated using the Cardiff dataset (Table.4). This leaves Network 10 at the bottom, the only network not containing any Cardiff influence (Table.4). Therefore, I concluded that there was a need to update the Cambridge data set.

b) Cambridge validated MN evaluated

The Cambridge ground truth was evaluated in the same way as the Cardiff data set, with the MN assessed using the following script in MatLab®:

Idx=find(imdsValidation.Labels==’Mononucleates with MN’ &YPred==’Mononucleates with MN’).

By assessing the MN produced, it was possible to re-assign cells into the correct category if they were previously placed incorrectly. By re-assessing the Cambridge ground truth, both ground truths had been evaluated and updated, ensuring greater accuracy levels. This process was even more important due to the smaller total number of MN cells in the ground truth due to the rarity of the MN phenotype, 0.32-1.38% (Lovell *et al., 2018).*

When assessing the MN found in the original 'others included' ground truth, 150 Mononucleated MN cells were included and 154 Binucleated MN cells (See Fig. 9b)). Post-updated ground truth, these numbers were reduced to 127 Mononucleated MN cells and 121 Binucleated MN cells (See Fig. 9c)). The Others category was increased on the other hand, from 1257 originally to 1372, suggesting that some MN cells may have been too ambiguous in morphology and thus confused the network and thus produced a lower accuracy (See Fig. 9b) and Fig. 9c)). The ratio of mononucleated MN cells to Mononucleated cells pre-update showed a ratio of 1:12.88. This was increased to 1:16.20, owing to the more specific regulations placed on MN scoring, resulting in a greater frequency of others and mononucleated cells. The binucleated MN cell to binucleated cell ratio pre-update was 1:25.14, which was increased once again to 1:30.47.

Thus, a reduction in MN cells was shown in the final Cambridge ground truth, increasing the specificity of MN scored cellular category and ensuring that all the MN assessed to be MN are placed in the correct category, thus ensuring optimal accuracy levels.

c) Re-analysis of network combinations following Cambridge evaluation.

Following the analysis of the Cambridge ground truth and the introduction of the updated ground truth, network creation was carried out. Whereas previously, 4 network combinations were carried out (2^2 due to 2 datasets being used), 9 network combinations were used for this network creation (3^2 due to 3 dataset being used

(Cardiff + Cambridge combined equates to the 3rd dataset)). This was possible due to the Cardiff and Cambridge datasets being combined to form 1 larger dataset. This allowed for a greater quantity of ground truth images to be used in the training and validation of cellular images and thus led to greater accuracy levels.

The 'DeepFlow' neural network was used to generate the 9 network combinations. The most accurate overall network was Network D at 93.6% (Table. 5). Network C produced the greatest accuracy levels for both mononucleated cells (98.6%) and binucleated cells (97.7%) (Table. 5). Network D also produced the highest binucleated MN cell accuracy (89%) (Table. 5). Network A gave the greatest accuracy level to mononucleated MN cells (77.3%)(Table. 5).

When evaluating overall accuracy, it could be shown that the overall accuracy of the networks peaked after the initial Cardiff dataset was amended with Network 1 showing the highest accuracy levels out of all networks produced (93.8%) (Table. 4). However, this accuracy level only dropped by 0.2% in Network D to 93.6%, whereas binucleated MN accuracy increased from 71.8% to 89%, therefore justifying the minute decrease in overall accuracy (Table.4, Table. 5). There was also an increase in accuracy from the most accurate binucleated MN cell network pre-Cambridge dataset update, Network 2 with an 86.5% accuracy, to 89% accuracy in Network D (Table. 4, Table. 5).

Network D, was a network trained using the Cardiff ground truth, validated using the Cambridge ground truth and accuracy levels were shown to be positively affected by updating the ground truth when compared to Network 5, which was trained and validated using the same datasets; but pre-Cambridge ground truth update. In Network D, there was a 0.1% decrease in overall cell accuracy and a 0.6% decrease in binucleated cell accuracy (Table. 5). However, there was also a 7.1% increase in binucleated MN cell accuracy, which is integral to the use of the network in MN analysis (Table. 5 When combining binucleated cells and binucleated MN cells together into an accuracy score out of 200 (Maximum 100% for Binucleated Cell accuracy and 100% for Binucleated MN accuracy), Network 5 scored 180.9, an average of 90.45% (Table. 4). This grouped percentage was 187.4%, an average of 93.7% for Network D (Table. 5). There was also an increase in accuracy levels in mononucleated cells, up from 95.9% to 97.4%, and in mononucleated MN cell

accuracy, up from 67.9% to 72.6% (Table. 4, Table. 5). Therefore, by using the training on Cardiff, validating on Cambridge networks, an increase was shown in the cellular accuracies of the cellular categories most important for MN analysis.

When evaluating mononucleated and mononucleated MN cellular accuracies, increases were shown post-Cambridge update also. The previous highest mononucleated cell accuracy was shown in Network 1 at 96.9%, with Network 5 showing the previous highest 2 channel mononucleated cell accuracy level at 95.9% (Table. 4). These levels were increased by 0.8% and 1.8% respectively to 97.7% in Network C (Table. 5). The previously highest rated mononucleated MN cell accuracy was Network 9 at 72.4%, produced in the 3-channel approach, and 71.4% shown in the 3-channel approach produced by Network 2 (Table. 4). An increase of 4.9% and 5.9% was shown to generate the 77.3% accuracy level shown in Network A (Table. 5). Interestingly, the most accurate network for mononucleated MN cells in all the network analysis was trained using a combination of the Cambridge and Cardiff ground truth, Network A (Table. 5). This allows for a greater quantity of images to be used to determine the cellular categories each cellular image should be placed in.

It must be noted however, the most accurate network for binucleated MN cells, Network D was trained on just one dataset, the Cardiff ground truth and validated on only the Cambridge ground truth (Table. 5). A simple explanation could have been that the differences, however negligible they appear, between the nuclear stains used in the Cardiff and Cambridge ground truths (Draq5 vs Hoescht 33342) may have accounted for some confusion when training the network using the combined ground truth. However, the mononucleated MN cell category thrived on using the ground truth combination approach, therefore the difference in nuclear stains appears to have a negligible effect. Network H produced the highest accuracy levels for binucleated MN cell accuracy from the pool of combined ground truth networks, with an 84.5% accuracy shown (Table. 5). However, when viewing the mononucleated MN cell accuracy, it was only 53.6% accurate (Table. 5). An explanation for this may be the confusion of the network in confusing mononucleated MN cells as mononucleated/binucleated cells, given the slight drop shown in both mononucleated and binucleated cells when compared to the singularly trained Network D (1.5% in binucleated cells and 2% in mononucleated cells (Table. 5). This may seem an insignificant decrease, but Fig. 11 shows 211 mononucleated MN cells were scored

initially, an accuracy level of 53.6% denoting 98 of the 211 mononucleated MN cells were mis-categorised. Thus showing how a potential difference in nuclear stains could have caused the neural network to mis-categorise a selection of mononucleated MN and possibly binucleated MN cells and therefore explaining why the combined ground truth did not necessarily produce the most accurate networks.

In conclusion, owing to the nature of the MN assay, which can be carried out with and without cytochalasin-B and thus assessing binucleated or mononucleated MN cells, the most accurate network for each cellular category was researched, analysed and compared for use in dose response analysis. Thus, Network D was used going forward in the cytochalasin-B MN assay due to the highest binucleated MN cell accuracy level (89%) and highest combination of binucleated cell and binucleated MN cell accuracy (187.4) (Table. 5). Network A was thus chosen for the non-cytochalasin-B MN assay due to the highest mononucleated MN cell accuracy (77.3%) and highest combination of mononucleated cell and mononucleated MN cell accuracy (174.1) (Table. 5).

ix) Dose response analysis

To fully evaluate the neural network approach and accuracy to analysis in the MN assay, a dose response was carried out using Carbendazim dosed on TK6 cells with cyto-B in a GSK laboratory and analysed using a variety of networks to showcase the accuracy of the neural network approach (Fig. 14). Manual scoring of the cellular images produced from the imaging flow cytometer were also analysed to form a comparison of neural network accuracy to a method comparable to the ‘gold standard’ of manual light microscopy (Verma *et al.*, 2017). The background levels of MN testing are between 0.32%-1.38% (Lovell *et al*., 2018). Therefore, the

background levels were evaluated using both the manual scoring of images approach and neural network also. It is of note that the Swansea historical lab background levels for MN rate in TK6 cells is 0.9%, which sits in between the historical 0.32%-1.38%. The background manual scoring levels shown were 0.72% (Fig. 14), which sits in between the historical figures and slightly lower than the Swansea historical value. This can be explained, due to the dosing taking place in the GSK laboratory and thus more variables existing involved in determining the change in background MN levels. The most accurate Swansea Network ('NETWORK D' Table. 5), had background MN levels of 0.66% (Fig. 14), which is slightly lower than the manual scoring average but sits in between the historical MN rates of 0.32%-1.38% (Lovell *et al.,* 2018). This shows the potential for neural network use in MN analysis and the ability of the network to accurately score background MN levels in different laboratories, showing the cross compatibility of the approach.

a) i) Stats

When assessing the normality of the GSK dataset and the MN distribution between replicates, the Shaprio-Wilks test showed that all the networks used in the GSK dose response testing (Fig. 14) had a P value of >0.05 and therefore the data was normally distributed.

When carrying out the Dunnett's test, each of the three doses used, 0.8μg/ml, 1.2μg/ml and 1.6μg/ml, showed a significant increase when compared to the control. Therefore, the lowest observed effect level (LOEL) was the 0.8μg/ml dose of Carbendazim. Since only three doses were carried out in the dose response analysis for the GSK dataset, a determination could not be made on the 0.4μg/ml dose which was used in the Newcastle and Cardiff analysis (Fig. 15).

To calculate the network with the greatest accuracy in respect to the manual scoring of cellular images, a variation distribution was taken. Comparing the spread of ranges and variation distribution between each network and the 'control' of manually scored cells. The spread of: range, standard deviation, standard error and coefficient of variation were all analysed and interpreted to understand which network provided the greatest results in practical terms when used in the dose response setting, for which they were created for. Therefore, it is important to note, that despite specific networks showing higher accuracy levels, these accuracy levels were not always translated into he dose response shown when compared to the manually scored images, which in this case is the 'gold standard' (Verma *et al*., 2017). The variations between network results and manually scored images was compared to the variation and range levels shown when just assessing the manually scored images, to gage values to form a comparison based on.

The Newcastle dataset focused on assessing mononucleated cells with a MN to carry out a dose response, due to the lack of cytochalasin-B. Four doses were used in the dose response, a control, 0.4μg/ml, 0.8μg/ml, 1.2μg/ml and 1.6μg/ml (Fig. 20). Thus, similar doses were used in comparison to the GSK dose response, with the extra

addition of the lower 0.4µg/ml dose. The same doses were used in Cardiff analysis of the dose response also and therefore allowed for a direct comparison to be made between the Cardiff and Newcastle datasets. In the Newcastle dose response, the data was once again found to be normalised and therefore a Dunnett's test could once again be carried out. Interestingly, since a lower dose was added, the 0.4µg/ml, it was found that a significant increase in MN frequency was not found for this dose. Therefore, the no observed effect level (NOEL) for the Newcastle dataset was 0.8µg/ml and the LOEL was 1.0µg/ml (Fig. 15, Fig. 16).

ii) Further Dose response analysis

A dose response was carried out on the Cardiff network to assess network reproducibility across different laboratories. By assessing the dose response of Carbendazim on mononucleated cells, a comparison was made between the Newcastle and Cardiff networks using the same chemical (Carbendazim), same cell type (mononucleated cells, no cytochalasin-B used) and using the same neural network (Table. 5, Network A).

When assessing the range and coefficient of variations found when comparing the mean of the Cardiff and Newcastle non-cyto-B Carbendazim samples, a cumulative 36.73% coefficient of variation was shown. This figure was lower than any previously carried out in the GSK data analysis, where the lowest coefficient of variation shown was 50.19%, which was used in comparing all the values of the most up to date network and the manual scoring approach. Considering the coefficient of variation was 52.92% when comparing the values of the manual scoring triplicates, great confidence can be taken from the 36.73% variation of coefficient value produced in the dose response comparison of the mean values shown following analysis of the Cardiff and Newcastle laboratories. Moreover, the figure can be further examined to show a 0.66% and 3.04% coefficient of variations produced for the middle two doses of 0.8µg/ml and 1.0µg/ml in the comparison of Cardiff and Newcastle networks, showing large degrees of similarity, gaining confidence in the application of the same neural network into multiple datasets.

Therefore, the production of dose responses across three laboratories (GSK, Cardiff and Newcastle) provided the basis for the reproducibility of using a ground truth based neural network to assess dose responses across different laboratories and using the same neural network. All three networks focused on the use of TK6 cells treated with Carbendazim with/without the presence of Cytochalasin-B and all three networks showed a significant increase in MN frequency at higher doses, in line with expected results. This shows a proof of concept study into the use of deep learning, ground truth based neural networks in the dose response setting of the MN assay and the potential to revolutionise this assay by streamlining the procedure, whilst maintaining accuracy.

b) Assessment of MCL-5 and AHH-1 populations

The neural networks produced were used in the assessment of MCL-5 and AHH-1 populations to identify the far-reaching uses of the network. Network 'A' was used to determine the background levels of MN due to the MCL-5 and AHH-1 cells being cultured without the use of cytochalasin-B and thus mononucleated cells and mononucleated MN cells were used in the determination of MN frequency.

Therefore, Network 'A', with an accuracy of 77.3%, was used in the analysis. A background rate of 0.74% was observed in the AHH-1 cells and 0.66% in the MCL-5

cells (Figure. 22). This is consistent with the background rate of MN in literature as it is between 0.32% and 1.38% (Lovell *et al*., 2018). However, this is lower than the historical background MN levels in our lab, which sit at 1.3% for AHH-1 cells and 1.48% for MCL-5 cells. This is an almost exact doubling of the results shown by Network 'A'. There are multiple reasons for this being the case. MCL-5 ad AHH-1 cells are not the same size as TK6 cells, this is highlighted when using a coulter counter to measure the quantity of cells in a sample, 5-17μm is used as standard for MCL-5 cells, whereas TK-6 cells are 10-17μm. The difference in size may account for the lower rates shown compared to the norm. With the training and test data comprising of different sized cells and thus adding to the confusion. However, it must be noted, that despite the background levels being lower than the historical average for our Swansea lab, the results still sit within the norm of 0.32-1.38% (Lovell *et al*., 2018).

When O^6-BG was added to the MCL-5 cell line respectively, a major MN frequency increase was shown from 0.66% to 1.92% (Figure. 18). This is somewhat unsurprising, given the role of O^6-BG in preventing DNA repair via MGMT (Estellar *et al., 1999*)**.** Despite this being the control sample and exogenous damage not being present, endogenous damage does still take place in the cell lines and by adding O^6-BG and inhibiting DNA repair, some of this endogenous damage remains un-repaired and gives rise to chromosomal damage in the form of breaks and addition which presents itself as MN (Ochs and Kaina, 2000). Therefore, there is an increase in MN frequency shown in Fig. 18. Moreover, the MCL-5 cell line contains 5 cytochromes p450s, including CYP1A1, this allows for greater levels of metabolic activity, which can heighten DNA damage via metabolising endogenous sources at a far greater rate (Crofton-Sleigh *et al., 1993)*. It is the metabolic products which tend to cause DNA damage rather than the original source itself. More work is needed on the application of neural networks to MCL-5 and AHH-1 dosing. This initial work shows the potential to use the neural network to test background levels in these cell lines, a full dose response using MCL-5 and AHH-1 cell lines would be the next step and a comparison to be made to the TK-6 dose response evaluated here, in both samples with and without cytochalasin-B. There would also be the potential to create a ground truth using the MCL-5 and AHH-1 networks and to assess how this affects network accuracy and compared to using the TK-6 based ground truth55.

x) GSK ground truth population

Following the manual scoring of the GSK dataset, in order to provide a comparison to the automated neural network approach, each individual cell was attributed a cellular phenotype in accordance with the categories used in the ground truth. Despite this being carried out as a comparison and to calculate the replicative index (RI), the result was the annotation of 25,805 cellular images (Fig. 12).

This data has the potential to be used in future neural network analysis as a 3rd dataset, once the word document containing the phenotypes of the cellular categories has been applied to a set of code. This allows for MatLab® to differentiate the images into their separate categories and to use as a ground truth. This has the possibility to further increase the robustness of the neural network approach. By applying a third dataset, there can be a greater volume of cellular images of the rarer phenotypes, especially the ‘Mononucleated MN cells’ and ‘Binucleated MN’ cells, which are of great interest in dose response analysis. By adding a third dataset, a larger quantity of MN images will be obtained overall, allowing for a further increase in ‘Mononucleated’ and ‘Binucleated MN cell’ accuracy. This is of particular interest in the ‘Mononucleated MN’ cell category, where the greatest dose response produced by the network was 77.3% (Table. 5), showing room for an improvement in accuracy levels.

5.0 Conclusion

In conclusion, a clear comparison was shown between the deep learning approach to the automation of the *in vitro* MN assay and the manual assessment of cellular images for carrying out the *in vitro* MN assay with cyto-B, shown using the assessment of Carbendazim and compared to manual scoring approaches and the historical gold standard (Verma *et al*., 2017).

The GSK cyto-B dose response showed a clear comparison between automated and manual scoring approaches, showing that the integrity of the assessment had not been compromised in the streamlining of the scoring approach. The results were comparable when using the deep learning neural network for assessment, despite the process being far less laborious and time consuming than the gold standard of manual scoring using light microscopy (Fenech, 2000).

The neural network was clearly developed, resulting in a model with peak accuracy levels achieved and showing an 89% accuracy on Binucleated MN cells despite only training from a ground truth population of 155 Binucleated MN cells. It is no small feat to achieve accuracy levels this high using this little quantity of ground truth images. Moreover, the applicability of the network was shown using the Cardiff and Newcastle datasets, whereby dose response results were found to be comparable to one another. This streamlines the process even further, as it allows the user to use a pre-created ground truth and therefore network to analyse samples. The reproducibility of the dose responses across the three laboratories shown, shows great promise for this technique in the future. More chemicals are required to be testing, outside the scope of Carbendazim, to allow this method to be used on a wider scale. Moreover, a dataset comprising of manually scoring a non cyto-B MN assay could be carried out in the future to provide a direct comparison when carrying out the assessment of mononucleated and mononucleated with MN cells.

It is thought that aneugens and clastogens produce slightly different shaped MN, this can therefore prove a potential stumbling block in the application of the ground truth to other datasets. However, the Cardiff network was trained using a mixture of

aneugens and clastogens and therefore one would expect dose response accuracy levels to be maintained following the assessment of clastogens also.

Initial studies were carried out on MCL-5 and AHH-1 cells and showed great similarity to one another and potential for this cell line to be assessed using a ground truth formed entirely of TK-6 cells. This would once again streamline the method further, as the creation of a separate ground truth to assess MCL-5 and AHH-1 cells may not be required. This could be a huge addition, as MCL-5 and AHH-1 cells are commonly used in the assessment of genotoxic and carcinogenic compounds requiring metabolic activation, such as NDMA. Therefore, an interesting future study can be centred around the assessment of NDMA and other nitrosamines using this deep learning neural network approach and to determine the use of this method in producing a dose response to MCL-5 and AHH-1 cell lines using different chemical compounds.

More future work could be carried out in carrying out a manually scored dose response triplicate for the non cyto-B MN assay as mentioned previously, which would allow for a direct comparison between the automated deep earning assessment of mononucleated cells and mononucleated MN cells in determining a dose response. In this study, the non cyto-B results were compared to other laboratories results and then again to the historical background figures to determine correlation to the ‘gold standard’. However, carrying out a manually scored comparison of such mononucleated and mononucleated with MN cells would be something to consider yet.

Moreover, forming greater ground truth populations, as this would lead to the increase in the rarer cellular types, such as MN, and allow for a further increase in neural network accuracy. The greatest Mononucleated MN cell accuracy was 77.3%, showing plenty of room for the improvement of the accuracy. This was highlighted in the Newcastle and Cardiff dose responses, whereby only the top Cardiff dose was found to have a statistically significant dose response to the control sample and only the top doses were found to have a statistically significant dose response in the Newcastle dataset. The dose responses recorded were slightly lower than the historical MN average for Carbendazim in the non-cytochalasin-B MN assay and

were also shown too be far lower than the values found in the GSK cytochalasin-B dose response. Therefore, the Mononucleated MN cell accuracy can be attributed to part of the reason for the reduction in dose response, suggesting that in this case, the Mononucleated MN accuracy levels were not high enough and therefore too specific in this case in the analysis of cellular phenotypes. By adding more ground truth images, it would be possible to allow for a greater quantity of Mononucleated MN cells to be manually assessed and then used in the training for a neural network.

The addition therefore of the GSK ground truth to future neural network assessments should allow for an increase to be shown in all cellular categories, but especially Mononucleated MN cell accuracy. This GSK ground truth, which I manually scored when forming a comparison to the automated accuracy of Network A, has the potential to further increase the accuracy, specificity and sensitivity of this method in the *in vitro* MN assay assessment.

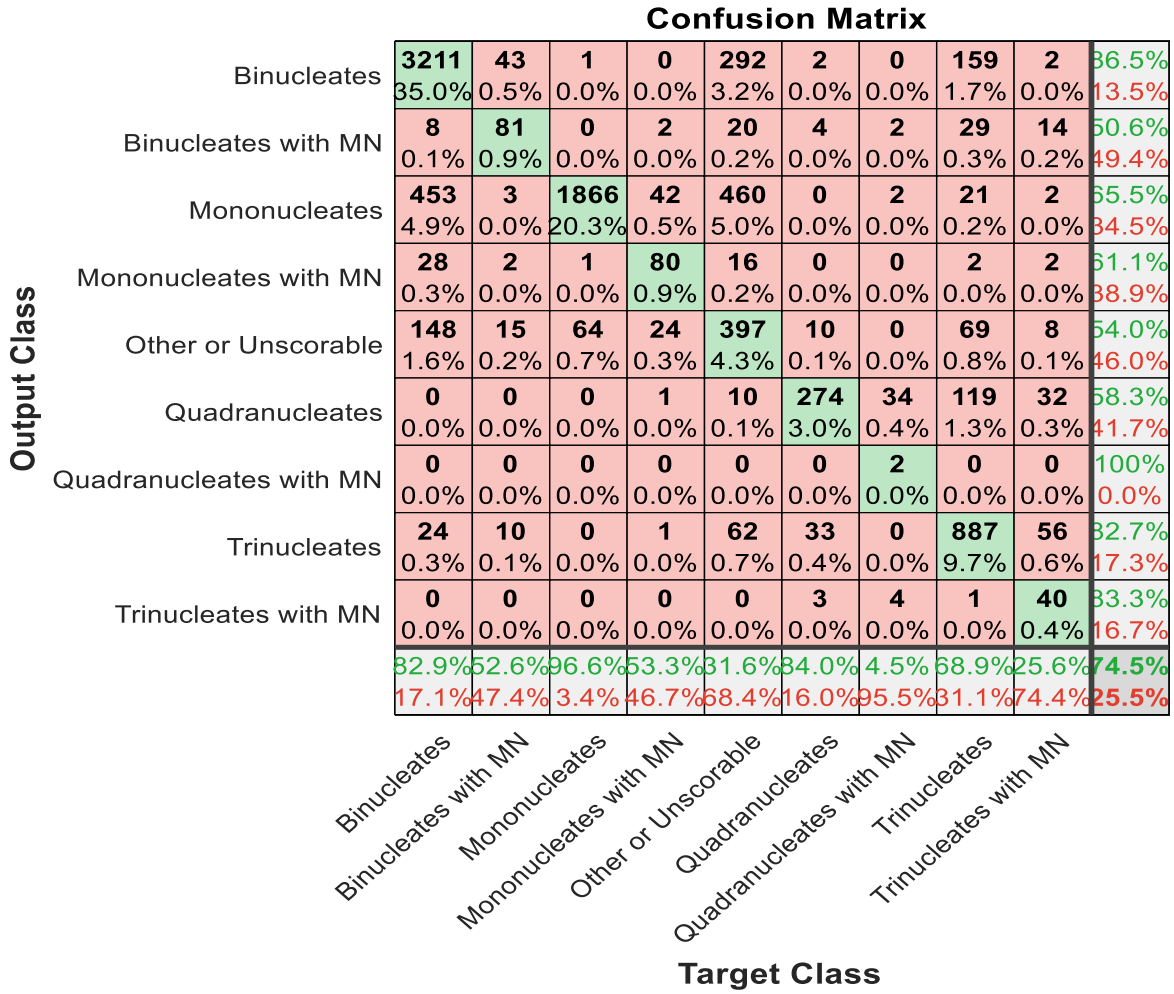
The deep learning neural network approach, a novel approach in this research setting, was therefore shown to produce a dose response following Carbendazim treatment. This dose response was shown to be comparable to the manual scoring of cellular images produced following imaging flow cytometry processing (a comparable method to the gold standard of manual light microscopy assessment).

By definition therefore, this approach is comparable with the gold standard and maintains the integrity of the results, whilst streamlining the method and the time taken to historically complete the *in* vitro MN assay.

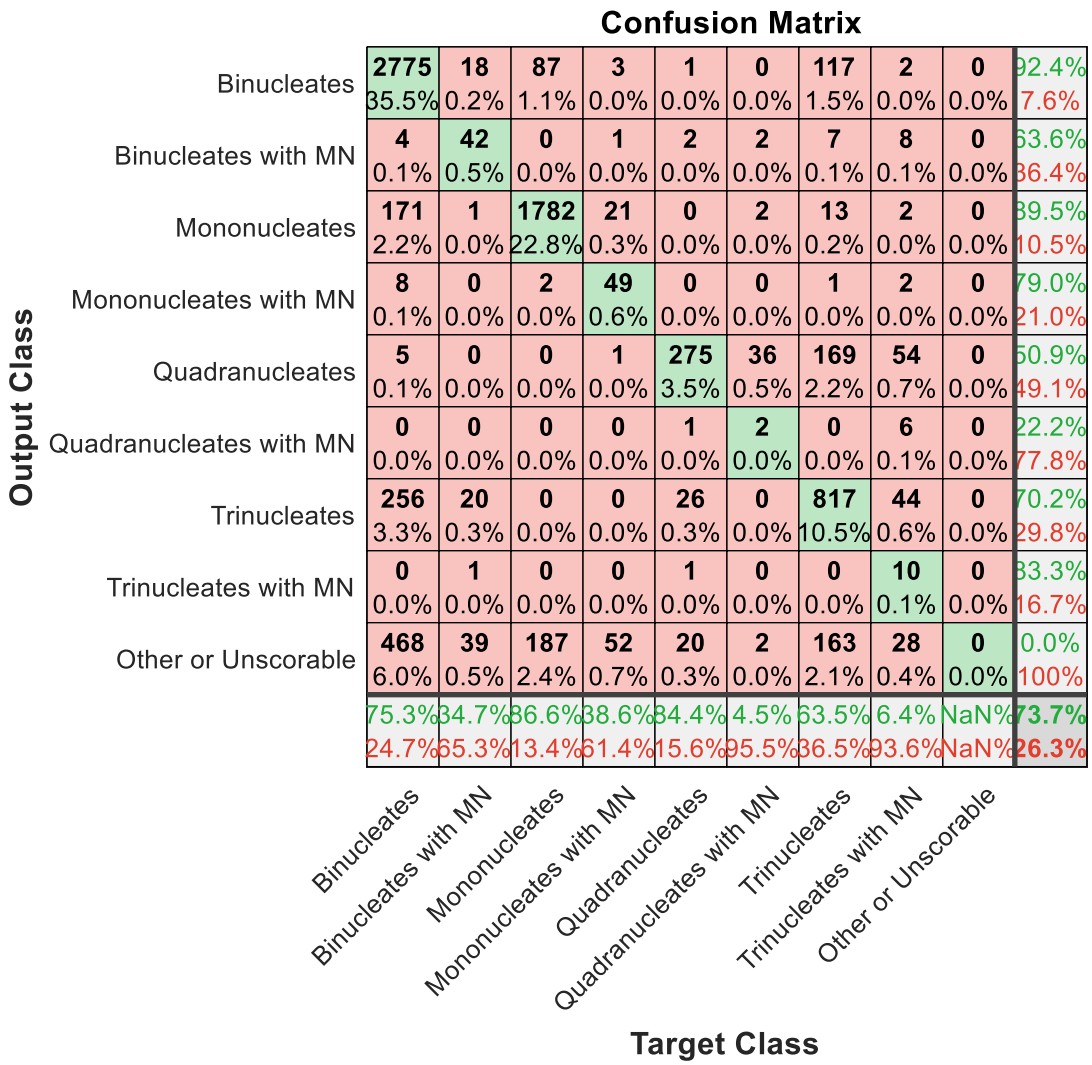
Appendix

Confusion matrices and neural network formation provided in addition to the training with Cardiff, validating with Cardiff (Fig. 13) which is provided in the main text under results. The following datasets are included here: 1) Training with Cardiff, validating with Cambridge (Fig. 20), 2) Training with Cambridge, validating with Cardiff (Fig. 21) and 3) Training with Cambridge, validating with Cambridge (Fig. 22).

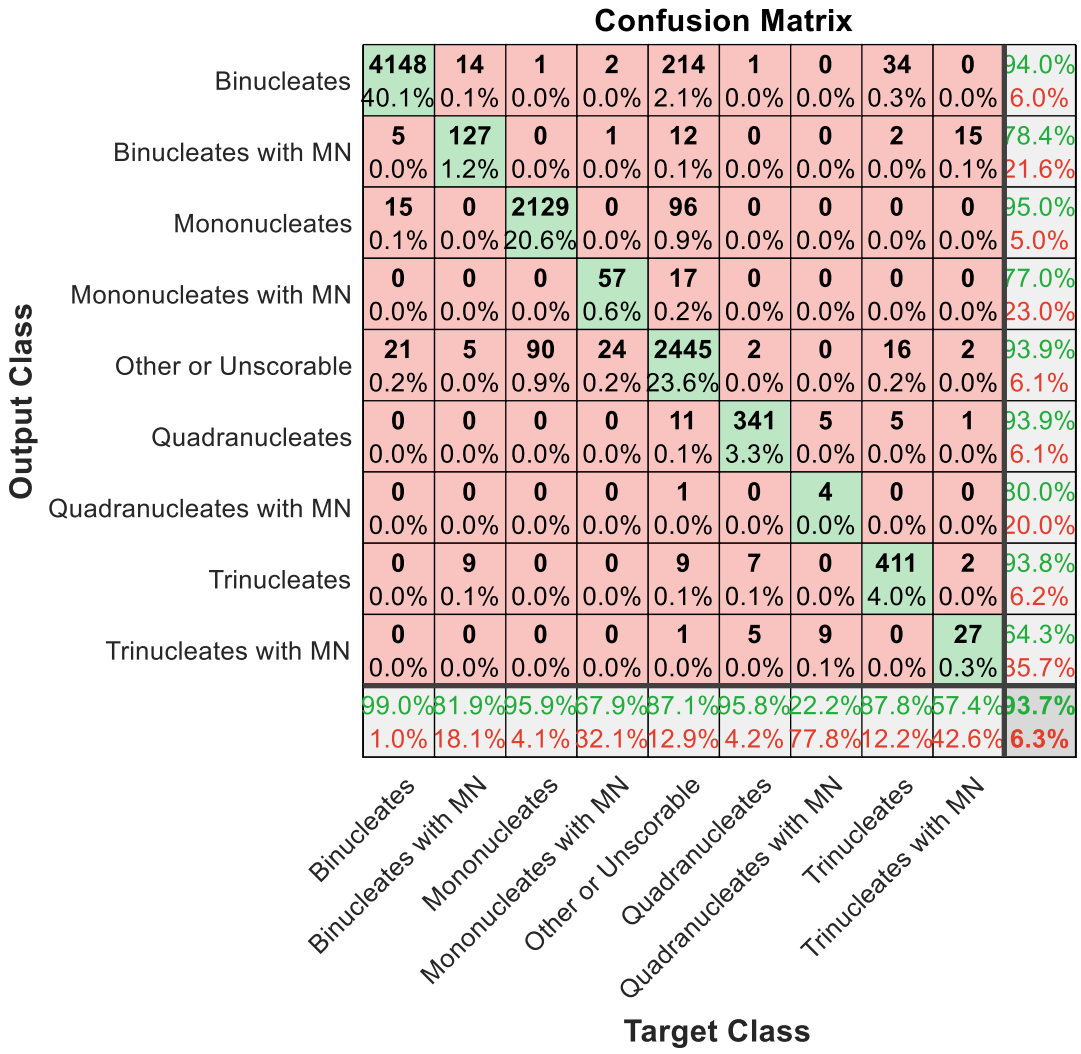
A



B

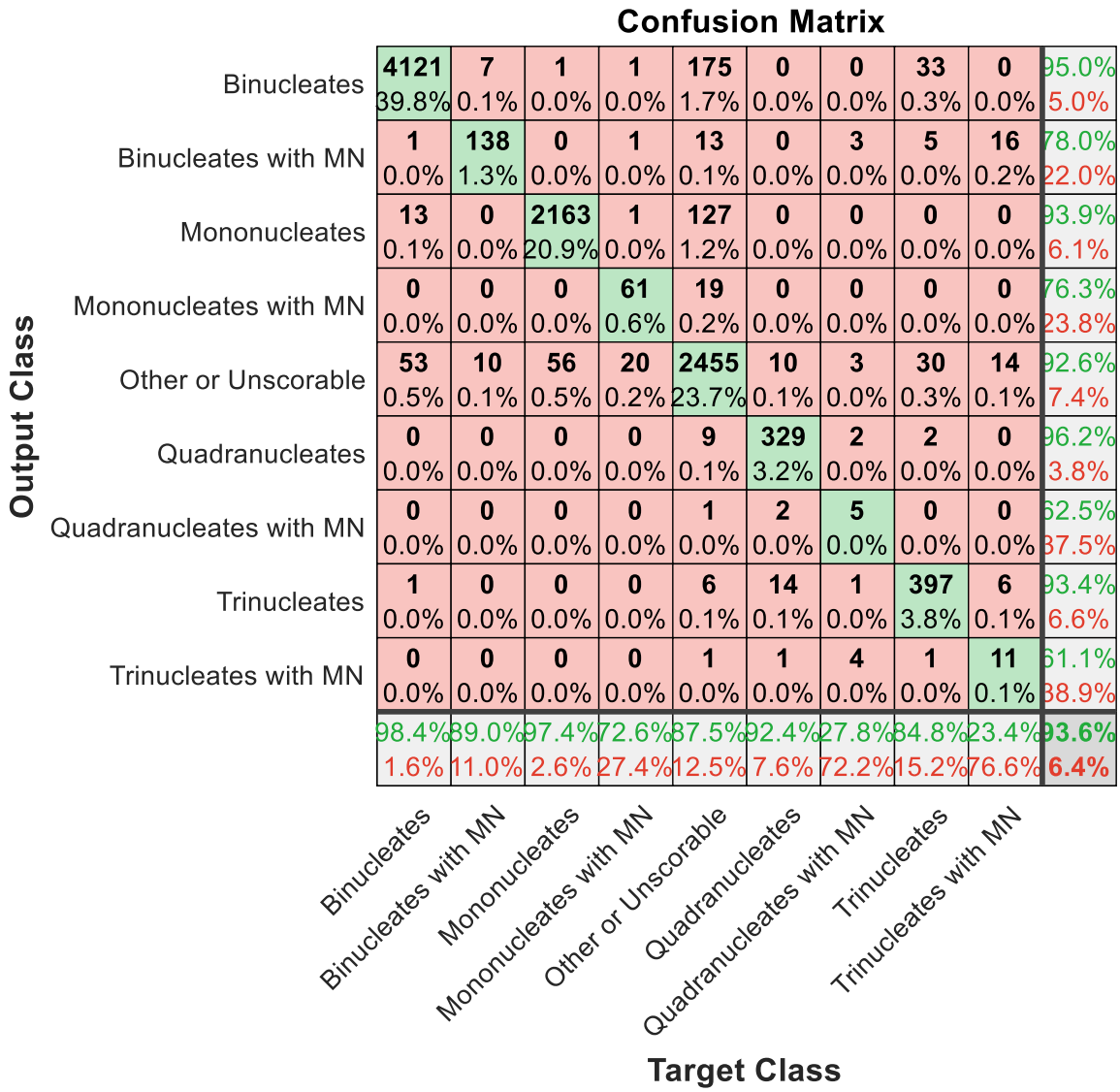

Confusion Matrix
Output Class
Target Class
Binucleates
Binucleates with MN
Mononucleates
Mononucleates with MN
Quadranucleates
Quadranucleates with MN
Trinucleates
Trinucleates with MN
Other or Unscorable
2775 18 87 3 1 0 117 2 0 92.4%
35.5% 0.2% 1.1% 0.0% 0.0% 0.0% 1.5% 0.0% 0.0% 7.6%
4 42 0 1 2 2 7 8 0 63.6%
0.1% 0.5% 0.0% 0.0% 0.0% 0.0% 0.1% 0.1% 0.0% 36.4%
171 1 1782 21 0 2 13 2 0 89.5%
2.2% 0.0% 22.8% 0.3% 0.0% 0.0% 0.2% 0.0% 0.0% 10.5%
8 0 2 49 0 0 1 2 0 79.0%
0.1% 0.0% 0.0% 0.6% 0.0% 0.0% 0.0% 0.0% 0.0% 21.0%
5 0 0 1 275 36 169 54 0 50.9%
0.1% 0.0% 0.0% 0.0% 3.5% 0.5% 2.2% 0.7% 0.0% 49.1%
0 0 0 0 1 2 0 6 0 22.2%
0.0% 0.0% 0.0% 0.0% 0.0% 0.0% 0.0% 0.1% 0.0% 77.8%
256 20 0 0 26 0 817 44 0 70.2%
3.3% 0.3% 0.0% 0.0% 0.3% 0.0% 10.5% 0.6% 0.0% 29.8%
0 1 0 0 1 0 0 10 0 83.3%
0.0% 0.0% 0.0% 0.0% 0.0% 0.0% 0.0% 0.1% 0.0% 16.7%
468 39 187 52 20 2 163 28 0 0.0%
6.0% 0.5% 2.4% 0.7% 0.3% 0.0% 2.1% 0.4% 0.0% 100%
75.3% 34.7% 86.6% 38.6% 84.4% 4.5% 63.5% 6.4% NaN% 73.7%
24.7% 65.3% 13.4% 61.4% 15.6% 95.5% 36.5% 93.6% NaN% 26.3%

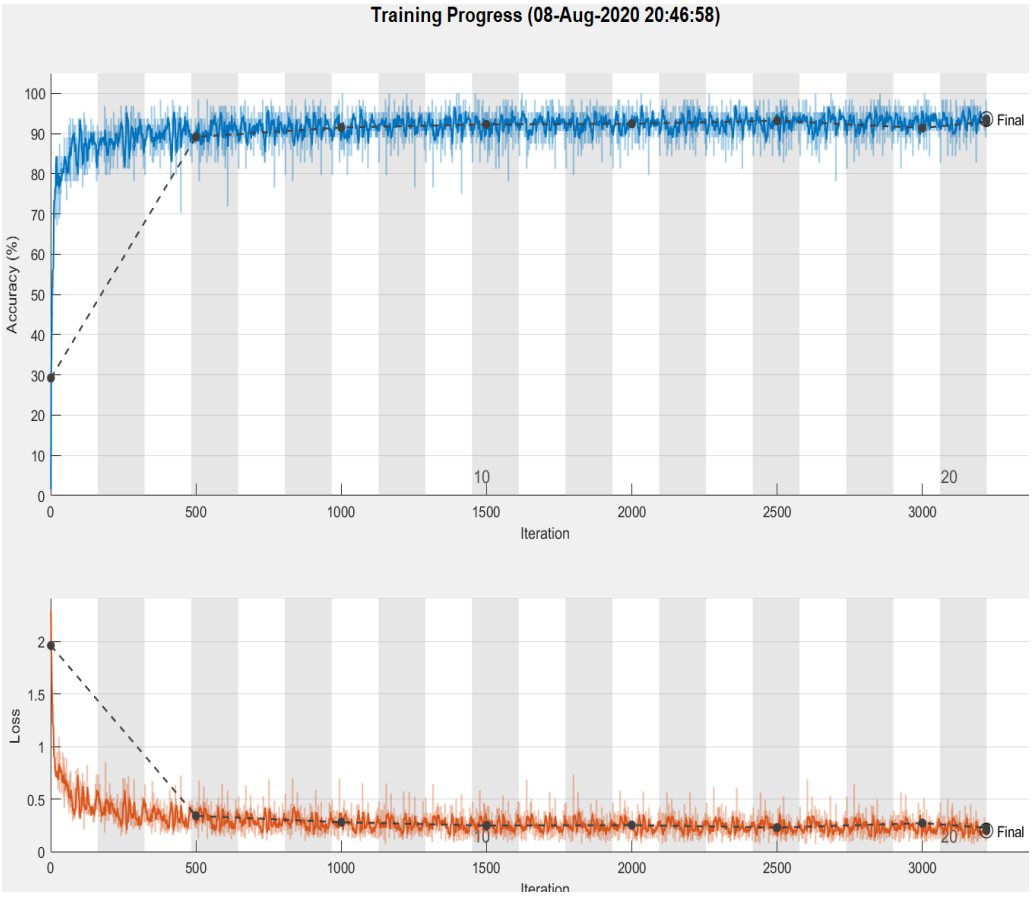
C


Confusion Matrix
Output Class
Binucleates
Binucleates with MN
Mononucleates
Mononucleates with MN
Other or Unscorable
Quadranucleates
Quadranucleates with MN
Trinucleates
Trinucleates with MN
4148 40.1% 14 0.1% 1 0.0% 2 0.0% 214 2.1% 1 0.0% 0 0.0% 34 0.3% 0 0.0% 94.0% 6.0%
5 0.0% 127 1.2% 0 0.0% 1 0.0% 12 0.1% 0 0.0% 0 0.0% 2 0.0% 15 0.1% 78.4% 21.6%
15 0.1% 0 0.0% 2129 20.6% 0 0.0% 96 0.9% 0 0.0% 0 0.0% 0 0.0% 0 0.0% 95.0% 5.0%
0 0.0% 0 0.0% 0 0.0% 57 0.6% 17 0.2% 0 0.0% 0 0.0% 0 0.0% 0 0.0% 77.0% 23.0%
21 0.2% 5 0.0% 90 0.9% 24 0.2% 2445 23.6% 2 0.0% 0 0.0% 16 0.2% 2 0.0% 93.9% 6.1%
0 0.0% 0 0.0% 0 0.0% 0 0.0% 11 0.1% 341 3.3% 5 0.0% 5 0.0% 1 0.0% 93.9% 6.1%
0 0.0% 0 0.0% 0 0.0% 0 0.0% 1 0.0% 0 0.0% 4 0.0% 0 0.0% 0 0.0% 80.0% 20.0%
0 0.0% 9 0.1% 0 0.0% 0 0.0% 9 0.1% 7 0.1% 0 0.0% 411 4.0% 2 0.0% 93.8% 6.2%
0 0.0% 0 0.0% 0 0.0% 0 0.0% 1 0.0% 5 0.0% 9 0.1% 0 0.0% 27 0.3% 64.3% 35.7%
99.0% 1.0% 81.9% 18.1% 95.9% 4.1% 67.9% 32.1% 87.1% 12.9% 95.8% 4.2% 22.2% 77.8% 87.8% 12.2% 57.4% 42.6% 93.7% 6.3%
Binucleates
Binucleates with MN
Mononucleates
Mononucleates with MN
Other or Unscorable
Quadranucleates
Quadranucleates with MN
Trinucleates
Trinucleates with MN
Target Class

D

I


Confusion Matrix
Output Class
Target Class
Binucleates
Binucleates with MN
Mononucleates
Mononucleates with MN
Other or Unscorable
Quadranucleates
Quadranucleates with MN
Trinucleates
Trinucleates with MN
4121 39.8% | 7 0.1% | 1 0.0% | 1 0.0% | 175 1.7% | 0 0.0% | 0 0.0% | 33 0.3% | 0 0.0% | 95.0% 5.0%
1 0.0% | 138 1.3% | 0 0.0% | 1 0.0% | 13 0.1% | 0 0.0% | 3 0.0% | 5 0.0% | 16 0.2% | 78.0% 22.0%
13 0.1% | 0 0.0% | 2163 20.9% | 1 0.0% | 127 1.2% | 0 0.0% | 0 0.0% | 0 0.0% | 0 0.0% | 93.9% 6.1%
0 0.0% | 0 0.0% | 0 0.0% | 61 0.6% | 19 0.2% | 0 0.0% | 0 0.0% | 0 0.0% | 0 0.0% | 76.3% 23.8%
53 0.5% | 10 0.1% | 56 0.5% | 20 0.2% | 2455 23.7% | 10 0.1% | 3 0.0% | 30 0.3% | 14 0.1% | 92.6% 7.4%
0 0.0% | 0 0.0% | 0 0.0% | 0 0.0% | 9 0.1% | 329 3.2% | 2 0.0% | 2 0.0% | 0 0.0% | 96.2% 3.8%
0 0.0% | 0 0.0% | 0 0.0% | 0 0.0% | 1 0.0% | 2 0.0% | 5 0.0% | 0 0.0% | 0 0.0% | 62.5% 37.5%
1 0.0% | 0 0.0% | 0 0.0% | 0 0.0% | 6 0.1% | 14 0.1% | 1 0.0% | 397 3.8% | 6 0.1% | 93.4% 6.6%
0 0.0% | 0 0.0% | 0 0.0% | 0 0.0% | 1 0.0% | 1 0.0% | 4 0.0% | 1 0.0% | 11 0.1% | 61.1% 38.9%
98.4% 1.6% | 89.0% 11.0% | 97.4% 2.6% | 72.6% 27.4% | 87.5% 12.5% | 92.4% 7.6% | 27.8% 72.2% | 84.8% 15.2% | 23.4% 76.6% | 93.6% 6.4%



III

Results

Validation accuracy:	80.58%
Training finished:	Reached final iteration

Training Time

Start time:	08-Aug-2020 20:46:58
Elapsed time:	41 min 46 sec

Training Cycle

Epoch:	20 of 20
Iteration:	3220 of 3220
Iterations per epoch:	161
Maximum iterations:	3220

Validation

Frequency:	500 iterations

Other Information

Hardware resource:	Multiple GPUs
Learning rate schedule:	Piecewise
Learning rate:	0.001

Figure. 20 Confusion matrices produced post neural network creation using MatLab®. Training on the 'Cardiff' dataset and validation on the 'Cambridge' data set. Overall accuracies shown as well as accuracies per individual subgroups.

A) Neural network formed after using the initial ground truth population and a 3-channel approach of: Brightfield, Fluorescence, Fluorescence. Network 3 produced this confusion matrix.

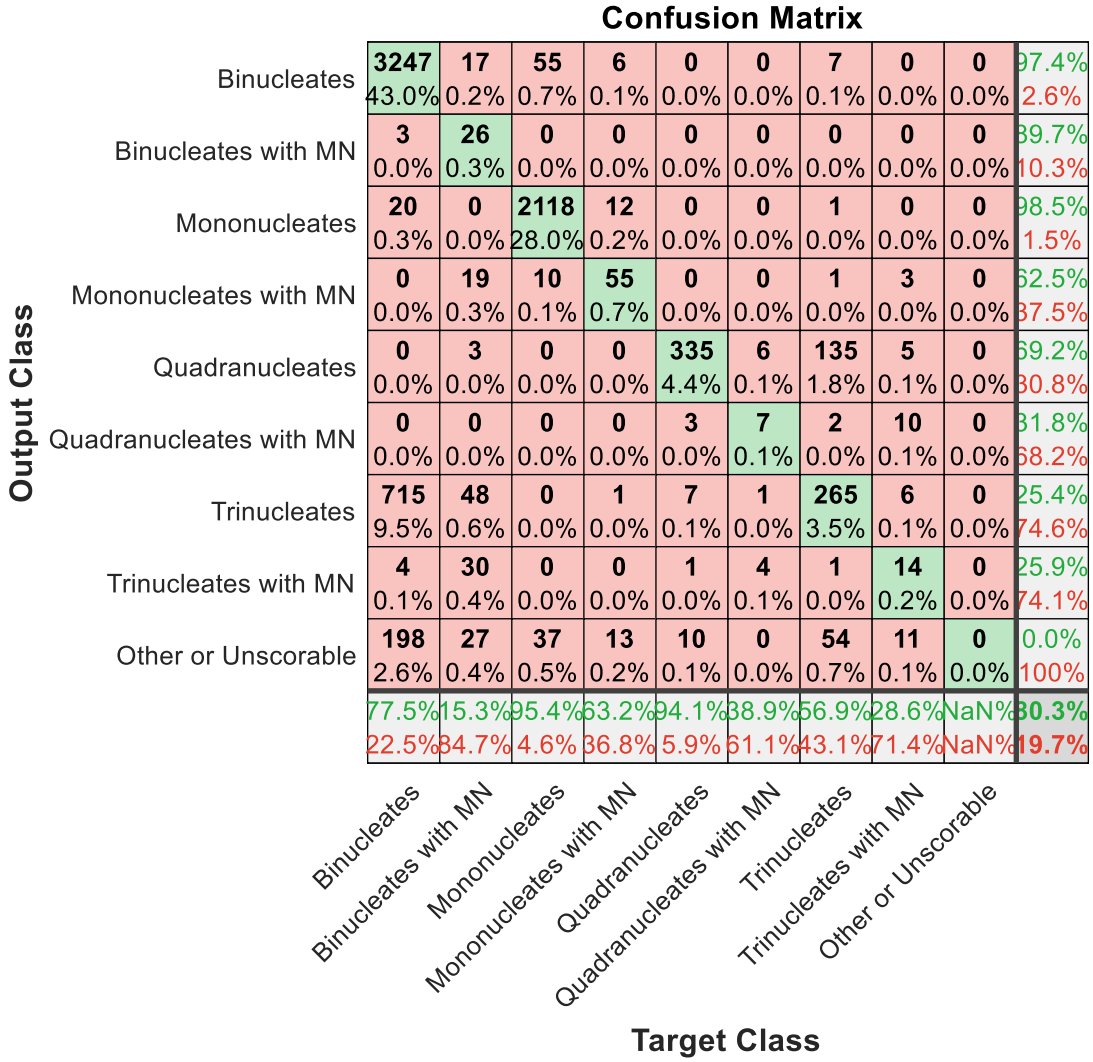
B) Neural network formed after using the first updated ground truth and using a 3-channel approach of Brightfield, Fluorescence, Fluorescence, Network 4 produced this confusion matrix.

C) Neural network formed after using the first updated ground truth and the 2-channel approach of: Brightfield and Fluorescence. Network 5 produced this confusion matrix.

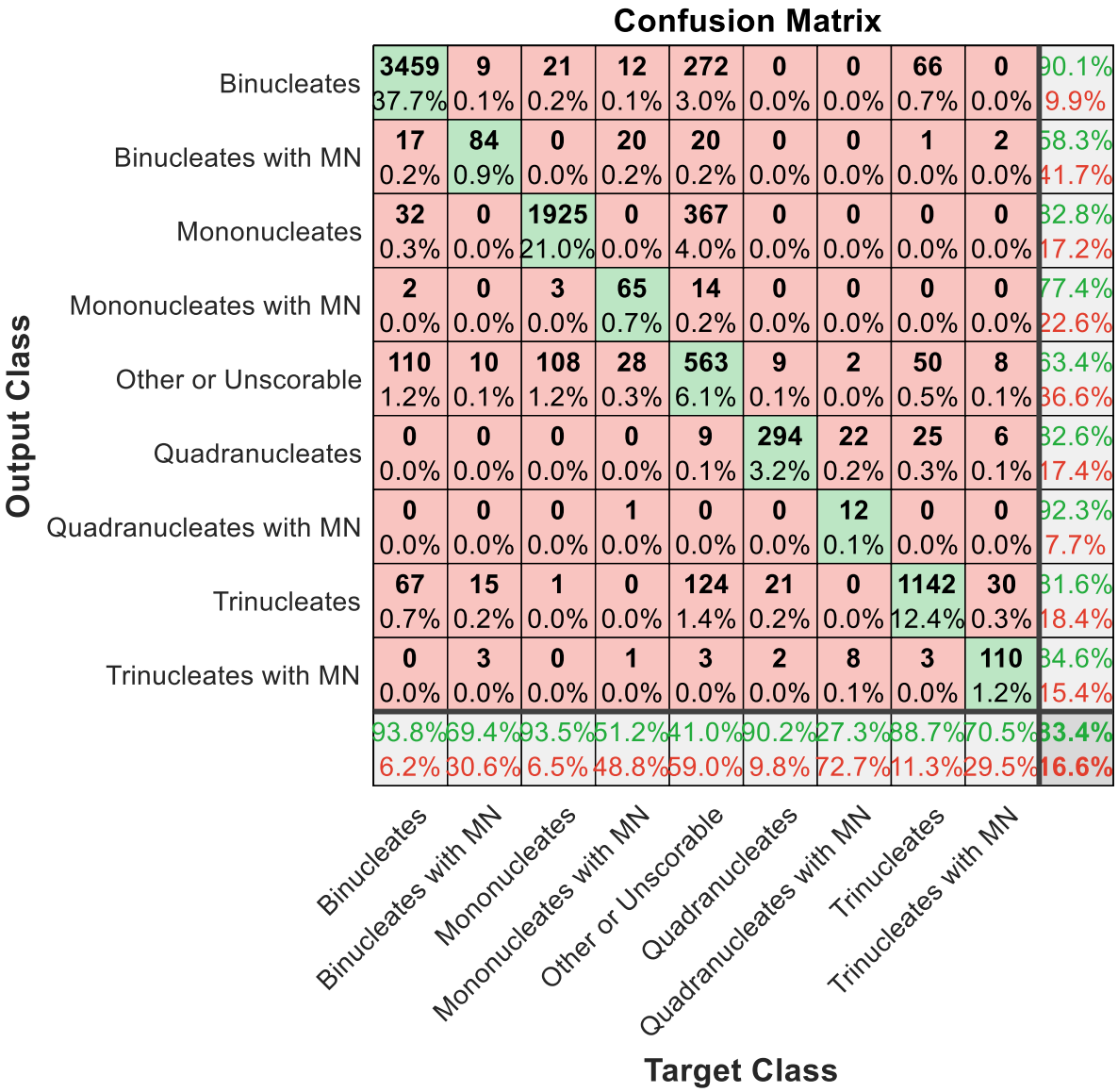
D) i) Neural network formed post Cardiff and Cambridge ground truth updates and the 2-channel approach of: Brightfield and fluorescence (1, 11). Network D produced this network

ii) Figure showing the training development of Network D up to 20 epochs. Accuracy and error rate are both shown.

iii) The results section after completing a network run, showing the accuracy rate of the network produced after completing 20 epoch cycles, other key statistics are also shown, such as iterations taken and time elapsed.

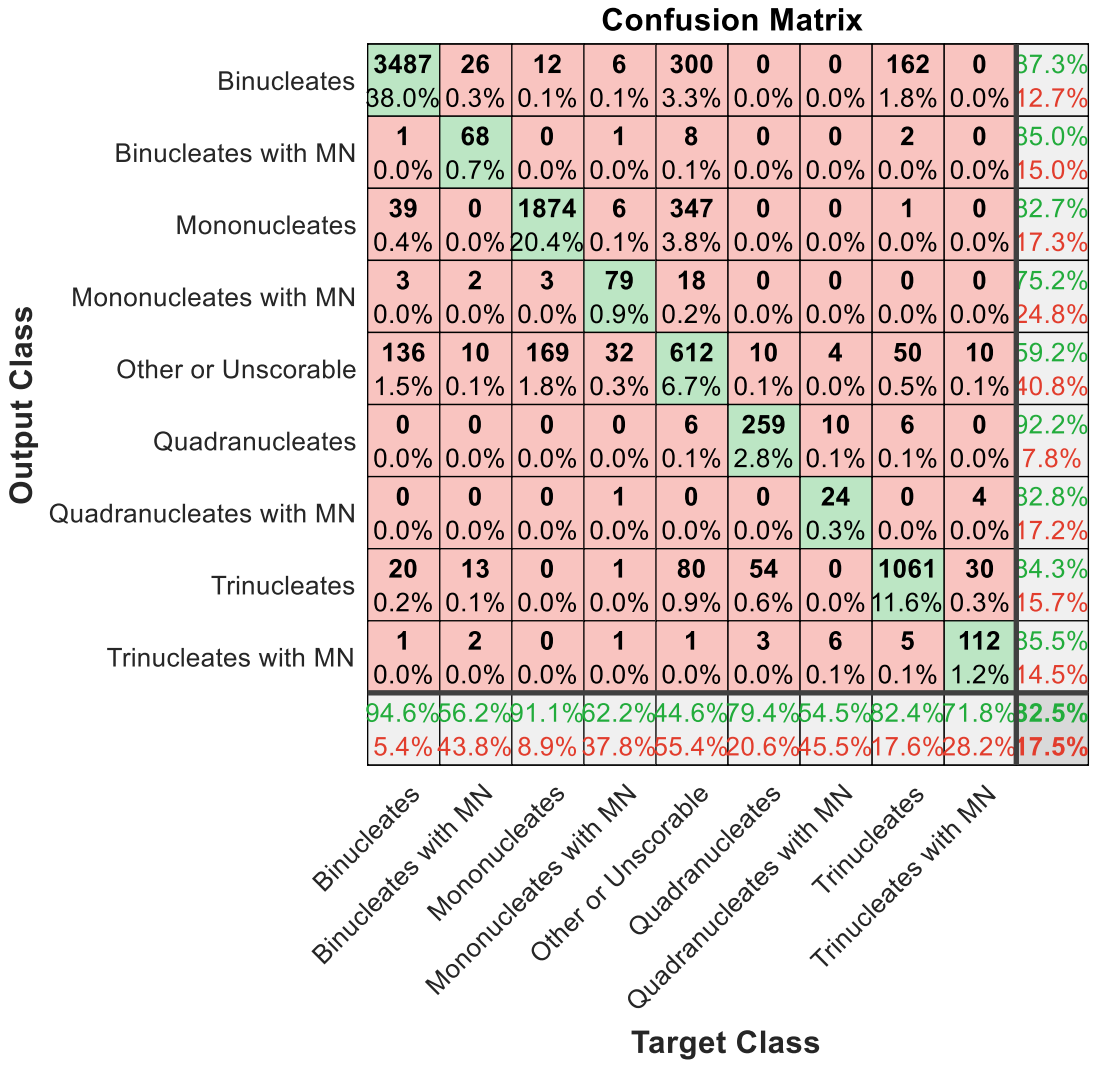

Confusion Matrix
Output Class
Target Class
Binucleates
Binucleates with MN
Mononucleates
Mononucleates with MN
Quadranucleates
Quadranucleates with MN
Trinucleates
Trinucleates with MN
Other or Unscorable
3247 17 55 6 0 0 7 0 0 97.4%
43.0% 0.2% 0.7% 0.1% 0.0% 0.0% 0.1% 0.0% 0.0% 2.6%
3 26 0 0 0 0 0 0 0 89.7%
0.0% 0.3% 0.0% 0.0% 0.0% 0.0% 0.0% 0.0% 0.0% 10.3%
20 0 2118 12 0 0 1 0 0 98.5%
0.3% 0.0% 28.0% 0.2% 0.0% 0.0% 0.0% 0.0% 0.0% 1.5%
0 19 10 55 0 0 1 3 0 62.5%
0.0% 0.3% 0.1% 0.7% 0.0% 0.0% 0.0% 0.0% 0.0% 37.5%
0 3 0 0 335 6 135 5 0 69.2%
0.0% 0.0% 0.0% 0.0% 4.4% 0.1% 1.8% 0.1% 0.0% 30.8%
0 0 0 0 3 7 2 10 0 31.8%
0.0% 0.0% 0.0% 0.0% 0.0% 0.1% 0.0% 0.1% 0.0% 68.2%
715 48 0 1 7 1 265 6 0 25.4%
9.5% 0.6% 0.0% 0.0% 0.1% 0.0% 3.5% 0.1% 0.0% 74.6%
4 30 0 0 1 4 1 14 0 25.9%
0.1% 0.4% 0.0% 0.0% 0.0% 0.1% 0.0% 0.2% 0.0% 74.1%
198 27 37 13 10 0 54 11 0 0.0%
2.6% 0.4% 0.5% 0.2% 0.1% 0.0% 0.7% 0.1% 0.0% 100%
77.5% 15.3% 95.4% 63.2% 94.1% 38.9% 56.9% 28.6% NaN% 80.3%
22.5% 84.7% 4.6% 36.8% 5.9% 61.1% 43.1% 71.4% NaN% 19.7%

B

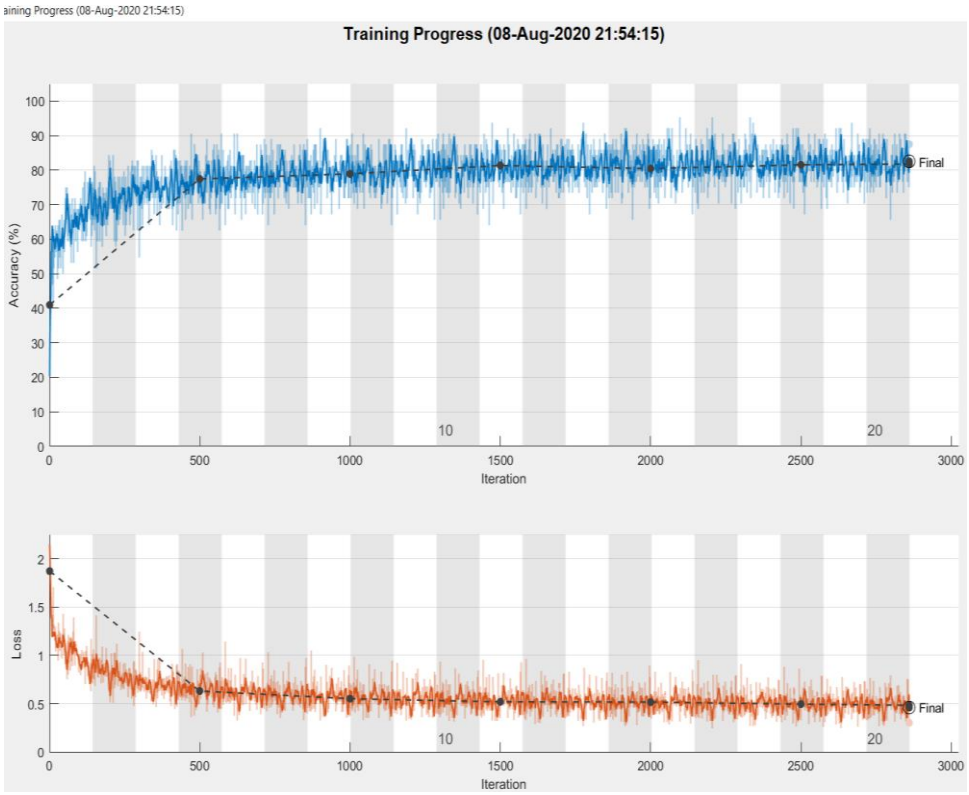

Confusion Matrix
Output Class
Binucleates
Binucleates with MN
Mononucleates
Mononucleates with MN
Other or Unscorable
Quadranucleates
Quadranucleates with MN
Trinucleates
Trinucleates with MN
3459 37.7%
9 0.1%
21 0.2%
12 0.1%
272 3.0%
0 0.0%
0 0.0%
66 0.7%
0 0.0%
90.1% 9.9%
17 0.2%
84 0.9%
0 0.0%
20 0.2%
20 0.2%
0 0.0%
0 0.0%
1 0.0%
2 0.0%
58.3% 41.7%
32 0.3%
0 0.0%
1925 21.0%
0 0.0%
367 4.0%
0 0.0%
0 0.0%
0 0.0%
0 0.0%
82.8% 17.2%
2 0.0%
0 0.0%
3 0.0%
65 0.7%
14 0.2%
0 0.0%
0 0.0%
0 0.0%
0 0.0%
77.4% 22.6%
110 1.2%
10 0.1%
108 1.2%
28 0.3%
563 6.1%
9 0.1%
2 0.0%
50 0.5%
8 0.1%
63.4% 36.6%
0 0.0%
0 0.0%
0 0.0%
0 0.0%
9 0.1%
294 3.2%
22 0.2%
25 0.3%
6 0.1%
82.6% 17.4%
0 0.0%
0 0.0%
0 0.0%
1 0.0%
0 0.0%
0 0.0%
12 0.1%
0 0.0%
0 0.0%
92.3% 7.7%
67 0.7%
15 0.2%
1 0.0%
0 0.0%
124 1.4%
21 0.2%
0 0.0%
1142 12.4%
30 0.3%
81.6% 18.4%
0 0.0%
3 0.0%
0 0.0%
1 0.0%
3 0.0%
2 0.0%
8 0.1%
3 0.0%
110 1.2%
84.6% 15.4%
93.8% 6.2%
69.4% 30.6%
93.5% 6.5%
51.2% 48.8%
41.0% 59.0%
90.2% 9.8%
27.3% 72.7%
88.7% 11.3%
70.5% 29.5%
83.4% 16.6%
Binucleates
Binucleates with MN
Mononucleates
Mononucleates with MN
Other or Unscorable
Quadranucleates
Quadranucleates with MN
Trinucleates
Trinucleates with MN
Target Class

C

I


Confusion Matrix
Output Class
Target Class
Binucleates
Binucleates with MN
Mononucleates
Mononucleates with MN
Other or Unscorable
Quadranucleates
Quadranucleates with MN
Trinucleates
Trinucleates with MN
3487 26 12 6 300 0 0 162 0 37.3%
38.0% 0.3% 0.1% 0.1% 3.3% 0.0% 0.0% 1.8% 0.0% 12.7%
1 68 0 1 8 0 0 2 0 35.0%
0.0% 0.7% 0.0% 0.0% 0.1% 0.0% 0.0% 0.0% 0.0% 15.0%
39 0 1874 6 347 0 0 1 0 32.7%
0.4% 0.0% 20.4% 0.1% 3.8% 0.0% 0.0% 0.0% 0.0% 17.3%
3 2 3 79 18 0 0 0 0 75.2%
0.0% 0.0% 0.0% 0.9% 0.2% 0.0% 0.0% 0.0% 0.0% 24.8%
136 10 169 32 612 10 4 50 10 59.2%
1.5% 0.1% 1.8% 0.3% 6.7% 0.1% 0.0% 0.5% 0.1% 40.8%
0 0 0 0 6 259 10 6 0 92.2%
0.0% 0.0% 0.0% 0.0% 0.1% 2.8% 0.1% 0.1% 0.0% 7.8%
0 0 0 1 0 0 24 0 4 82.8%
0.0% 0.0% 0.0% 0.0% 0.0% 0.0% 0.3% 0.0% 0.0% 17.2%
20 13 0 1 80 54 0 1061 30 84.3%
0.2% 0.1% 0.0% 0.0% 0.9% 0.6% 0.0% 11.6% 0.3% 15.7%
1 2 0 1 1 3 6 5 112 85.5%
0.0% 0.0% 0.0% 0.0% 0.0% 0.0% 0.1% 0.1% 1.2% 14.5%
94.6% 56.2% 91.1% 62.2% 44.6% 79.4% 54.5% 82.4% 71.8% 82.5%
5.4% 43.8% 8.9% 37.8% 55.4% 20.6% 45.5% 17.6% 28.2% 17.5%

II



III

Results

Validation accuracy:	82.55%
Training finished:	Reached final iteration

Training Time

Start time:	08-Aug-2020 21:54:15
Elapsed time:	36 min 44 sec

Training Cycle

Epoch:	20 of 20
Iteration:	2860 of 2860
Iterations per epoch:	143
Maximum iterations:	2860

Validation

Frequency:	500 iterations

Other Information

Hardware resource:	Multiple GPUs
Learning rate schedule:	Piecewise
Learning rate:	0.001

Figure. 21. Confusion matrices produced post neural network creation using MatLab®. Training on the 'Cambridgef' dataset and validation on the 'Cardiff' data set. Overall accuracies shown as well as accuracies per individual subgroups.

A) Neural network formed after using the first updated ground truth and using a 3-channel approach of Brightfield, Fluorescence, Fluorescence, Network 6 produced this confusion matrix.

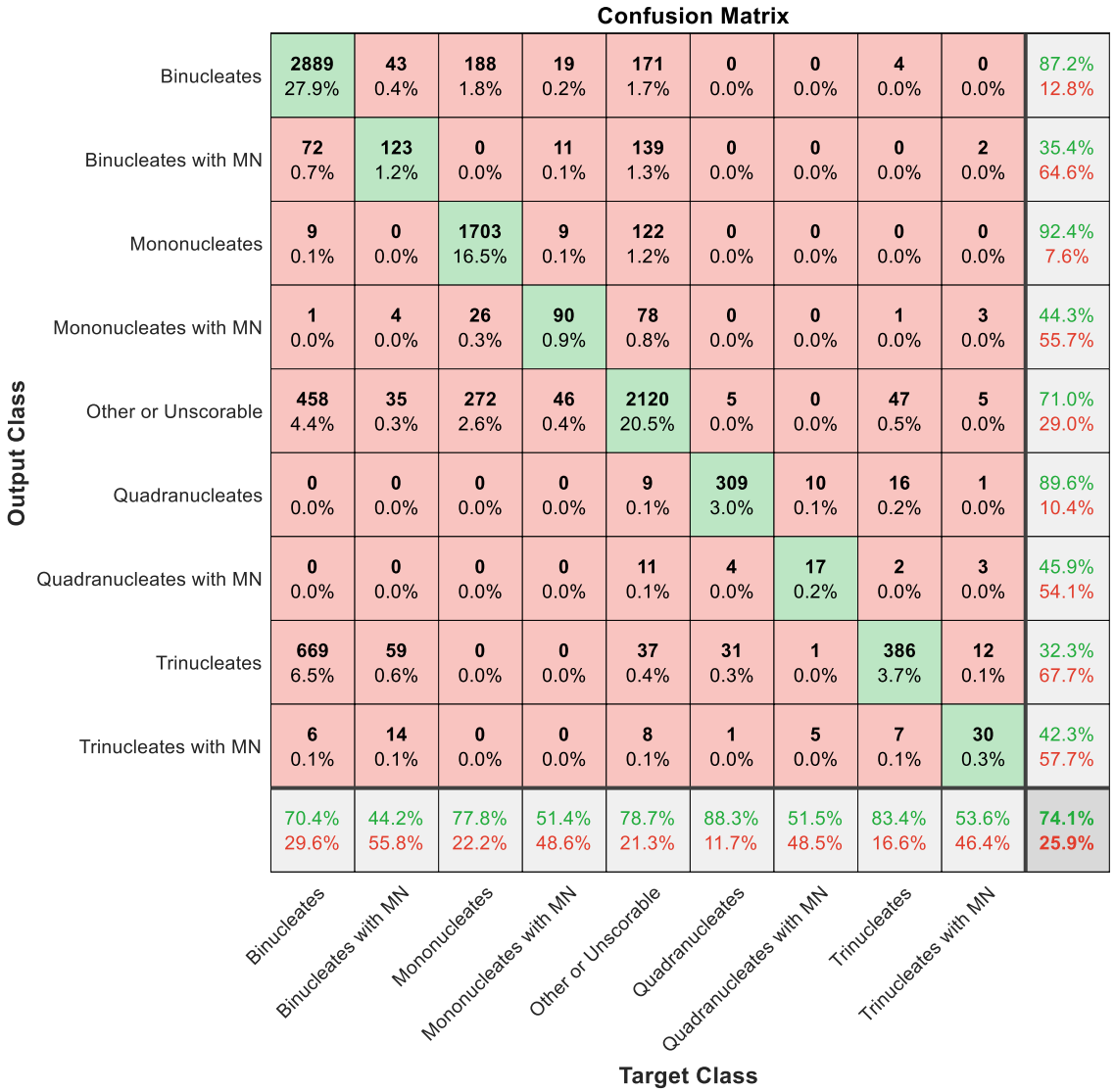
B) Neural network formed after using the first updated ground truth and the 2-channel approach of: Brightfield and Fluorescence. Network 7 produced this confusion matrix.

C) i) Neural network formed post Cardiff and Cambridge ground truth updates and the 2-channel approach of: Brightfield and fluorescence (1, 11). Network E produced this network

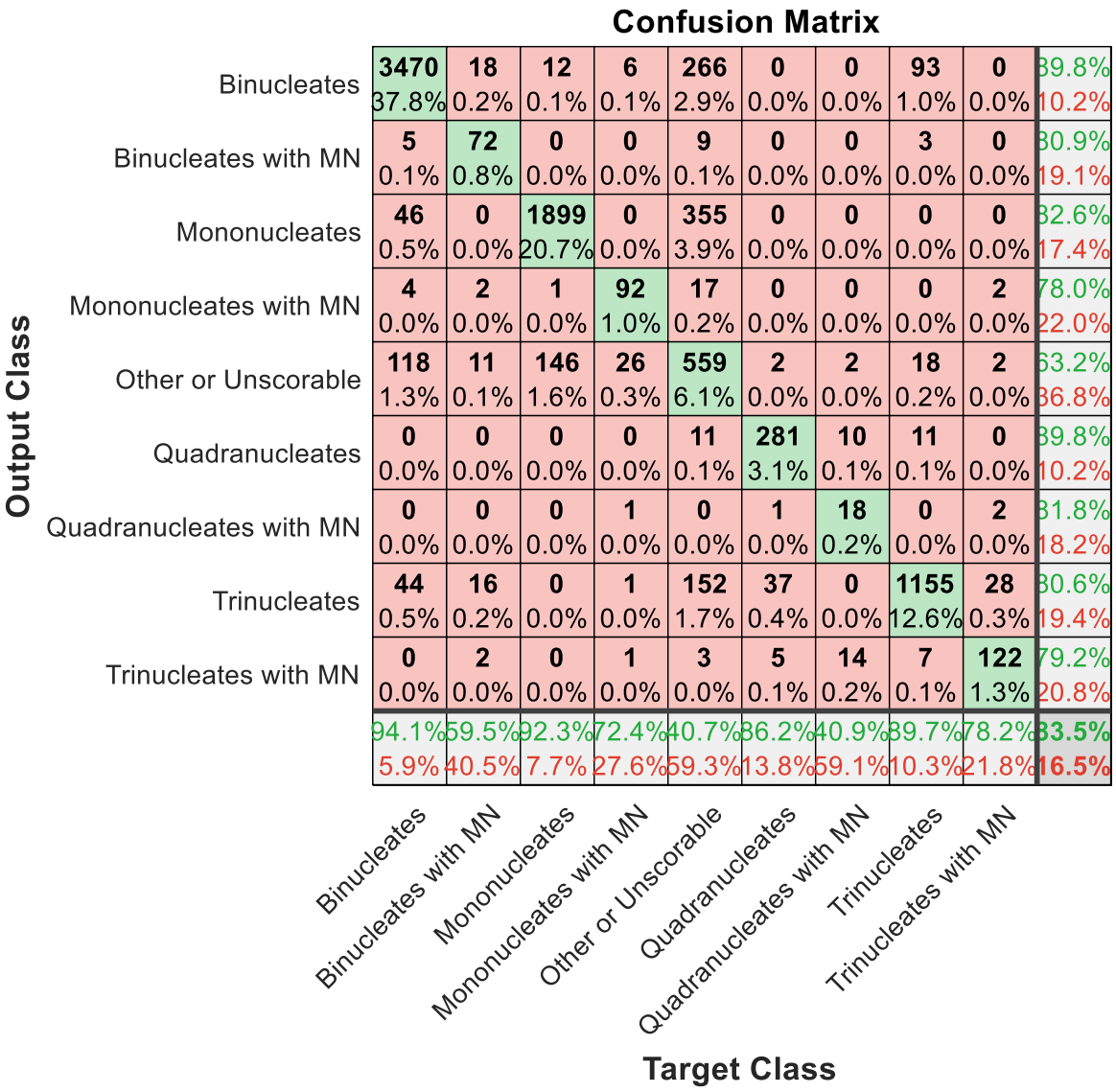
ii) Figure showing the training development of Network E up to 20 epochs. Accuracy and error rate are both shown.

iii) The results section after completing a network run, showing the accuracy rate of the network produced after completing 20 epoch cycles, other key statistics are also shown, such as iterations taken and time elapsed.

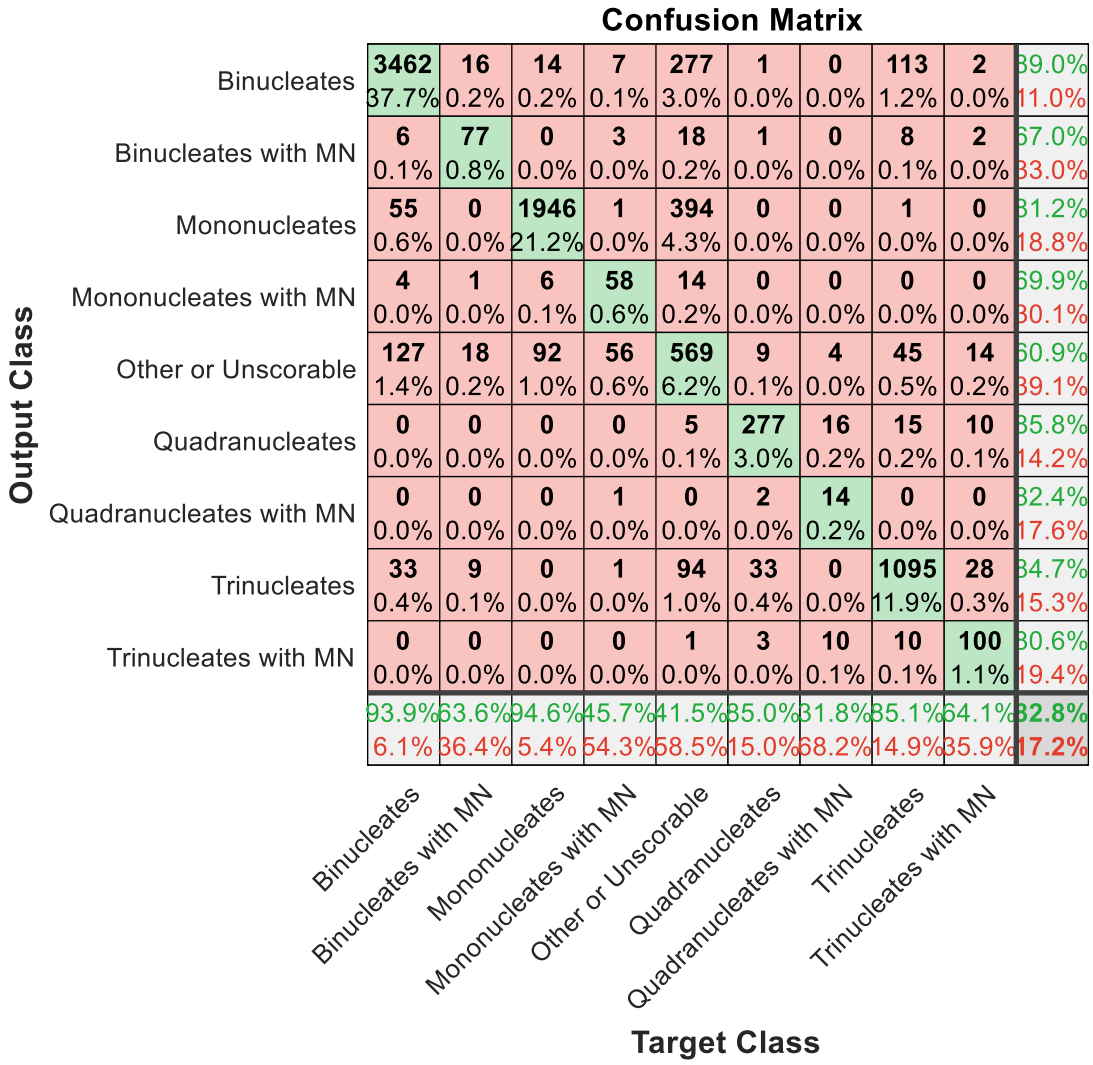
A


Confusion Matrix
Output Class
Target Class
Binucleates
Binucleates with MN
Mononucleates
Mononucleates with MN
Other or Unscorable
Quadranucleates
Quadranucleates with MN
Trinucleates
Trinucleates with MN
2889 27.9%
43 0.4%
188 1.8%
19 0.2%
171 1.7%
0 0.0%
0 0.0%
4 0.0%
0 0.0%
87.2% 12.8%
72 0.7%
123 1.2%
0 0.0%
11 0.1%
139 1.3%
0 0.0%
0 0.0%
0 0.0%
2 0.0%
35.4% 64.6%
9 0.1%
0 0.0%
1703 16.5%
9 0.1%
122 1.2%
0 0.0%
0 0.0%
0 0.0%
0 0.0%
92.4% 7.6%
1 0.0%
4 0.0%
26 0.3%
90 0.9%
78 0.8%
0 0.0%
0 0.0%
1 0.0%
3 0.0%
44.3% 55.7%
458 4.4%
35 0.3%
272 2.6%
46 0.4%
2120 20.5%
5 0.0%
0 0.0%
47 0.5%
5 0.0%
71.0% 29.0%
0 0.0%
0 0.0%
0 0.0%
0 0.0%
9 0.1%
309 3.0%
10 0.1%
16 0.2%
1 0.0%
89.6% 10.4%
0 0.0%
0 0.0%
0 0.0%
0 0.0%
11 0.1%
4 0.0%
17 0.2%
2 0.0%
3 0.0%
45.9% 54.1%
669 6.5%
59 0.6%
0 0.0%
0 0.0%
37 0.4%
31 0.3%
1 0.0%
386 3.7%
12 0.1%
32.3% 67.7%
6 0.1%
14 0.1%
0 0.0%
0 0.0%
8 0.1%
1 0.0%
5 0.0%
7 0.1%
30 0.3%
42.3% 57.7%
70.4% 29.6%
44.2% 55.8%
77.8% 22.2%
51.4% 48.6%
78.7% 21.3%
88.3% 11.7%
51.5% 48.5%
83.4% 16.6%
53.6% 46.4%
74.1% 25.9%

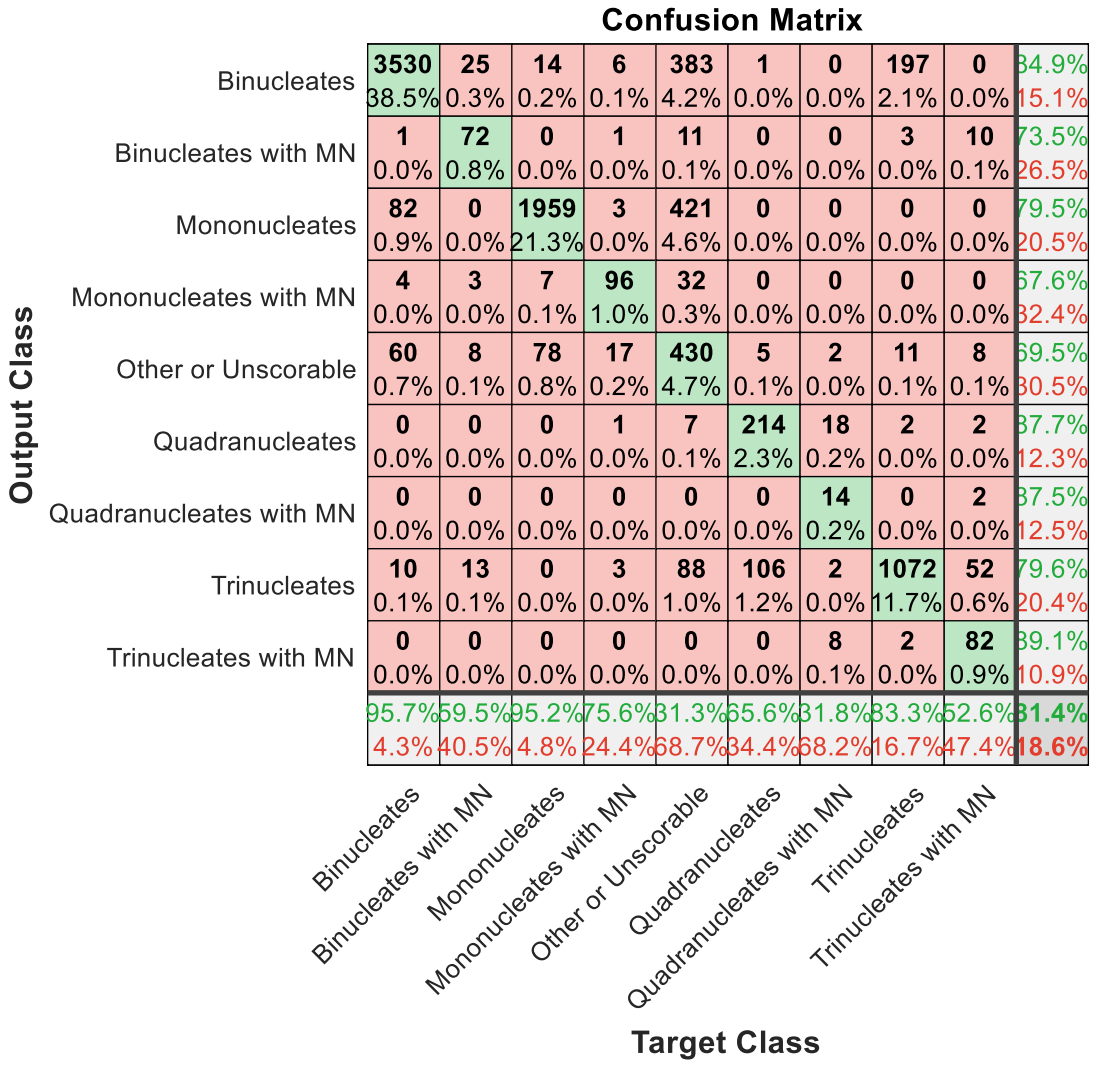
B


Confusion Matrix
Output Class
Target Class
Binucleates
Binucleates with MN
Mononucleates
Mononucleates with MN
Other or Unscorable
Quadranucleates
Quadranucleates with MN
Trinucleates
Trinucleates with MN
3470 37.8% | 18 0.2% | 12 0.1% | 6 0.1% | 266 2.9% | 0 0.0% | 0 0.0% | 93 1.0% | 0 0.0% | 39.8% 10.2%
5 0.1% | 72 0.8% | 0 0.0% | 0 0.0% | 9 0.1% | 0 0.0% | 0 0.0% | 3 0.0% | 0 0.0% | 30.9% 19.1%
46 0.5% | 0 0.0% | 1899 20.7% | 0 0.0% | 355 3.9% | 0 0.0% | 0 0.0% | 0 0.0% | 0 0.0% | 82.6% 17.4%
4 0.0% | 2 0.0% | 1 0.0% | 92 1.0% | 17 0.2% | 0 0.0% | 0 0.0% | 0 0.0% | 2 0.0% | 78.0% 22.0%
118 1.3% | 11 0.1% | 146 1.6% | 26 0.3% | 559 6.1% | 2 0.0% | 2 0.0% | 18 0.2% | 2 0.0% | 63.2% 36.8%
0 0.0% | 0 0.0% | 0 0.0% | 0 0.0% | 11 0.1% | 281 3.1% | 10 0.1% | 11 0.1% | 0 0.0% | 89.8% 10.2%
0 0.0% | 0 0.0% | 0 0.0% | 1 0.0% | 0 0.0% | 1 0.0% | 18 0.2% | 0 0.0% | 2 0.0% | 81.8% 18.2%
44 0.5% | 16 0.2% | 0 0.0% | 1 0.0% | 152 1.7% | 37 0.4% | 0 0.0% | 1155 12.6% | 28 0.3% | 80.6% 19.4%
0 0.0% | 2 0.0% | 0 0.0% | 1 0.0% | 3 0.0% | 5 0.1% | 14 0.2% | 7 0.1% | 122 1.3% | 79.2% 20.8%
94.1% 5.9% | 59.5% 40.5% | 92.3% 7.7% | 72.4% 27.6% | 40.7% 59.3% | 86.2% 13.8% | 40.9% 59.1% | 89.7% 10.3% | 78.2% 21.8% | 83.5% 16.5%

C

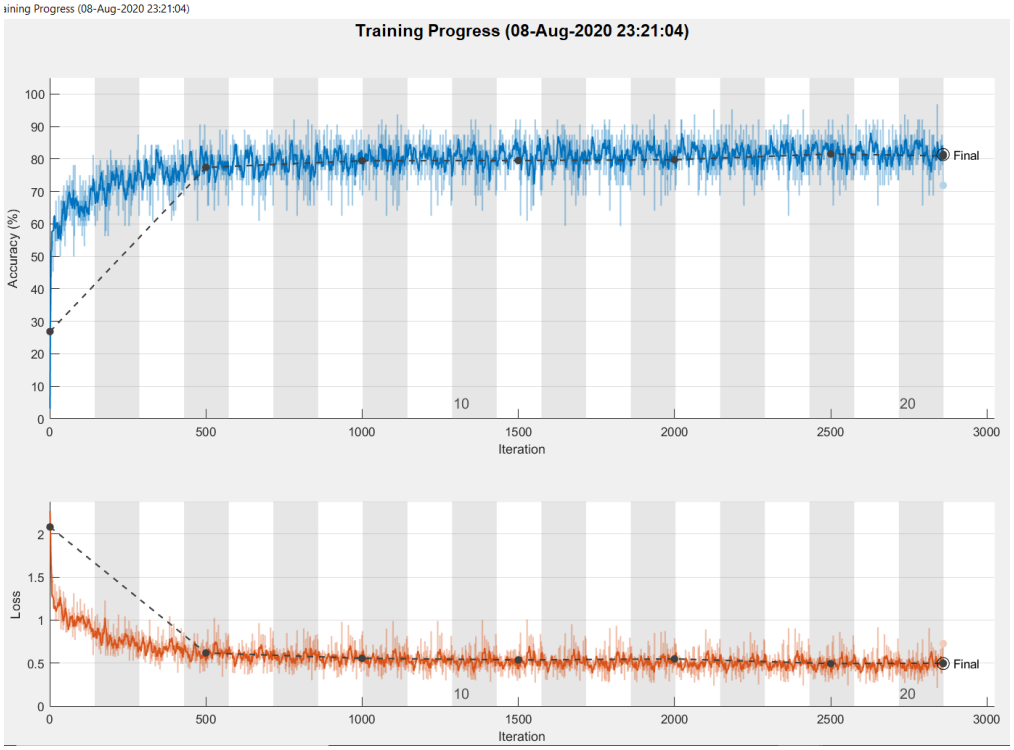

Confusion Matrix
Output Class
Binucleates
Binucleates with MN
Mononucleates
Mononucleates with MN
Other or Unscorable
Quadranucleates
Quadranucleates with MN
Trinucleates
Trinucleates with MN
3462 37.7% | 16 0.2% | 14 0.2% | 7 0.1% | 277 3.0% | 1 0.0% | 0 0.0% | 113 1.2% | 2 0.0% | 89.0% 11.0%
6 0.1% | 77 0.8% | 0 0.0% | 3 0.0% | 18 0.2% | 1 0.0% | 0 0.0% | 8 0.1% | 2 0.0% | 67.0% 33.0%
55 0.6% | 0 0.0% | 1946 21.2% | 1 0.0% | 394 4.3% | 0 0.0% | 0 0.0% | 1 0.0% | 0 0.0% | 81.2% 18.8%
4 0.0% | 1 0.0% | 6 0.1% | 58 0.6% | 14 0.2% | 0 0.0% | 0 0.0% | 0 0.0% | 0 0.0% | 69.9% 30.1%
127 1.4% | 18 0.2% | 92 1.0% | 56 0.6% | 569 6.2% | 9 0.1% | 4 0.0% | 45 0.5% | 14 0.2% | 60.9% 39.1%
0 0.0% | 0 0.0% | 0 0.0% | 0 0.0% | 5 0.1% | 277 3.0% | 16 0.2% | 15 0.2% | 10 0.1% | 85.8% 14.2%
0 0.0% | 0 0.0% | 0 0.0% | 1 0.0% | 0 0.0% | 2 0.0% | 14 0.2% | 0 0.0% | 0 0.0% | 82.4% 17.6%
33 0.4% | 9 0.1% | 0 0.0% | 1 0.0% | 94 1.0% | 33 0.4% | 0 0.0% | 1095 11.9% | 28 0.3% | 84.7% 15.3%
0 0.0% | 0 0.0% | 0 0.0% | 0 0.0% | 1 0.0% | 3 0.0% | 10 0.1% | 10 0.1% | 100 1.1% | 80.6% 19.4%
93.9% 6.1% | 63.6% 36.4% | 94.6% 5.4% | 45.7% 54.3% | 41.5% 58.5% | 85.0% 15.0% | 31.8% 68.2% | 85.1% 14.9% | 64.1% 35.9% | 82.8% 17.2%
Binucleates
Binucleates with MN
Mononucleates
Mononucleates with MN
Other or Unscorable
Quadranucleates
Quadranucleates with MN
Trinucleates
Trinucleates with MN
Target Class

D I


Confusion Matrix
Output Class
Target Class
Binucleates
Binucleates with MN
Mononucleates
Mononucleates with MN
Other or Unscorable
Quadranucleates
Quadranucleates with MN
Trinucleates
Trinucleates with MN
3530 38.5% | 25 0.3% | 14 0.2% | 6 0.1% | 383 4.2% | 1 0.0% | 0 0.0% | 197 2.1% | 0 0.0% | 34.9% 15.1%
1 0.0% | 72 0.8% | 0 0.0% | 1 0.0% | 11 0.1% | 0 0.0% | 0 0.0% | 3 0.0% | 10 0.1% | 73.5% 26.5%
82 0.9% | 0 0.0% | 1959 21.3% | 3 0.0% | 421 4.6% | 0 0.0% | 0 0.0% | 0 0.0% | 0 0.0% | 79.5% 20.5%
4 0.0% | 3 0.0% | 7 0.1% | 96 1.0% | 32 0.3% | 0 0.0% | 0 0.0% | 0 0.0% | 0 0.0% | 67.6% 32.4%
60 0.7% | 8 0.1% | 78 0.8% | 17 0.2% | 430 4.7% | 5 0.1% | 2 0.0% | 11 0.1% | 8 0.1% | 69.5% 30.5%
0 0.0% | 0 0.0% | 0 0.0% | 1 0.0% | 7 0.1% | 214 2.3% | 18 0.2% | 2 0.0% | 2 0.0% | 87.7% 12.3%
0 0.0% | 0 0.0% | 0 0.0% | 0 0.0% | 0 0.0% | 0 0.0% | 14 0.2% | 0 0.0% | 2 0.0% | 87.5% 12.5%
10 0.1% | 13 0.1% | 0 0.0% | 3 0.0% | 88 1.0% | 106 1.2% | 2 0.0% | 1072 11.7% | 52 0.6% | 79.6% 20.4%
0 0.0% | 0 0.0% | 0 0.0% | 0 0.0% | 0 0.0% | 0 0.0% | 8 0.1% | 2 0.0% | 82 0.9% | 89.1% 10.9%
95.7% 4.3% | 59.5% 40.5% | 95.2% 4.8% | 75.6% 24.4% | 31.3% 68.7% | 65.6% 34.4% | 31.8% 68.2% | 83.3% 16.7% | 52.6% 47.4% | 81.4% 18.6%

II

ining Progress (08-Aug-2020 23:21:04)



III

Results

Validation accuracy:	91.38%
Training finished:	Reached final iteration

Training Time

Start time:	08-Aug-2020 23:21:04
Elapsed time:	36 min 55 sec

Training Cycle

Epoch:	20 of 20
Iteration:	2860 of 2860
Iterations per epoch:	143
Maximum iterations:	2860

Validation

Frequency:	500 iterations

Other Information

Hardware resource:	Multiple GPUs
Learning rate schedule:	Piecewise
Learning rate:	0.001

Figure. 22 Confusion matrices produced post neural network creation using MatLab®. Training on the 'Cambridge' dataset and validation also on the 'Cambridge' data set. Overall accuracies shown as well as accuracies per individual subgroups.

A) Neural network formed after using the initial ground truth population and a 3-channel approach of: Brightfield, Fluorescence, Fluorescence. Network 8 produced this confusion matrix.

B) Neural network formed after using the first updated ground truth and using a 3-channel approach of Brightfield, Fluorescence, Fluorescence, Network 9 produced this confusion matrix.

C) Neural network formed after using the first updated ground truth and the 2-channel approach of: Brightfield and Fluorescence. Network 10 produced this confusion matrix.

D) i) Neural network formed post Cardiff and Cambridge ground truth updates and the 2-channel approach of: Brightfield and fluorescence (1, 11). Network F produced this network

ii) Figure showing the training development of Network E up to 20 epochs. Accuracy and error rate are both shown.

The results section after completing a network run, showing the accuracy rate of the network produced after completing 20 epoch cycles, other key statistics are also shown, such as iterations taken and time elapsed.

Glossary

Adobe Bridge®: Software commonly used by photographers which allows for the ground truth to be created in this case and cellular images analysed on an individual basis and grouped accordingly.

Batch size: the quantity of samples which will be put through the system at a time

Carbendazim: An aneugenic agent commonly used in dose response analysis due to its well-studied mode of action.

Confusion Matrix: A figure produced after assessing a Neural Network on a dataset. Accuracy levels are shown so that the user can relay the information going forwards on what category/categories require improvement. Normally produced when attempting to increase the accuracy of the network or when carrying out a dose response.

Cytochalasin-B: A spindle poison commonly used in the MN assay as it disrupts cytokinesis and therefore not allowing the cells to divide their cytoplasms, whilst nuclear division continues, giving the cells their binucleated cell appearance.

Deep Learning: An artificial intelligence approach which mimics the workings of a human brain by using neural networks to recognise patterns from training data sets.

Epoch: A complete training cycle on an entire data size. Shown during the neural network creation, optimal level required. Too high an epoch frequency can lead to the error rate increasing. Too low an epoch accuracy can cause the epoch count to not reach optimal accuracy levels.

Ground truth: A set of images which have been manually assessed by the user and confirmed to be displaying a specific phenotype or shape required. The Network is trained using this dataset.

Imaging Flow Cytometry: A machine used which carries out analysis on samples by suspending samples in fluid and analysed by the machine following excitation of fluorescent markers by light, causing the light to be scattered and high throughput analysis to be undertaken. The imaging flow cytometer allows the user to click on individual cellular images for analysis and provides extra confidence into the results and added sample integrity.

Micronucleus: A smaller than normal nuclei, $1/3^{rd}$ to $1/16^{th}$ the diameter of a regular nucleus. Occur at a resting level in healthy cells, but levels are elevated following exposure to agents causing chromosomal damage and this can be used to calculate a dose response using the MN assay.

Neural Network: A series of algorithms linked to one another, much how neurones are in the brain, working in tandem to recognise patterns and efficiently analyse data.

Training Set: The Ground Truth dataset used in order to teach the neural network how to identify a specific pattern.

Validation Set: The images used to assess the Neural Network based on what the training set has taught it. This should not be the same as the training set as can result in false positives. If using the same dataset for training and validation, then split this dataset into two distinct groups, one to train the dataset, one to validate the dataset.

Bibliography

References

Amnis imaging flow cytometry. (2016). Inspire for FlowSight Software. User's Manual.

Aranda A, Bezunartea J, Casales E *et al.* (2014). A quick and efficient method to generate mammalian stable cell lines based on a novel inducible alphavirus DNA/RNA layered system. *Cell. Mol. Life Sci:* **71,** 4637–4651.

Auclair C, Gouyette A, Levy A, et al. (1990). Clastogenic inosine nucleotides as components of the chromosome breakage factor in scleroderma patients. Arch Biochem Biophys.**278**:238–4.

BD Biosciences. (2020). Hoeschst 33342 Solution. Date accessed: 25/08/20. https://www.bdbiosciences.com/eu/applications/research/cell-and-tissue-microscopy/kits-buffers-and-stains/hoechst-33342-solution/p/561908 .

BD Pharmingen™. (2017). Technical Data Sheet, DRAQ5™. BD Biosciences, 564903 Rev.4. Date accessed: 25/08/20, https://www.bdbiosciences.com/ds/pm/tds/564903.pdf.

Biostatus. (2017). DRAQ5™. Date accessed: 25/08/20, http://www.biostatus.com/DRAQ5/.

Bonner W. M, Redon C. E, Dickey J. S, Nakamura A. J, Sedelnikova O. A, Solier S, Pommier Y. (2008). Gamma H2AX and cancer. *Nat Rev Cancer; **8**(12):957-67.*

Brücher B.L and Jamall I.S. (2016). Somatic Mutation Theory- Why it's Wrong for Most Cancers. Cell Physiol Biochem **38(5)**: 1663-80.

Bui D.T and Li J.J. (2019). DNA re-replication is susceptible to nucleotide level mutagenesis. *Genetics* **10**.1534.

Chapman KE, Thomas AD, Wills JW, Pfuhler S, Doak SH and Jenkins GJ (2014) Automation and validation of micronucleus detection in the 3D EpiDerm human reconstructed skin assay and correlation with 2D dose responses. *Mutagenesis* **29**(**3**):165–175.

Chatterjee N and Walker G.C. (2017) Mechanisms of DNA damage, repair and mutagenesis. *Environ Mol Mutagen* **58(5):** 235-263.

Christmann M, Verbeek B, Roos W.P, Kaina B. (2011). O^6-Methylguanine-DNA methyltransferase (MGMT) in normal tissues and tumours: Enzyme activity, promoter methylation and immunohistochemistry. Biochimica et Biophysica Acta (BBA)- Reviews on Cancer **1816(2):** 179-190.

Copper G.M. (2000). Chapter 14: The Eukaryotic Cell Cycle. *The cell: a molecular approach*, 2nd edition. Sunderland (MA): Sinauer Associates.

Crofton-Sleight C, Doherty A, Ellard S, Parry E.M and Venitt S. (1993). Micronucleus assays using cytochalasin-blocked MCL-5 cells, a proprietary human cell line expressing five human cytochromes P-450 and microsomal epoxide hydrolase. Mutagenesis: **8(4):** 363-372.

Cross Validated. (2020). Batch size in neural network. (Date accessed: 29/08/20).

https://stats.stackexchange.com/questions/153531/what-is-batch-size-in-neural-network

Decordier I, Papine A, Plas G, Roesems S, Vande Loock K, Moreno-Palomo J, Cemeli E, Anderson D, Fucic A, Marcos R, Soussaline F, Kirsch-Volders M. Automated image analysis of cytokinesis-blocked micronuclei: an adapted protocol and a validated scoring procedure for biomonitoring. Mutagenesis. **24(1**):85–93.

Doherty AT, Hayes J, Fellows M, Kirk S, & O'Donovan M. (2011). A rapid, semi-automated method for scoring micronuclei in mononucleated mouse lymphoma cells. *Mutat Res*, *726*(1), 36-41.

Dolan M. E, Young G. S and Pegg A. E. (1986) The effect of O^6 alkylguanine pretreatment on the sensitivity of human colon tumor cells to the cytotoxic effects of chloroethylating agents. Cancer Res.. **46**: 4500-4504.

Dolan M. E, Moschel R. C and Pegg A. E. (1990). Depletion of mammalian 06-alkylguanine-DNA alkyltransferase activity by ft-benzylguanine provides a means to evaluate the role of this protein in protection against carcinogenic and therapeutic alkylating agents. Proc. NatI. Acad. Sci. USA, **87**: 5368-5372.

Dolan M.E and Pegg A.E. (1997). O^6-Benzylguanine and its Role in Chemotherapy. Clinical Cancer Research, **3:** 837-847.

Emerit I, Khan SH, Esterbauer H. (1991). Hydroxynonenal, a component of clastogenic factors? *Radic Biol Med.* **10(6**):371-7.

Emerit I, Fabiani JN, Levy A, Ponzio O, Conti M, Brasme B, Bienvenu P, Hatmi M. (1995). Plasma from patients exposed to ischemia reperfusion contains clastogenic factors and stimulates the chemiluminescence response of normal leukocytes. *Free Radic Biol Med. **19(4):**405-15.*

Emerit I. (2007). Clastogenic Factors as Potential Biomarkers of Increased Superoxide Production. *Biomark Insights* **11(2)**: 429-38.

EPA. (2017). Technical Fact Sheet- N-nitroso-dimethylamine (NDMA). United States Environmental Protection Agency, Office of Land and Emergency: **EPA**: 505-F-17-005.

Esteller M, Hamilton S.R, Burger P.C *et al.* (1999). Inactivation of the DNA Repair Gene O^6-Methylguanine-DNA Methyltransferase by promoter Hypermethylation is a Common Event in Primary Human Neolasia. *Cancer Res***:59:793-797**.

Evans H.J. (1977). Molecular mechanisms in the induction of chromosome aberrations. Progress in Genetic toxicology, pp. 57-74.

Fan C-H, Liu W-L, Cao H, Wen C, Chen L, Jiang G. (2013). O^6-methylguanine DNA methyltransferase as a promising target for the treatment of temozolomide-resistant gliomas. *Cell death Dis*: **4,** e876.

Fenech M. (1997). The advantages and disadvantages of the cytokinesis-block micronucleus method. *Mutat. Res.* **1;392(1-2**):11-8.

Fenech M. (2000). The *in vitro* micronucleus technique. *Mutat. Res.* **455(1–2):**81–95.

Fenech M, Chang W.P, Kirsch-Volders M, Holland N, Bonassi S and Zeiger E. (2003). HUMN project: Detailed description of the scoring criteria for the cytokinesis-block micronucleus assay using isolated human lymphocyte cultures. *Mutat Res* **534:** 65-75.

Fenech M. (2007). Cytokinesis-block micronucleus cytome assay. *Nat Protoc***2:**1084- 1104.

Fenech M, Kirsch-Volders M, Natarajan A. T, Surralles J, Crot J. W, Parry J *et al* (2011). Molecular mechanisms of micronucleus, nucleoplasmic bridge and nuclear bud formation in mammalian and human cells. Mutagenesis **26** (**1):** 125-132.

Ganai, R. A., and Johansson E. (2016). DNA Replication- A Matter of Fidelity. Mol Cell **62**:745-755.

Hanahan D and Weinberg R.A. (2000). Hallmarks of cancer. Cell; 100(1): 57-70.
Hanahan D and Weinberg R.A. (2011). Hallmarks of cancer: the next generation. Cell, 4;144(5): 646-74.
Haxhiraj Q, Verma J.R, Johnson G.E. (2018). Assessing chemically induced DNA damage in human lymphoblastoid cells, *in vitro,* using FlowSight® imaging flow cytometry. Swansea University PM304 Research Project.

Heddle JA. (1973). A rapid in vivo test for chromosome damage. *Mutat. Res*. (**18)** 187–192.

Hintzsche H, Hemmann U, Poth A, Utesch D, Lott J, Stopper H. (2017). Fate of micronuclei and micronucleated cells. Mutation Research - Reviews in *Mutat. Res:***771**:85-98.

ICHS2(R1). (2012). International Conference on Harmonisation; guidance on S2(R1) Genotoxicity Testing and Data Interpretation for Pharmaceuticals intended for Human Use; availability. Notice. *Fed Regist, 77*(110): 33748-33749.

Jardim D. L, Groves E. S, Breitfeld P. P, Kurzrock R. (2017). Factors associated with failure of oncology drugs in late-stage clinical development: A systematic review. Cancer treat Rev **52**: 12-21.

Johnson G.E, Slob W, Doak S.H, Fellows M.D, Gollapudi B.B, Heflich R.H, *et al.*, (2014). New approaches to advance the use of genetic toxicology analyses for human health risk assessment. *Toxicol Res* 4**(3)**: 667-676.

Kaina B, Margison GP, Christmann M. (2010). Targeting O . (6)-methylguanine-DNA methyltransferase with specific inhibitors as a strategy in cancer therapy. *Cell Mol Life Sci*; **67**: 3663–3681.

Kunkel, T. A.. (2009). Evolving views of DNA replication (in)fidelity. Cold Spring Harb. *Symp.Quant. Biol*. **74**:91-101

Lane DP. (1992). Cancer. P53, guardian of the genome. Nature **358(6381**): 15-6.

Liteplo R.G, Meek M.E, Windle W. (2002). Concise International Chemical Assessment Document 38. N-Nitrosodimethylamine, CID=6124. World Helath Organisation.

Lorge E, Moore MM, Clemens J, O'Donovan M, Fellows MD, Honma M et al. (2016). Standardised cell sources and reccomendatios for good cell culture practices in genotoxicity testing. Mutation Research/Genetics Toxicology and Environmental Mutagenesis (**809**): 1-15.

Lovell D. P, Fellows M, Marchetti F, Christiansesn J, Elhajouji A, Hashimoto *et al*. (2018). Anlaysis of negative historical control group data from the *in vitro* micronucleus assay using Tk6 cells. *Mutat Res*: **825**: 40-50.

Lijinsky W. (1999). N-Nitroso compounds in the diet. *Mutat Res* **443**: 129-138.

Matlab DL. (2020). Train Object Detector using R-CNN Deep Learning. Deep Learning Toolbox. MathWorks, MatLab 2020a.

MatLab DL. (2020). Train Deep Learning Network to Classify New Images. Deep learning toolbox. Mathworks, Matlab 2020a.

Matlab St. (2020). Statistics and Machine Learning Toolbox. MathWorks, MatLab 2020a.

Mortelmans K, Zeiger E. (2000). The Ames Salmonella/microsome mutagenicity assay. *Mutat Res* ;**455(1-2**): 29-60.

Natarajan A.T, Obe G. (1982). Mutagenicity testing with cultured mammalian cells: cytogenetic assays. Mutagenicity: New Horizons in Genetic Toxicology, pp. 171-213.

National Cancer Institute. (2018). National Institute of Health. (Date accessed: 18/07/20). https://www.cancer.gov/about-cancer/understanding/statistics.

NC3Rs. (2020). The 3Rs. National Centre for the Replacement, Refinement & Reduction of Animals in Research. (Date accessed: 25/09/20).

https://www.nc3rs.org.uk/the-3rs .

OECD. (2011). QASR Toolbox User Manual. Strategies for grouping chemicals to fill data gaps to assess genetic toxicity and genotoxic carcinogenicity. Available from: http://www.oecd.org/chemicalsafety/risk-assessment/46985336.pdf.

OECD. (2013). QASR Toolbox User Manual. Strategies for grouping chemicals to fill data gaps to assess genetic toxicity and genotoxic carcinogenicity. Available from: https://www.oecd.org/chemicalsafety/risk-assessment/genetic%20toxicity.pdf .

OECD. (2014) OECD TG487 Guideline for the Testing of Chemicals, In Vitro Mammalian Cell Micronucleus Test. Organisation for Economic Cooperation and OECD, Paris.

Ochs K, Kaina B. (2000). Apoptosis induced by DNA damage O^6-methylguanine is Bcl-2 and caspase-0/3 regulated and Fas/caspase-8 independent. *Cancer Res*, **60:**5815-5824.

Olive P and Banáth J. (2006). The comet assay: a method to measure DNA damage in individual cells. *Nat Protoc* **1:** 23-29.

Phillips D.H, Arlt V.M. (2009). Genotoxicity: damage to DNA and its consequences. *EXS*,**99**: 87-110.

Rodrigues M. A, Beaton-Green L. A, Kutzner B. C, Wilkins R.C. (2014). Automated analysis of the cytokinesis-block micronucleus assay for radiation

biodosimetry using imaging flow cytometry. *Radiat. Environ. Biophys., 53, 273-282.*

Rodrigues M.A, Probst C.E, Beaton-Green L.A and Wilkins R.C. (2016). The effect of an optimized imaging flow cytometry analysis template on sample throughput in the reduced culture cytokinesis-block micronucleus assay. *Radiat Prot Dosim*: **172**: 223-229.

Rodrigues M.A, Probst C.E, Beaton-Green L.A and Wilkins R.C. (2016). Optimized automated data analysis for the cytokinesis-block micronucleus assay using imaging flow cytometry for high throughput radiation biodosimetry. *Cytom A*: **89**:653-662.

Rodrigues M. A. (2018) Automation of the *in vitro* micronucleus assay using the Imagestream® imaging flow cytometer. *Cytom A*, 93: 7.

Roos W.P, Thomas A.D, Kaina B. (2016). DNA damage and the balance between survival and death in cancer biology. Nat Rev Cancer **16**, 20-33.

Rosefort, C, Fauth E, Zankl H. (2004). Micronuclei induced by aneugens and clastogens in monocytes and binucleate cells using the cytokinesis block assay. Mutagenesis **19(4):** 277-284.

Rudd N. L, Hoar D. I. (1991). Kietochore analysis of micronuclei allows insights into the actions of colcemid and mitomycin C. Mutat Res, **261(1)**: 57-68.

Savage J. R. K. (1993). Update on target theory as applied to chromosomal aberrations. *Env. Mol. Mutagen*., **22**, pp. 198-207.

Schmid, W. (1975) The micronucleus test. *Mutat. Res*. **(31):** 9–15.

Verma J. R, Rees B. J, Wilde E. C, Thornton C. A, Jenkins G. J. S et al. (2017). Evaluation of the automated MicroFlow® and Metafer™ platforms for high-throughput micronucleus scoring and dose response analysis in human lymphoblastoid TK6 cells. *Arch. Toxicol*, **91(7**): 2689-2698.

Verma J. R, Harte D. S. G, Ume-Kulsoom S, Summers H, Thornton C. A et al. (2018). Investigating FlowSight® imaging flow cytometry as a platform to assess

chemically induced micronuclei using human lymphoblastoid cells *in vitro*. *Mutagenesis*, **33, 4:** 283-289.

Warden P. (2017). How many images do you need to train a network. Date accessed: 10/09/2020. https://petewarden.com/2017/12/14/how-many-images-do-you-need-to-train-a-neural-network/

World Health Organization. (2008). Guidelines for Drinking-Water Quality. 3rd edition including 1st and 2nd addenda.

Xu-Welliver and Pegg A.E. (2002). Degradation of the alkylated form of the DNA repair protein, O^6-alkylguanine DNA alkyltransferase. Carcinogenesis, **23**:823-830.

Yamamoto,K.I. and Kikuchi,Y. (1980) A comparison of diameter of MN induced by clastogens and by spindle poisons. *Mutat. Res.* , 71, 127–131.