FACULTY
OF SOCIAL SCIENCES
Charles University

Institute of Political Studies

Department of Security Studies

Master's Thesis

2023 **Antonín Kanát**

Institute of Political Studies

FACULTY
OF SOCIAL SCIENCES
Charles University

Department of Security Studies

ML-OSINT: Classifying Russian Vehicle Losses in Ukraine

Master's Thesis

Author of the Thesis: Antonín Kanát

Study programme: Security Studies

Supervisor: Mgr. Petr Špelda, Ph.D.

Year of the defence: 2023

Declaration

1. I hereby declare that I have compiled this thesis using the listed literature and resources.
2. I hereby declare that my thesis has not been used to gain any other academic title.
3. I fully agree to my work being used for study and scientific purposes.
4. During the preparation of this thesis, the author used Claude 2.1 and GitHub Copilot in order to troubleshoot and suggest code, find appropriate English terminology, discuss and brainstorm ideas, check spelling and styling, answer factual and general questions, and other helper tasks. After using this tool/service, the author reviewed and edited the content as necessary and takes full responsibility for the content of the publication.

In Prague on Janary 2nd 2024 Antonín Kanát

References

KANÁT, Antonín. *ML-OSINT: Classifying Russian Vehicle Losses in Ukraine.* Praha, 2023. 55 s. Master's thesis (Mgr). Charles University, Faculty of Social Sciences, Institute of Political Studies, Department of Security Studies. Supervisor Mgr. Petr Špelda, Ph.D.

Length of the Thesis: 115,737 characters (incl. spaces)

Abstract

This thesis explores the potential of applying machine learning (ML) to assist with open source intelligence (OSINT) analysis. As the shared input of both disciplines, data is the primary lens through which the topic is examined. To understand the entire process of deploying an ML model from data collection to analysis, an image classifier of Russian vehicle losses in the invasion of Ukraine was trained and tested. Trained on a dataset of over 50,000 labelled images from the WarSpotting database, the classifier achieved a decent accuracy of 79% on evaluation data on the five most populous categories of images. On testing data from a later period, the performance dropped to 62%. One explanation offered is that the static frontlines and the prominence of drones led to most of the recent imagery being aerial, while the training data was captured mainly from the ground. That result demonstrated how inevitable changes, even in seemingly well-curated data, can lead to the low performance of ML models in deployment. Beyond changes on the battlefield, deeper data issues came to light, including the cascading effects of early data management decisions and dataset imbalance. Overall, current image classification methods do not work well on the noisy data available.

Abstrakt

Tato práce se zabývá možnostmi využití strojového učení (ML) při analýze zpravodajských informací z otevřených zdrojů (OSINT). Vzhledem k tomu, že data jsou společným vstupem obou oborů, jsou data hlavní optikou, kterou je téma zkoumáno. Pro pochopení celého procesu nasazení ML modelu od sběru dat až po jejich analýzu byl vycvičen a otestován obrazový klasifikátor ztrát ruských vozidel při invazi na Ukrajinu. Tento klasifikátor, vycvičený na více než 50 000 obrázcích z databáze WarSpotting, dosáhl slušné přesnosti 79 % na tréninkových datech pěti nejpočetnějších kategorií snímků. Na testovacích datech z pozdějšího období klesl výkon na 62 %. Jedno z nabízených vysvětlení je, že statické frontové linie a rozšíření dronů vedly k tomu, že většina nedávných snímků byla pořízena ze vzduchu, zatímco tréninková data byla pořízena převážně ze země. Tento výsledek ukázal, jak nevyhnutelné změny i ve zdánlivě dobře spravovaných datech mohou vést k nízkému výkonu ML modelů při nasazení. Kromě změn na bojišti vyšly najevo i hlubší problémy s daty, včetně kaskádových účinků raných rozhodnutí o procesování dat a jejich nevyváženosti. Celkově lze říci, že současné metody klasifikace obrazu nefungují dobře na dostupných nedokonalých datech.

Keywords

OSINT, Machine Learning, Image Classification, Ukraine War, Data, Loss Tracking

Klíčová slova

OSINT, Strojové učení, Klasifikace obrazu, Válka na Ukrajině, Data, Sledování ztrát

Název práce

ML-OSINT: Rozpoznávání Ruské vojenské techniky zničené na Ukrajině

Acknowledgement

I would like to express my gratitude to Mgr. Petr Špelda, Ph.D., for his immense flexibility and good will throughout the duration of the writing process, to Mgr. Adam Harmanec for providing consultations regarding the experiment and to developers and maintainers of libraries and tools used in this project. I would also like to thank all volunteers involved in the Oryx and WarSpotting projects for their immense work and willingness to share their data.

Contents

1 Introduction

Open source intelligence (OSINT) has experienced significant growth and popularity in recent years. Using only publicly available data, journalists and independent analysts were able to make groundbreaking discoveries. The ongoing war in Ukraine has propelled OSINT into prominence, with dozens of analysts tracking equipment losses, troop movements, and war crimes to inform public discourse and even Western government agencies.

However, the sheer volume of data available poses challenges for the analysts, who are overwhelmed by the workloads. Machine learning (ML), especially computer vision techniques, holds promise to facilitate the analysis by automating tedious tasks. Yet the suitability and practicality of applying ML to messy, unstructured OSINT data remain unclear. Furthermore, the impacts and effectiveness of such systems must be examined, given their sensitive application.

This thesis explores the viability of using ML image classification methods to contribute to OSINT analyses of the war in Ukraine. A supervised image classifier will be trained on a dataset of visual evidence documenting destroyed Russian vehicles. Documenting Russia's mounting losses helped to make the case for supporting Ukraine at the beginning of the war, and the assessed numbers have been used for both analysis and strategic communication by Ukraine and its allies ever since.

The model's performance, evaluated via quantitative metrics and qualitative analysis, will serve as a proxy for assessing the quality and consistency of OSINT data sources. Shortcomings identified throughout the process will provide some insights into similar applications but make no claim to be universally applicable to all contexts and modes of OSINT and ML.

The following analysis adopts a post-positivist stance, recognising that objective reality is obscured by the inherent limitations of observation. Rather than optimising model accuracy, the focus lies in critically examining the process to assess the suitability and impacts of incorporating ML into OSINT investigations. This direction is motivated by the ultimate dependency of algorithmic systems on the quality and representativeness of the available data.

The thesis hypothesises that the vehicle loss data, while abundant, are too noisy for widely available ML techniques to provide any meaningful contribution to an experienced analyst.

2 Literature review

2.1 OSINT

Open source intelligence can be defined as the process of collecting, processing, and analysing information from public data sources, such as mass media, social networks, public government data, or commercial sources. Academic papers would often attribute its beginning to World War 2, during which the British were analysing public radio broadcasts of Nazi Germany. According to Lakomy, a true boom of OSINT occurred after 9/11 due to several reasons: the broadening of the concept of security, the perceived insufficiency of human and signal intelligence to prevent the attacks, and the growth of the Internet as an ever-growing source of data that can be collected and analysed (Lakomy, 2022). In a way, this development is obvious. As a greater portion of our lives is lived through our devices and their networks, it follows that collecting data in this realm will get increasingly significant.

Over the past two decades, the concept has become broader. It can be used to refer to a subset of intelligence disciplines, along with HUMINT, IMINT or SIGINT. Ghioni et al. (2023) state that in 2018, between 70 and 90 per cent of law enforcement intelligence came from OSINT. From that perspective, OSINT is researched in relation to how law enforcement and intelligence agencies (and sometimes even corporations) collect and use public information. Ghioni et al. describe two levels of investigating ethical and legal aspects of OSINT. The first one is the macro level – the way OSINT impacts society and politics, and the second is the micro level – the impacts on individuals and organisations.

For most people, the term OSINT is arguably linked with the application of OSINT methods by investigative journalists or citizen activists. Using only publicly available data, individual researchers (or small groups thereof) have been able to solve major cases and bring light to new ones transparently and convincingly. The group Bellingcat has become synonymous with OSINT for their role in finding the perpetrators of the MH17 plane downing in 2014, identifying the Russian GRU assassins of Sergey Skripal in 2019, and discussing the details of an attempted murder of Alexei Navalny with one of the GRU agents in 2020 (Bellingcat, 2020).

But Bellingcat is not the only actor. Throughout the war in Syria, numerous analysts have been documenting and geolocating Assad's war crimes against the Syrian population. Various groups have been tracking the war in Ukraine since its inception in 2014, with the focus increasing since the invasion in 2022 (Block, 2022). A documented vehicle losses database run by a

handful of enthusiasts at Oryx has been the definitive source on this topic. In one video, a Ukrainian soldier filming a burned-out vehicle is heard saying, "This is for the Oryx website." Since 2017, NYT has been operating a Visual Investigations branch, with many of the staff having links for Bellingcat (NYT, 2023).

Most of the insights of this paper will likely be relevant to all applications of OSINT. Due to the context of the current war in Ukraine and the role of independent OSINT analysts therein, emphasis will be put on their situation, and relevant examples will be used.

2.1.1 The OSINT process

As OSINT has matured into an established discipline, it has developed its own methodology, with set workflows and a host of tools. Pastor-Galindo et al. (2020) describe how an investigation usually stems from an available piece of information, be it name, picture or location. A wide range of sources (search engines, public databases, social media) is then used to extract additional information, which is then in turn used to find and integrate further data. A wide selection of tools is available to aid in finding and connecting other pieces of information, but their description is beyond the scope of this thesis. To illustrate the process of an OSINT investigation, two short case studies will be presented. The first by Bellingcat relies on combining various leaked databases and social media, while the other combines social media discovery with geolocation and subject matter knowledge.

2.1.1.1 Identifying the First of Skripal's Assassins

The investigation is described in detail in a blog post (Bellingcat, 2018). The Bellingcat team began the investigation with a photograph of one of the suspects and his cover identity (Ruslan Boshirov). First, reverse image search was attempted via several online services, but to no avail. Then, no telephone numbers were found to be registered to the cover name. Therefore, the investigators focused on yearbooks and reunion galleries of the academy, from which a GUR operative with a focus on Western Europe would likely graduate.

There, they found a group picture with someone who *might be* Boshirov, with an annotation that the photo from Chechnya depicts recipients of the "Hero of the RF" award. Searching for the name of the academy, Chechnya, and the award led to the page of a Volunteer Union, where a certain Anatoliy Chepiga was linked to all three of those terms. However, a Google and Yandex search found no information related to such a person, which was highly suspicious for someone who had received the highest state honour.

Telephone databases were searched once more with the new name, showing an entry from 2003 with the address of "Unit 20662" in Khabarovsk, where the elite GRU unit is located. Another entry from 2012 in Moscow included Chepiga's date of birth, which was exactly one year later than the cover documents stated.

To confirm the identity, the researchers needed to find a picture of Chepiga. Strangely, his pictures were missing from all articles about the Hero of the RF award ceremony, even if the other recipients mentioned in the article were depicted. The systematic omission of Chepiga from the photographs further suggested that he may be a secret service officer.

The final confirmation came when Bellingcat obtained passport files from two separate sources (these tend to be corrupt government workers), where Chepiga's picture closely resembled Boshirov's pictures. The date of birth and place of residence (Khabarovsk) matched the rest of the evidence.

This example illustrates how a wide range of public and semi-public sources is used in an investigation. While it may not be clear from this simplified walkthrough, the process also involves dead ends and countless hours of combing through databases, websites, and social media posts.

2.1.1.2 Tracking Lost Vehicles in Ukraine

Although open source equipment loss tracking came into prominence after the 2022 invasion of Ukraine, many analysts have been following the Syrian civil war and other conflicts long before then (Block, 2022). Oryx blog and (now defunct) twitter account, which became synonymous with loss tracking, was started as early as 2013 (Mitzer, 2023). Other notable trackers include @UAWeapons, @Rebel44, @Danspiun, @naalsio26 and @Warspotting, among many others.

The general workflow consists of following a wide range of both Russia- and Ukraine-ran Telegram accounts, where videos are posted by soldiers and civilians from both sides, either to boast, document, or just to get some recognition. X (formerly Twitter) users also send what they believe to be previously unseen footage to relevant accounts.

Once an analyst receives an apparently previously unseen picture (and comes back from their day job), the first step is to identify the depicted piece of equipment. Over the years, analysts became quite adept at it, using their knowledge of the vehicles to identify even the most heavily

damaged or partially pictured machines. Guides on differentiating between different models or variants of equipment are also publicly available (Mäkelä, 2019).

Once the vehicle is identified, another round of checking whether the same piece of equipment has not been already documented from a different angle or in a different location ensues. Despite that, some duplicates find their way into the database, but volunteers regularly wade through the sizeable databases to draw attention to such duplicates (Jadrný, 2022).

The image, along with identification (and possibly some other labels), is then uploaded to the team's website (oryxspioenkop.com or warspotting.org), with the analysts publishing regular updates on newly reported equipment.

As mentioned above, all of the relevant analysts are volunteers who go through the data alongside their regular work. The workload, lack of recognition and duration of the war has already seen some analysts (such as @oryxpionenkop of the Oryx Blog and @CalibreObscura of Ukraine Weapons Tracker) have recently retired, with other groups or individuals taking over their work (Mitzer, 2023).

Even though the existing tools make data discovery and analysis easier, the 30 OSINT researchers interviewed by Ganguly (2022) confirm that the process is still incredibly laborious, requires a lot of tedious work, and is time intensive. This is not surprising – it merely reflects the centuries-old challenge of intelligence analysis. A wealth of data is available from IMINT, HUMINT, SIGINT and other sources. The data by itself, however, is of no use and needs to be understood and applied to get actionable intelligence. The heaps of data intelligence agencies today have at their disposal are both a blessing and a curse. The situation is no different in OSINT, possibly with the complication that even obtaining the data required a lot of sleuthing and clicking around. Could machine learning help?

2.2 Machine learning

2.2.1 Define ML

Machine learning (ML) is a field of computer science enabling machines to learn from data without explicit programming. The process involves exposing algorithms to large datasets to statistically uncover patterns and relationships (Alzubi et al., 2018). Many of the methods have been known for decades, but only the recent increases in computing power and data availability have made ML feasible for most applications. Using these simple methods, ML can learn to be useful in a host of real-world applications, including predictions (stocks, weather),

personalisation (product recommendations, search results), classification (medical diagnosis) and speech/image recognition (virtual assistants, autonomous vehicles).

A related term is artificial intelligence (AI). AI is a broader concept covering all systems mimicking human cognitive abilities. For example, the dominant paradigm of AI before the rise of machine learning was symbolic AI, which aimed to manually describe the logic, rules, and representations of human knowledge to produce expert systems. Along with hard-coded if-then rules and search algorithms, this approach could then, in theory, reason (Hofstadter, 1999). Apart from niche uses, symbolic AI proved impractical when handling real-world situations, including uncertainty and noisy data. With ML (where the algorithm uncovers the patterns underlying the data itself) being the dominant paradigm in AI, these two terms are often used as synonyms (Jakhar & Kaur, 2020). For the sake of variety, this thesis will do the same.

2.2.2 Approaches to machine learning

The major machine learning approaches involve supervised, unsupervised and reinforcement learning. Since the field is massive, I will provide just a brief introduction to the relevant topics below.

In supervised learning, the models are trained on input-output pairs, with the input being, for example, a picture and the output a label describing the picture ("cat"). The model starts at a predefined state, takes in a picture, and makes a prediction of the label. If the answer is wrong, the algorithm makes slight changes to the model's parameters and tries another input to see if the performance improves (Hastie et al., 2001).

Supervised learning can be very effective but requires the labelled data to train on. It also requires the task to be easily defined, with clear outcomes – e.g. classifying or identifying input into predefined categories. Indeed, one of the early ML applications was identifying pastry in Japan to save cashier's time (Somers, 2021). What if we have some data and hope to uncover underlying patterns or group those data points? Showing the model other piles of data, along with labels describing the patterns, would be unfeasible.

This is where **unsupervised learning** comes in. Its algorithms group unlabelled data points based on discovered similarities (Murphy, 2022). Clustering data points can be useful, for example, when a retailer wants to better understand different kinds of customers and tailor

marketing and services towards those groups. Another application is anomaly detection: when a clear outlier is identified in the data, it may point to malfunctions or unwanted activity.

But what if we know what we want the system to do but are not sure how it can be achieved? **Reinforcement learning (RL)** lets the model learn by trial and error, giving feedback in various ways (Sutton & Barto, 2018). Video games were a fit environment for early developments of these methods, so simple game scores as given to any player often sufficed. For games with scarcer rewards, developers had to devise alternative feedback, such as uncovering previously unseen parts (frames) of the game (Christian, 2020). One of the reasons behind the success of ChatGPT was that its developers let an existing language model provide several alternative answers to a prompt, and human contractors gave feedback on which response was most useful. The model then learnt by itself what makes a good response (Ouyang et al., 2022).

RL is very powerful but comes with certain limitations. First of all, it is very good at doing exactly what we reward it for, which may not always be what we want it to do. This can lead to a whole host of practical and existential challenges, the magnitude of which overshadows anything else written on these pages (Christian, 2020). Second, this approach requires a lot of trials (although, in the case of ChatGPT, the feedback process reportedly proved effective even after a relatively low number of rounds). This means that simulated environments where actions can be sped up and run millions of times (video games, board games, simulations) are ideal for RL. For problems constrained by the laws of physics, such as robotics, RL has still proven effective, but the number of rounds is severely limited (Kober et al., 2013).

2.2.3 The ML process

The general process of machine learning research and application is fairly standardised (Murphy, 2022). Due to its experimental and iterative nature, it sits somewhere on the line between science and engineering. There are voices frustrated with the current flood of academic papers blindly applying ML techniques to arbitrary scientific problems without any other added value or insight (Lipton & Steinhardt, 2018). The process below describes the typical supervised learning experiment, which is the approach this thesis will take.

The first step is usually gathering, cleaning and labelling data. As will be discussed later on, this is by far the least glamorous but also arguably the most important part of the entire pipeline (Sambasivan et al., 2021). If the relevant data does not exist, applying ML may simply not be

possible. This is the case of terrorism, nuclear risk, and large-scale conflict. Although these are important issues to investigate, the data covering them is so scarce that no reasonable models can be trained on them. Unlike, say, social media posts or baseball games. Given the importance of the issue of data, a large part of the literature review will focus on this topic.

The next step comprises picking a relevant ML algorithm with defined parameters and feeding the data into it. Prior to that, the data is split into three parts – one for training, one for evaluation during training, and the last for the final assessment of performance. One thing to note is that this and the following step can require a significant amount of computing power, which translates to time and costs.

Optimisation follows, with tweaking different parameters of the algorithm. The researcher can also trial different models or algorithms, as well as go back to the dataset. Given the importance of data, the best way to improve a model's performance is often data augmentation – producing additional data points by slightly modifying the existing ones – or other operations such as excluding poorly represented classes, filtering problematic data, etc. This phase usually includes multiple rounds of training, intending to reach optimal performance without overfitting – building a model that is optimised for the specific training dataset without an ability to generalise (Goodfellow et al., 2015).

Finally, the resulting model is tested with data that was not part of the training. This should, in theory, provide a measure of the model's performance on unseen data and evaluate the level of overfitting. For the reasons described below, obtaining testing data that reflects the deployment environment is problematic. Different performance in deployment conditions should always be expected, and the entire process should be able to account for any adjustments that need to be made.

The process described above is a good representation of the experiment included in the thesis. A more detailed and practical look at the individual steps will be provided in the methodology section.

2.3 Specific research on AI-OSINT

In 2020, Evangelista et al. conducted a systematic review of academic literature on the intersection of AI and OSINT. The resulting paper provides an interesting perspective on some

of the long-term trends in the field. By searching databases of scholarly literature using relevant keywords, they arrived at a body of 244 publications.

One interesting, if unsurprising, observation made by the authors is that not only did the topics of the academic literature follow the current topics and issues of society, but the studies often start from the latest available solutions or tools. This then dictates which phenomena are studied. A good example of this is the 2010 – 2015 spike in social network analysis. Thanks to the generosity and openness of Twitter (now known as X) APIs, the methods for such analysis became widely available, which led to an avalanche of papers utilising this technique (Zafarani et al., 2014).

It is arguable whether social connections on Twitter did indeed have such an impact on the outside world or if it suddenly appeared so due to the ease with which it could be studied. Nevertheless, this era may provide valuable insights into the current day. The trend of the past several years, where readily available ML frameworks are applied to any problem without a closer justification (including this thesis), may be later seen in a similar light.

Another interesting insight is that 88% of the published papers come from the US or Europe, with only 10 out of 244 from China. Given the volume of academic work coming out of China, combined with the country's emphasis on AI, this is a striking figure. While the author does not investigate this issue further, the explanation appears straightforward. A similar trend has been underway with the topic of cyber security, where Russian and Chinese academics mostly focused on information security, while the broader field has been operationalised as cyber security in the West. Importantly, this is not merely semantics – the fields do differ in their conceptualisation (Drazdovich, 2023).

As mentioned above, the term OSINT has two somewhat distinct meanings. The first is the OSINT citizen movement of investigative journalism and war tracking. Needless to say, such activities are unlikely to be legitimised and officially studied by Chinese academics. Understanding OSINT as a more general approach to analysing publicly available data introduces an interesting question. Within China, given the way public and private organisations are blended, is there even a meaningful distinction between publicly available and state-collected data? The concept of OSINT can be used by China when investigating foreign actors. Such research, however, is unlikely to get published.

Another interesting observation is the trends identified over the years. Evangelista notes that pre-2010, the papers focused largely on terrorism, then 2010 -2015 mainly on social media analysis, with 2015 leaning heavily towards cyber security and employing actual ML models. Interestingly, papers from 2005 to 2010 talked about AI as a potential solution to problems presented, while from around 2010, such solutions started being presented. Looking back, we can observe how the terms Machine Learning and Artificial Intelligence started to overtake older terms such as data science and big data (Vijayakumar & Sheshadri, 2019).

One thing to note is that ChatGPT was only launched a year ago (November 30, 2022), with the more capable GPT-4 update coming out four months later. Therefore, even two years ago, it was not at all clear that large language models (LLMs) would become the dominant approach to AI. For this reason and many other failed predictions, it is important to remember how unpredictable and dynamic the field is (Armstrong et al., 2014). In a sense, one could argue that LLMs are the peak of OSINT development, at least in the sense that they were trained on a vast swathe of publicly available data.

2.4 Data in OSINT

As can be seen, the two disciplines mentioned above seem to suffer from issues that complement each other. The main bottleneck of deep learning for many applications is a lack of a large volume of quality data. Intelligence analysis, traditional or OSINT, faces the challenge of navigating vast amounts of data available and distilling lessons from them.

To investigate whether combining these two fields holds the promise of resolving their issues, we need to take a close look at the uniting factor – data. In the following sections, we will examine the main features of data that AI, especially deep-learning models, require and the associated challenges. Then, the characteristics of the data available via OSINT channels will be described.

2.4.1 Sources of OSINT data

In line with the thesis' accent on data, OSINT data sources will be described in some detail, compared to existing tools, which will be largely omitted. Data that is used in OSINT analysis comes from a wide range of sources (Lakomy, 2022). The most commonly used ones will be described in the following section to illustrate their nature and foreshadow both their potential and shortcomings.

2.4.1.1 Social media

Posts and media posted on social networks are some of the notorious sources of OSINT data. The range of information that can be extracted from an X (formerly known as Tweet) or a VKontakte (Russian equivalent of Facebook) post is immense. First, a profile picture can help connect a name with a face. Users share their residence, workplace, and education history. Some posts may have the user's current location attached, or it can be geolocated from uploaded media. Users' friends, likes, and group membership can help in placing a person within a community or learning further information about them. Writing samples can be used for stylometry. The uploaded media can contain valuable information, such as class photos of military and intelligence academies (Higgins, 2021).

This data, however, comes with serious drawbacks. As the teen mental health crisis stemming partly from unrealistic portrayals of one's life on social media demonstrates, there is no guarantee the data voluntarily posted on social media is an accurate reflection of the underlying reality. The user can, intentionally or not, post misleading information. An analyst is also constantly at risk of succumbing to the observation bias and assuming that the social links and interactions that are documented online are the only (or the most) relevant aspect of an individual's social relationships (Ruths & Pfeffer, 2014).

Another major drawback comes from the fact that the data is hosted on private platforms. The extent to which the data is accessible for analysis is, therefore, highly dependent on the policy of a for-profit company. Advertisers, spammers, and actors other than OSINT researchers also use data for their ends (Humphreys & Wang, 2018). The platform is, therefore, incentivised to hinder data scraping and collection, which limits the volume and quality of data that can be obtained. To prevent scraping, the data on the platform is often obfuscated. Updates can be made to the platform structure, either due to technical development or to hinder data collection (Turk et al., 2020). So, while most data providers strive to provide documented, backwards-compatible and well-structured datasets, the goals of social media platforms are the opposite. Existing data pipelines can be broken overnight.

Another issue that will be described in more detail later is that the platform is free to delete, moderate, or modify any existing content without any notice. URLs cannot be assumed to be alive forever, and media hosted on the platform must be archived.

Furthermore, the platform's terms and conditions can be restrictive regarding the use of data posted on the network. While that may not be a concern for some vigilante researchers, the more institutionalised forms of OSINT have to be wary of this fact (Humphreys & Wang, 2018).

2.4.1.2 Other apps with social features

Other mobile apps and web platforms that may not be considered primarily social networks can be used for collecting valuable data. Every so often, a new article comes out, stating that a military base has been located (Pérez-Peña & Rosenberg, 2018) or a Russian officer has been assassinated due to the sharing features of the sport-tracking app Strava (Martin, 2023). Or that US nuclear secrets have been exposed on a flashcard website (Postma, 2021), and that the Polish army border buildup could be seen on the Belarussian Tinder, a dating app (Coakley, 2021).

However, I would argue that similar innovative use cases tend to be one-offs before either the users or the platform plug the relevant holes. And even when they prove to be useful, rather than just being interesting curiosities, the bottleneck appears to be in the innovative use of such a platform rather than the main point of this thesis – an abundance of data.

2.4.1.3 Satellite imagery

Over the past years, commercial satellite imagery providers have emerged. These companies, such as MAXAR or PlanetLabs, offer relatively affordable satellite imagery (in the range of hundreds of dollars per small region) to anyone down to an individual. Such a technology would be available only to major powers just a decade ago. Currently, the picture resolution of private satellites is limited to 30cm per pixel – this limitation lies in policy rather than technical capabilities (Bump, 2021). This resolution is sufficient for analysing buildings, vehicles, and large crowds, but not individual humans.

Even smaller organisations can now afford subscriptions to these services, with individual analysts sometimes receiving pictures for analysis for free. As the field grows, the imagery is likely to get cheaper and of better quality. Automated analysis of this imagery at scale is already possible, with companies like Orbital Insight or SpaceKnow offering such services to commercial customers. For example, their systems can automatically count the number of cars in parking lots or set up a watchdog for changes in equipment stationed at a military base (Bartošová, 2023).

The first limitation of these images is a relatively low revisit frequency – most satellites can only capture a given place every 1 to 3 days (Liu et al., 2021). With the time added by image processing and distribution, OSINT communities can often be seen impatiently waiting for the first images to come out. This frequency also means that objects present for only a short period may not be captured by any satellite.

Cloudiness still poses problems, as it covers the direct line of view for the satellite. Be it due to chance or not, the attack on Nordstream 2 happened during a period of dense cloudiness. Fortunately, photography satellites here can be complemented by radar imagery, which can penetrate clouds, and it was able to identify suspicious vessels with their maritime tracker disabled (Mareš, 2022).

While the civilian world has not yet adapted to the reality of widely available satellite imagery, governments had decades to develop techniques and procedures to avoid surveillance by other states. So, while satellite imagery can provide great insights into the fluctuations of the number of cars parked at the employee parking of Tesla factories or even larger deployments of conventional troops, it can be misleading for many military applications. Secret projects and activities take place underground or in hangars, and militaries have been known for decades to use decoy covers on experimental aircraft to evade detection by satellites. Decoys are omnipresent on the current battlefield, including very convincing ones (Graham-Harrison, 2023).

A recent and unfortunate example of this issue regards the work of prominent OSINT analysts before Ukraine's summer offensive of 2023. Given that some of the Russian fortifications could easily be visible on satellite imagery, comprehensive maps of the fortifications have been made, with the tacit assumption that they approximately capture the extent of Russian defences (Africk, 2023). When the operation started, however, much denser fortifications hidden underground or in treelines were present, which surprised many Western analysts. The density and quality of the fortifications were among the factors contributing to predictions of the possible outcomes of the operation (Zaluzhny, 2023). In this case, it appears that the analysis was misled by reliance on available but flawed data.

2.4.1.4 Maps and Google Streetview

While the previous data source is used mainly for damage assessment, object discovery, and identification, maps and orthophotographs (satellite maps) complement those investigations

and are used for geolocation. Unlike isolated satellite photos, services such as Google Maps or Google Earth Pro overlay satellite imagery over maps so that shapes in the photography correspond spatially to the corresponding features on the map. In addition to this, Google Earth Pro offers features such as comparing past satellite imagery of the same location and enabling the tracking of changes over time (Google, 2023).

These tools are most often used for geolocation, which is one of the main building blocks of OSINT analysis. Working from a hypothesis of where a photo or video could have been taken, major landmarks such as chimneys, power lines, major buildings or roads are identified. A map is then used to identify a location with a matching set of landmarks (ETI, 2022). While this is an arduous process, a pool of experienced geolocators can geolocate a picture based on very limited information. Google's street view, if available, is useful to check whether a view from a spot on the map matches what can be seen on the geolocated photograph.

A general limitation of maps is that the map author can decide to blur or exclude certain areas, often at the request of local authorities. Additionally, most mapmakers are for-profit companies that focus on profitable areas. The coverage and data quality are often much worse in locations where conflicts are happening. This is especially the case for Google Street View (Biljecki & Ito, 2021).

It is also very difficult to automatically collect data from Google and similar products, and such use may break the service's terms and conditions (Mostafi & Elgazzar, 2021).

2.4.1.5 Live-streaming cams, CCTVs and dashcams

This category bundles up two distinct sources of data. Dashcams (cameras mounted in the windscreen of a car) and local CCTVs tend to only save data locally, and the availability of the data they capture then depends on the user sharing the footage. Thus, they are similar in most aspects to the section on posting media to social networks. The one distinguishing feature, however, is that these cameras tend to operate automatically, capturing events that a user may not be able to capture reactively. Therefore, dashcams or local CCTVs would often be sources of footage on events such as plane crashes and overflights, military convoys and missile attacks (X, 2023).

Livestreaming cameras are connected to the Internet, and their footage can be watched remotely in near-real time. There are various ways in which the stream becomes publicly available. The first is weather and traffic cameras, intended for the public to be viewed.

The second category is cameras that are not intended to be publicly accessible, but they are. In the past two decades, the market has been flooded with cheap IP cameras, which businesses and individuals scattered around the built environment. While all devices connected to the Internet tend to lack proper security, IP cameras are one of the worst offenders, with factory settings defaulting to no authentication and being findable on the Internet. Even barring such basic oversights, cheap technology often has flaws that can be exploited. And since there is usually no functioning process of pushing security patches to the device (even if the vendor releases them), a large portion of the world's IP cameras are vulnerable (Cusack & Tian, 2017). Locating such cameras has been a pastime for internet users. For example, a YouTuber has found vulnerable CCTVs within a scam call centre he was investigating (Scambaiter, 2022).

Fig 1 A CCTV camera shows Russian vehicles crossing a Ukrainian border post. (BBC, 2022)

A third, specific category is private cameras, the access to which has been shared for some other purpose. The conflict in Ukraine is rich in cases where the regular border guard CCTVs were used for tracking the movements of Russian forces crossing the border (BBC News, 2022) or where a dislocated homeowner provided Ukrainian armed forces live footage from his house

to enable targeting of precise fire (Farberov & feed, 2022). Similarly to dashcams, these tend to be only available when shared, so they are listed here only for the sake of completeness.

These live feeds have various purposes in OSINT analysis. Cameras overlooking the city can be used for post-attack damage assessments and corroborating reports of strikes, fires, and explosions from other sources. Cameras overlooking roads or waterways can spot the movements of enemy troops.

Due to the greater bandwidth, automatically processing large volumes of video feeds remains resource-intensive. However, it is conceivable that object, event or change detection algorithms could be applied to the live stream or capture of such feeds to gain information (Canty, 2019).

Cheap IP cameras suffer from a whole range of limitations. Their resolution, frame rate, and image quality, especially in low-light conditions, limit the volume of information that can be extracted from the feed. Also, once the camera owner or the target of the surveillance learns about the feed, they can usually easily physically disable it, even if fixing the cyber security issues would be nearly impossible. Also, the footage is usually only available live, so the feed must be observed in real-time, which is laborious, or captures, which is resource intensive and unfeasible for a large number of disparate models of cameras.

2.4.1.6 Public datasets

This broad category covers many sources of data that are made available to the public for transparency, safety, research, commercial and other purposes. A creative use of such sources forms the backbone of many OSINT investigations. Since the data is intended to be consumed and further processed, it is often presented in a standardised way, along with documentation or clear formatting rules. Depending on the source institution, the data also tends to be quite reliable, at least in presenting the reality from the dataset's paradigm (GOV.UK, n.d.). Since data scraping, collection, wrangling, and processing are often the most laborious part of data analysis, this ease of use is a factor not to be underestimated.

Mandatory tracking systems, both for ships (AIS) and flights (ADS-B), are popular amongst hobbyists and analysts. Given that their use is not enforceable, but there is little reason to do so under normal circumstances, tracking *invisible* craft or detecting transponder shutdowns can provide more information than combing through the available data (Zuzanna et al., 2022). These sources were used to identify suspicious vessels around the Nordstream II attack, the

flights of Yevgeniy Prigozhin's jets or China's shadow fleets overfishing in territorial waters of neighbouring natures (Project & CBC, 2020).

Public and corporate registries, court documents, land cadastre, domain registry, company ownership structure, newspaper archives and others are useful for investigations regarding illegal activities. With this data being highly formalised and text-based, even early ML methods could be applied to similar issues (Marappan & Bhaskaran, 2022). As is usually the case, the analysis needs to be adjusted to the relevant context, as the assumptions regarding data quality and its relation to reality may differ across different nations and legal systems.

Systems intended primarily for research or specific purposes are often utilised in OSINT analysis. A great example of this is NASA's FIRMS dataset intended for tracking forest fires. Its worldwide coverage and open access allow analysts to track large fires, thus corroborating reports of large explosions and fires, even including general assessment of shelling intensity along the Ukrainian frontline. Weather models and archives can aid in the temporal placement of footage (Gonzales, 2022)

2.4.1.7 Leaked datasets

Datasets that are not meant to be public offer valuable insights. Leaked website databases can help link a username to an email or IP address. Corporate documents offer insights into the organisation's activities. These datasets usually become available due to whistleblower leaks, negligence, or criminal activity. In a way, this process brings data that would usually be only available to the government or an organisation into the public domain.

Bellingcat makes extensive use of similar datasets in their investigations of Russian intelligence operatives. Due to poor data practices and the ever-present corruption of Russian society, virtually any data is available for sale. Bellingcat's analysts have thus obtained the vehicle registry records, stolen database of contact records linking phone numbers to names, and even passport data. The latter was done via a seamless process of contacting a dedicated Telegram channel and paying a set price of around 20 USD for a corrupt official to perform an on-demand lookup (Bellingcat, 2018).

Leaked datasets offer valuable insights, but one can hardly rely on their authenticity. Some kinds of data can also be expensive to obtain (and thus likely are not considered open sources anymore). In some contexts, it is also questionable whether paying money to corrupt officials

and organised crime networks holds up to moral scrutiny. That is especially true if the funds come from the budgets of journalists or analysts fighting crime.

2.4.2 Data Limitations

The following section will cover issues relating both to data in general as well as the specific challenges of OSINT.

2.4.2.1 Raw Data Does not Exist

One of the coveted features of some deep learning methods is its ability to learn from *raw* data such as text corpora or image datasets. Older ML systems required often manually performed feature extraction, such as splitting text into arbitrary tokens and counting their prevalence, and similar statistics, such as word or sentence length. For images, edges would be detected and counted, colour prevalence measured, and contrast considered. Deep learning, however, can be fed the *unprocessed* data and extract the most important features of the data itself (Nixon & Aguado, 2019).

This characteristic often improves the model's performance, reduces menial and often high-skilled labour, and introduces applications that would otherwise not be possible. Additionally, there was initially this notion that this could even decrease the bias and blank spots of the models, as a human is not involved in the process. Thus, a truly impartial system could be built (Cremer, 2020).

However, this notion merely pushes the weak link one step down. While it is not a human anymore who decides what features of a data point are relevant, the data still came to be due to human action. Broad academic literature regards the issue of *raw* data, the meanings associated with this term, and how bias (both in the statistical and ethical sense) can be introduced in raw data (Räsänen & Nyce, 2013)

For example, Muller (2019) argues that data is never *raw,* despite this term being often used by analysts, investigators, processes, and scientists. Human is always in the loop. Muller presents a 5-step process of creating a dataset, at each step of which the data is influenced.

- **Discovery of data.** At this step, a source of data is discovered. Given the infinite complexity of the physical environment, it is impossible to be sure that all relevant data has been collected. This step inevitably contains the decision to stop searching for further data. In practical terms, later analysis may assume lack of data as evidence for lack of activity, whereas an adversary may have merely been

communicating on a different channel (radio frequency, online platform, mode of communication).

- **Capture of data**. At this step, the discovered data is captured to be processed. Due to capacity constraints, a decision has to be made regarding what data will be captured. Some forms of information cannot be captured with available tools. Measurement errors may be introduced. Importantly, all of these effects are unlikely to be uniformly distributed. For example, pictures of certain objects or areas will be of worse quality due to the environment, which could lead to underrepresentation of that area.
- **Curation of the data.** Curation here refers to integrating different sources of kinds of data. Again, an analyst or engineer must deem two data sources or forms to be relevant for them to be linked together.
- **Design of the data.** For the data to be usable down the line, it needs to be properly formatted and packaged. At this point, issues like missing data or obvious outliers have to be handled. In the context of warzone pictures, an analyst may decide to leave out images that are not geolocated. However, this favours easily identifiable areas at the expense of featureless countryside, potentially skewing the analysis.
- **Creation of the final dataset.** The way the dataset is then published or utilised is also influential. The Warspotting team now focuses mainly on Russian losses, as monitoring these is instrumental to the values and goals of the Warspotting team.

The myth of *raw* data appears to be reinforced by the *ubiquitous data, information, knowledge, and intelligence* (DIKI) charts depicting the intelligence process of going from data through information to intelligence. On the DIKI chart, the first (or zeroth) part – the signals, environment, reality - is missing. This chart is often on the first slide when data gathering and analysis are discussed. Luckily, the intelligence community appears to be cognizant of this issue, with the US 2013 Joint Intelligence Manual including the operational environment in a similar chart (JCS, 2013).

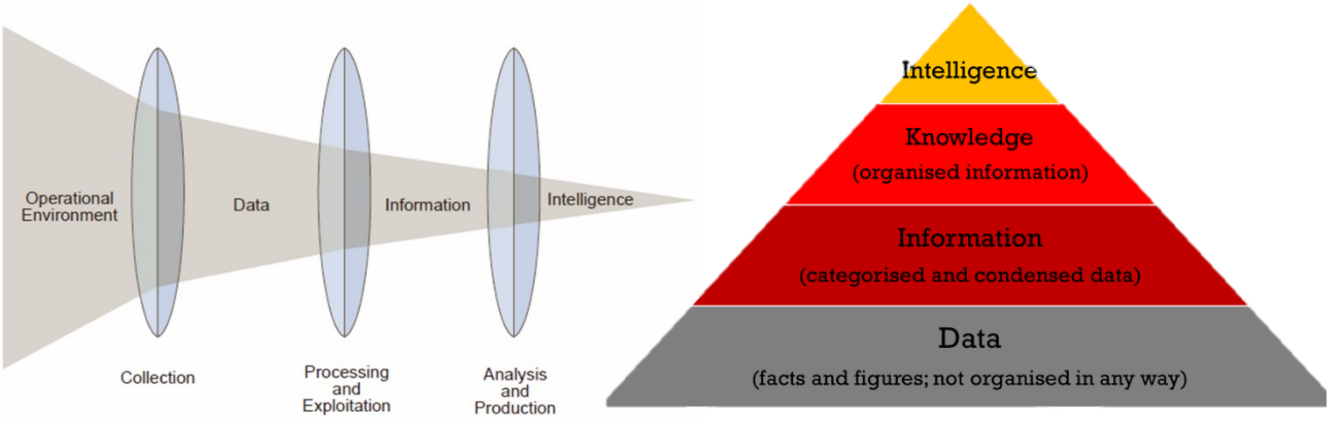


Fig 2 A comparison of a (good) ODII with the (bad) DITI chart (JCS, 2013)(Petavradzi, 2017)

Speaking of this chart and the intelligence process it describes, Tuomi (1999) goes even further and argues that it, in fact, runs backwards. Rather than working from arbitrary data they collected, intelligence agencies almost always start with a goal in mind, which then dictates what information is required and which data will be collected.

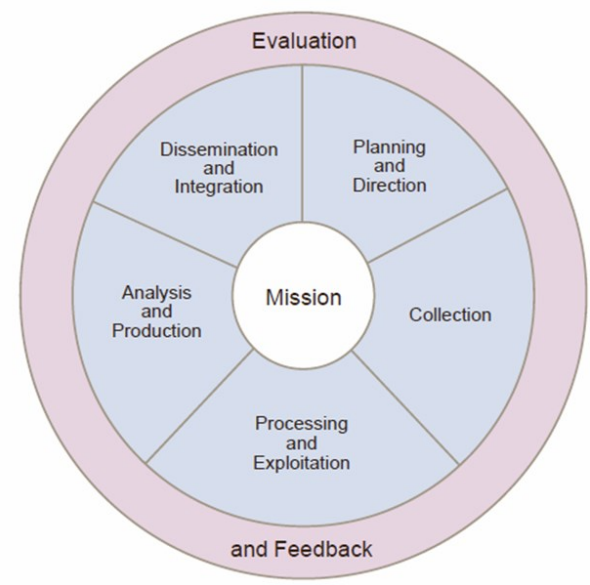


Fig 3 The intelligence process (JCS, 2013)

Tuomi's description corresponds to the first two steps of the intelligence cycle, which is the other notorious chart of Intelligence 101. Therefore, the point may appear trivial, as it merely restates one of the basic principles of intelligence analysis. Nevertheless, Tuomi's point is valuable in pointing out that even the mere fact that the data exists means that somebody has decided this data is to be stored and processed.

While the last point may appear trivial, the reality is often shaped by what data is collected. For example, the aforementioned NASA FIRMS imagery, intended for detecting forest fires, is publicly available and used to detect shelling and fires following missile attacks. We do have decent data on the movement of gold in and out of the US since the metal is easily detected and must be declared, while the data on moving valuable artwork is much spottier (Teichmann & Falker, 2020).

2.4.2.2 Data Work is Undervalued

An entire class of issues regarding data has been described in Sambasivan et al.'s seminal, aptly named paper *"Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI"* (Sambasivan et al., 2021).

The overarching issue is that the model work – developing and finetuning the neural network – is exciting, while the equally important data work is seen as grunt work, and no one wants to do it. Other factors come into play, such as the common sentiment of software developers to quickly move forward and deploy the model.

Unfortunately, not enough attention is then often given to the way the data is collected, processed, and analysed. The model's performance is seen as a proxy of the combination of the model's fit and data quality. In hobbyist and arguably even for-profit projects, this approach of fast deployment and good enough solutions may be acceptable. Very often, however, this approach is also brought to environments that are much more high-stakes, such as healthcare, criminal justice, or social policy.

When the data causes the model to underperform, the reason is investigated, and some of its shortcomings are remediated. Much worse are situations where the model performs reasonably well and the problems with the data manifest in unpredictable and hard-to-investigate ways. Sambasivan calls this downstream impact of data quality in high-stakes applications data cascades. According to the survey of 53 AI practitioners, 92% of large, high-stakes AI projects suffered from at least one data cascade. Investigation of those cascades produced interesting observations:

All the researchers were committed to data quality and had deep moral commitments to vulnerable groups. Therefore, these issues cannot be attributed merely to for-profit or malevolent actors. As mentioned above, gathering and processing data is laborious, so it is often performed by different people or organisations than the models and applications. It is common practice, for example, to outsource the labelling of datasets to Amazon's Mechanical Turk, which employs armies of workers from low-income countries (Samuel, 2018). In the context of wildlife preservation or health care, the work can often only be done by local groups and organisations.

This leads to a situation where no single person oversees the entire process. And while no individual involved would want their work to negatively impact someone's life, the workload is so compartmentalised that this perspective is lost. This can lead to major, well-intentioned projects suffering from issues one would accept in a market research study.

As is usually the case, the failure lies within the incentives and system rather than individuals. Those collecting the data are often not incentivised, educated, or trained well enough relative to the impacts of the resulting application. Model work has traditionally been more glamorous, lucrative, and impressive than data work. Sambasivan also found that very few students studying computer science or machine learning degrees received classes on data quality, at least relative to the attention paid to model training and finetuning.

Hobbyist OSINT is a great example of this issue. The data is collected, labelled, and presented by volunteers in their spare time. While diligent and conscientious, those analysts are probably not thinking about the potential impacts their work might have. As discussed elsewhere, governments are likely to take their analysis into account while making strategic decisions. Targets geolocated by hobbyists have previously been struck by missiles and artillery of the Ukrainian armed forces (Salerno-Garthwaite, 2022).

Simbasavan also notes that data is often either collected or analysed without subject matter expertise. That reduces the chance of spotting discrepancies in both the internal and external validity of the model. Similarly, posts on the social network X (formerly Twitter) can be seen visualising or presenting Oryx data by users lacking subject-matter expertise. Their implied or explicit conclusions are then often misleading (X, 2023).

Finally, as is usually the case, Simbasavan observes that these cascades usually disproportionately affect marginalised groups. This can create feedback loops, in which groups for which low-quality or little data is available receive less attention in the model, leading to further marginalisation.

2.4.2.3 Data Archiving

One of the most pertinent issues of OSINT is the retention of data. Social media posts, pictures and videos are often posted on private platforms. These platforms may then be incentivised to remove this content, be it for moderation, compliance, or economic reasons. Since individual researchers rarely can archive vast amounts of data, the entire community is at the mercy of private companies. In 2018, YouTube started automatically deleting videos its algorithms deemed related to ISIS propaganda. At least hundreds of videos documenting Bashar al-Assad's war crimes have been deleted as a byproduct of this system (Rosen, 2018). In May 2023, Elon Musk announced the Twitter platform (now known as X) would be deleting old inactive accounts, which also include accounts documenting the Syrian civil war (Omar, 2023).

2.4.2.4 Impact of Generative AI

The rise of generative AI, which can produce convincing text, images, audio, and video, will have a major impact on OSINT analysis.

The first and most obvious hurdle is the potential for fake or edited data to mislead analysts. Deception, image editing, and false flags have been a threat for as long as OSINT analysis existed. The issue is that generative AI may make this process much easier, thus being able to flood the information space with conflicting data. That could be an issue since Russia had so far resorted mainly to taking pictures of the same vehicles from different angles, which was easy to spot. Otherwise, they would move the vehicle and slightly modify it, but that is a very labour-intensive task that is easy to detect (Paul & Matthews, 2016).

On the other hand, AI systems may help identify generated media, point out similar or duplicate images, and help analysts with their work (Marra et al., 2018). All in all, the active debate on the role of generative AI in propaganda and the information space is out of the scope of this thesis.

One related issue is the availability of good-quality training data for AI models. Some argue that we as humankind are (have been a year ago) at the peak of data quality, where large volumes of original data – pictures, social media posts, videos, books, essays - were produced and stored. As generative AI becomes widely available, the information space is now flooded with AI-generated data that will inevitably get into the training datasets for AI models (Shumailov et al., 2023).

This vicious cycle, which will likely decrease the overall veracity and quality of model outputs, is a major issue with no clear solution. Employing metadata proving data originality currently appears infeasible at the required scale. Excluding data originating after the spread of generative AI would bar the models from learning about recent events and facts.

While this issue may not be specifically related to OSINT, it appears worth mentioning to understand the uniqueness of the current situation, where large, good-quality, labelled datasets are available but appear not to have yet been polluted by generated media.

2.4.2.5 Distribution Shifts

As will be described later in the methodology section, this experiment (and the entire field of ML) takes the general form of splitting the available training data into two parts, training the

model on one and then evaluating it on the other. The best-performing model is then broadly assumed to be optimal for deployment in the field. This assumption is challenged by the fact that the world is constantly evolving, and the data collected when a model goes live will almost certainly be different from the training data. If the model is overfitted on the training data, it may then perform significantly worse in deployment when data is even slightly different. This phenomenon is called distribution shifts (Spelda & Stritecky, 2022).

These distribution shifts occur even in most lab data, carefully controlled for outside interferences. For example, a model for detecting eye tumours flopped when the actual images contained specks of dust (Sambasivan et al., 2021). Given the dynamic nature of the conflict in Ukraine, distribution shifts are an absolute certainty. One thing to note is that we currently do not have reliable ways of understanding how a model classified a certain image. Unbeknownst to us, the model may be using (to us) entirely irrelevant features to produce its input. This means that even shifts that appear harmless to the human analyst (such as the season) may have a major impact on a model's performance.

Below is listed just a small sample of expected distribution shifts.

- As the weather changes, the surrounding environment will be different. As the war progresses, we also see shifts between urban and country terrain.
- With the frontlines being more static, images of destroyed enemy equipment will likely be mostly taken by drones rather than soldiers passing captured areas. Over time, the quality of drone footage improves.
- The equipment of both sides is changing, with Ukraine receiving Western vehicles and Russia taking old or rare models out of storage. Modifications such as reactive armour or protective copecages are widespread.
- The vehicle and unit markings of both parties develop over time as units are reformed, and new offensives are planned.
- Over time, the means of destruction are changing. A mostly ATGM-dominated phase of the war will produce vastly different footage from a phase of landmines and/or FPV (first-person view, also known as kamikaze) drones.

As can be seen throughout this paper, and given its post-positivist approach, this thesis works with the assumption that data coming out of Ukraine now is very much different from those received three months ago. On the more practical level, the experimental design and model training will try to take this fact into account.

2.4.2.6 Data Labelling

Data labelling, briefly mentioned before, is an important part of the ML pipeline. Given the volume of data modern models require, data collection and labelling become very laborious. Historically, this has led to either everyone reusing the same available datasets or outsourcing the labelling process.

The issue with extensive use of available datasets is that individual models, and potentially even entire subfields, can get overfitted on them. Additionally, many of the available datasets contain only good-quality data, which often translates to poor deployment performance (Koch et al., 2021).

Outsourcing labelling to workers in low to middle-income countries – or to internet users via Captchas (O'Malley, 2018) is economical even at scale but has its limitations. First, the labelling must not require subject-level or cultural knowledge. If it does, the labellers may establish flawed ground truth due to their lack of knowledge. Sourcing and employing qualified contractors can quickly become unscalable and expensive. The other issue is peculiar – while social sciences and humanities have long been criticised for sourcing most data from rich young students in high-income countries, a lot of the data ML systems are trained with come from poor people from low-income countries. Just like in psychology, this can introduce biases into the datasets (Samuel, 2018).

The previous paragraphs can help us appreciate the fact that the loss of datasets from the conflict in Ukraine exists. These have been put together by qualified and motivated volunteers in a transparent way that enables errors to be fixed. OSINT analysts can often distinguish between different production runs of a vehicle based on a blurry picture shot from the Orlan 10 drone. The amount of work that went into compiling the dataset used in this thesis and others is immense, and we are lucky to have it available.

The current situation, unfortunately, does not appear sustainable. The two accounts synonymous with loss tracking in Ukraine (@oryxpioenkop and @UAWeapons) have ceased their activity in the past months, citing the workload and minimal reward as the main reasons (Mitzer, 2023). The remaining analysts who took over their work also often voice their frustration about the situation where major intelligence agencies use their work without any compensation. One could argue that publishing data for free and encouraging followers to instead donate to relevant charities while also complaining that one is not compensated for their

work is in conflict. However, with the war coming into its third year, this situation just demonstrates the unsustainability of this level of tracking merely off the free time of enthusiasts (WarSpotting, 2023c).

2.4.2.7 Adversarial Inputs

Along with distributional shifts, adversarial inputs are one of the major threats facing deployed ML models. Adversarial are those inputs that are specifically tailored to make the model produce an erroneous output. Traditionally, these include an adversarial panda (Goodfellow et al., 2015) or adding small stickers to traffic signs, which makes self-driving cars misclassify them (Eykholt et al., 2018).

In Ukraine, we are seeing a widespread use of decoys. These range from very rudimentary boxes with tubes to high-fidelity decoys. In November 2023, Russian telegram channels posted footage of destroying what looked like a Ukrainian Su-25 aeroplane. After a day of back and forth, the OSINT community reached the conclusion that the aircraft was most likely a superb decoy, probably built with real spare parts (Newdick, 2023).

Fig 4 The likely Su-25 decoy, as captured by the attacking Lancet drone (Newdick, 2023).

Admittedly, these decoys are not built to fool an OSINT analyst but to make the adversary waste precious resources and give up position. A community of analysts is able to filter out many decoys upon closer inspections. In the case of FPV drones, the fact that the target is not

real may even become apparent to the drone's operator as the drone gets closer to the target seconds before the impact. For footage of low quality or shot from a great distance, a lack of secondary explosions of fuel or ammunition is often the only indicator that the target was either a decoy or was merely damaged.

One of the great strengths and weaknesses of ML systems is that they are very good at learning precisely what we reward them for. An experienced human analyst strives for accuracy but withholds judgment when necessary. An ML system rewarded only for correctly classifying an input into one of the predefined categories has learned to guess and fit reality into one of those boxes. In practice, models can express certainty, and classifications below a certain threshold may be withdrawn. If there are no decoys in the dataset, it can be expected that a model will classify an object with features similar to one in the dataset as belonging to its class (Hendrycks & Dietterich, 2019).

A hypothetical, practical ML system, which would be an extension of the one introduced in this thesis, would most likely still have to operate along with a human analyst to verify the outputs. The low quality of the footage, combined with the complexities of vehicle identification (designs are often based on each other), makes it unlikely that such a system could be reliable enough to even consider standalone deployment.

2.4.2.8 Impacts of Graphic Data on Mental Health

A final issue related to OSINT coming from conflict zones is that they often contain graphic content. Almost by definition, the more outrageous the content (executions, harm to children), the more important it is to investigate the incident. 26 of the 30 researchers interviewed by Ganguly (2022) reported suffering from PTSD and similar conditions due to their extended exposure to distressing media. This reality has two implications.

First, it raises the barrier of entry for new analysts – experienced researchers often urge their audiences not to look for the footage whenever a new atrocity comes up "unless it is your job" (Fiorella, 2022). Second, this media is very likely to be deleted or even prevented from being uploaded to social media platforms with content moderation. That is good for the wider society but makes data archiving and potentially even discovery more difficult.

Law enforcement officers, especially those involved in crimes involving minors, have been facing these issues for decades. One of the features included in investigation systems is the

automatic blurring of graphic images (Perez et al., 2010). While an analyst will never be able to fully avoid exposure to graphic content, further automation may make it possible to limit such exposure.

2.4.3 Literature Review Conclusion

As demonstrated in this chapter, the issues surrounding OSINT, its data sources, and AI pipelines are complex and often unpredictable. Therefore, an experiment that would emulate the entire process of integrating OSINT data with ML tools is proposed. Even if the resulting model may not be practical to use and is very unlikely to outperform human analysts, lessons identified throughout the process will be valuable.

3 Theoretical Framework

3.1 Post-positivist Approach

Based on the limitations and challenges identified in the literature review, a post-positivist approach will be taken in this thesis. Post-positivism holds that an objective reality does exist, but our understanding of it is always limited by our viewpoints as researchers (Ryan, 2006).

The first assumption does not need much defence – while more relativist approaches can be useful when studying social and other abstract phenomena, the objects depicted in the dataset weigh between 3 and 50 tonnes, which leaves very little space for discussions regarding their tangibility. The previous chapter also illustrated the sheer breadth of limitations in the way the data and tools available reflect the reality on the battlefield.

Interestingly, military theorists and practitioners are familiar with the concept of the *fog of war*, the foundations for which were laid by (whom else than) Clausewitz: *"War is the realm of uncertainty; three-quarters of the factors on which action in war is based are wrapped in a fog of greater or lesser uncertainty."* (1950). The concept is broadly used to refer to the uncertainty and lack of situational awareness inherent to military operations and thus is, in many aspects, aligned with the approach of post-positivism.

The Russian and Ukrainian general staff themselves have limited awareness regarding the state of their army and equipment. Therefore, it would be foolish to assume that one could get an accurate understanding of the situation on the battlefield based on images posted on social media. Nevertheless, trying to attain as accurate a representation of reality as possible is vital, both for the conflict parties and outside observers.

Another aspect of post-positivism is the focus on falsification – it seeks to disprove individual points rather than trying to exclude all but one hypothesis (Phillips & Burbules, 2000). Given the limited resources available for this thesis, this approach makes it possible to make valid, if minute, observations rather than attempt to support a grand theory.

3.2 Aims of Critical Analysis

Therefore, the main aim of the thesis is not to develop a highly accurate model for identifying vehicle losses. Given the constraints, such a goal would be almost inconceivable to reach and would likely lead to a result that underperforms and does not generalise. Instead, the primary aim will be to critically analyse the process and limitations of trying to understand the complex reality of battlefields through AI systems.

Because of that, the focus will be on the individual steps of the process of building the system rather than its finetuning and optimisation – most lessons are likely to become apparent regardless of 0.5% improvements in the model's accuracy. The model thus serves as an analytical lens to explore the nexus of machine learning and OSINT.

In line with this approach, issues that may be seen as mere oversights on the author's side will be documented (Ryan, 2006). While the average ML project may be conducted in a more professional way, all projects will struggle with human error, lack of knowledge, time pressure and similar issues. Brushing these away as not integral to the topic is a naïve approach that leaves much insight on the table.

As mentioned before, the core tenet of this text is data. Deep learning models are powerful, but their bottleneck lies in the availability of a large volume of good-quality data. At the same time, OSINT and intelligence analysis face the challenge of processing and making sense of a never-ending stream of new data. The question then is whether the open-source data is good and abundant enough for ML and whether the models will be effective enough in distilling meaning from the data or at least facilitating human analysis.

3.3 OSINT: Data Source and Phenomenon

While OSINT and general intelligence gathering have been conflated in the previous paragraph, a clear distinction should be made between them, as the thesis will be concerned with OSINT data and community. With OSINT being defined as intelligence and data collected from publicly available sources to inform decision-making, the term is concerned largely with the source of the data and can be considered a subset of intelligence collection and analysis (Ghioni et al., 2023).

However, it can be argued that the nature of OSINT data is so different from other collection methods (IMINT, HUMINT...) that a paper examining the applications of ML in general intelligence analysis would highly likely lead to different conclusions. For example, intelligence agencies have traditionally had control over a significant portion of the data pipeline – deploying wiretaps or developing spy satellites enabled them to influence what data is being collected and how (Sanger & Miller, 2014).

Comparing this to OSINT data, where one has to make do with whatever data is available, makes it clear that applying ML models to OSINT data is likely to be less scalable than to data

sources one has full control over. The thesis thus works the hypothesis that OSINT data will be messier and less consistent than traditional intelligence sources. At the same time, it could be argued that the growing ability of ML to process unstructured or multimodal data may disproportionately impact OSINT by enabling new or previously impractical approaches (Wu et al., 2023). At the same time, the impact of similar developments on well-structured data may be only incremental.

The second way in which the focus on OSINT makes the thesis take a different direction from a more general ML – intelligence paper is when we consider the broader context of the OSINT community. This thesis treats OSINT not as just one of many sources of data but as a broader social phenomenon, which is, in most people's minds, linked with its applications by activists and investigative journalists.

This will be reflected by assuming limited resources – human, computational, and capital – on the side of the analyst. Also, the angle informs the topics covered. Tracking equipment losses is a valuable enterprise with real impacts that are achievable for a handful of individuals. It has become one of the poster children of OSINT and will be the main focus of this thesis. A state agency, be it Ukrainian, American, Russian or NATO, highly likely chooses to devote its resources to more actionable intelligence, such as eavesdropping, tracking enemy movements, targeting, etc (Murauskaite, 2023).

3.4 Machine Learning

The second piece of the puzzle – machine learning – is approached rather widely as a field of computer sciences concerned with computers able to learn without being explicitly programmed. Despite the many differences between different ML approaches and models, the data requirements appear rather consistent. Nevertheless, this paper will follow the approach of supervised machine learning, which uses labelled training and evaluation sets to learn to classify new data. It is also the approach most relevant to the use case in question (Murphy, 2022).

The vast majority of applied ML research is quite light on theory, with the underlying principles and assumptions of the models being treated similarly to fundamental laboratory procedures in chemistry and biology papers (Lipton & Steinhardt, 2018). As will be described in more detail in the methodology section, the most common techniques and off-the-shelf

approaches will be used. The reason for that is to ensure the observations are applicable to the most common architectures present in academia and industry.

Once the model is trained, it will be evaluated by both quantitative and qualitative means. Quantitative classification metrics, such as the model's accuracy, are the default measures of a model's performance. These will be treated as an approximate measure of the model's performance and, by proxy, the quality and suitability of the data for ML application. However, as optimisation of the model's performance is not a goal of this project, more attention will be placed on the confusion matrix. This measure contrasts the actual and predicted classes of images, which allows a quantitative assessment of which classes tend to be misclassified and between which confusion most often happens (Deng et al., 2016).

Qualitative evaluation approaches are likely to provide more relevant insights. Error analysis manually examines a sample of incorrect predictions in order to try to identify specific limitations and potential biases. Edge cases – challenging or atypical examples that a human could handle well – may expose brittle reasoning. Finally, an analysis of the dataset, even before the model is trained, can bring valuable insights into gaps or misrepresentation issues.

To conclude, the thesis operates with the following primary hypothesis: Even though OSINT analysis struggles from data overload, the vehicle loss data coming from the war in Ukraine is so noisy that widely available image classification models are not suitable to contribute to the analysis.

4 Methodology

4.1 Training Dataset

4.1.1 Collection

The two major loss datasets – Oryx and Warspotting – are largely comparable and (according to their own words) largely build on each other's work (WarSpotting, 2023a). For the training data, preference was given to Warspotting's database, as it features some additional labels such as "Cope cage", "Loitering", or "Turretless", as well as multiple separate pictures of many losses. This decision also meant that the project would include only Russian losses, as that is what the team focuses on.

In October 2023, administrators of the website were contacted to request access to the database. To respect the volunteers' time, the author offered to scrape the website if there was no easy way of exporting the database available. The Warspotting team provided a Google Drive link with all losses until the start of July 2023. While the additional six months' worth of data that have been accumulated since then would have been useful in increasing the dataset size and including new developments on the battlefield, no further requests have been made to the team. To go against the general entitlement to the time of volunteers and their work by individuals and organisations, the author decided against making any further requests.

The folder contained a folder with one subfolder for each day when new vehicles were added, which contained the relevant images. One .csv file mapped each image path to a specific observation ID, with another .csv file providing the vehicle type, model, status and link to the Warspotting website for each observation ID. In total, there were 59,917 image files in 376 folders.

4.1.2 Processing

As was foreshadowed by the literature review, wrangling this data into a form suitable for the application of common ML libraries was a major undertaking. Below are listed some steps required, as well as related issues or limitations.

- The number of files is so high that manual review is not feasible. Therefore, there is no practical way of discovering if some images were mislabelled or if there was an issue in the data processing logic, causing limited misclassification. While random samples of images were manually verified throughout the entire experiment, smaller issues may have been missed.
- The observation ID file contained 15341 rows, with 18829 being the highest ID number. While that is not documented, consecutive IDs are presumably assigned,

which would already point to a considerable number of adjustments and corrections of the data points. A significant portion of the website links included in the file were not working anymore, and some equipment cannot currently be found on the website at all (i.e. the P-3537 Bar Lock radar).

- The image mapping file only contained 56422 rows, meaning that some 3500 images were not accounted for. Thus, ~6% of the images are excluded when merging the information before any analysis even starts.
- After dropping rows with missing image paths, July 2023 (for which the labels were available, but not the images), and a handful of corrupted anomalies, the total number of well-labelled images dropped to just 50717.
- To illustrate the perils of documentation not being available: the dataset also contained a field *area*, which was only present in around 10000 lines. A careless researcher rushing to deploy the model may decide to disregard this field due to its sparsity or reasonably assume it contains geolocation coordinates (which are available for some images on the Warspotting website). Upon closer inspection, however, it becomes clear that the area refers to the coordinates of squares which outline specific vehicles on pictures where multiple are visible. This is implemented in the browser in a rather uncommon way.
- The coordinate system thus had to be reverse-engineered, and a script was written to save a cropped copy of the relevant pictures for each relevant observation. If this step was overlooked, the model would get confused by being presented with multiple vehicles with one label, and the total number of training images would be lower.
- After cleaning the dataset, the class distribution can be assessed. As can be seen in Table 1, representation differs significantly among classes. *Imbalanced datasets* like this one often cause problems for ML algorithms, especially in classification tasks. Without proper adjustments, the model is likely to learn to be biased towards the largest classes, and many performance metrics may be dominated by the model's performance in those classes. The model may also end up overfitting on the little data available for small classes.
- While a more detailed dataset analysis will be presented in the Results section, one interesting observation can be made. The aeroplanes section, already modest at 385 pictures, contains a significant number of pilot obituaries and Telegram posts of condolence. While it is perfectly reasonable for crash site images to be corroborated by these posts, confirming aeroplane type, date and crash location, the extracted image data lacks this context. Without hyperbole, the model could learn to confidently classify pictures of middle-aged men in ceremonial uniforms as aeroplanes.

Table 1 Number of images per class in the training dataset

Category	*Number*
Infantry fighting vehicles	17700
Tanks	11273
Transport	7260
Infantry mobility	3100
Self-propelled artillery	2768
Command posts, communication	1327
Anti-aircraft systems	1224
Towed artillery	1177
Drones	1145
Engineering	1073
Multiple rocket launchers	872
Radars, jammers	572
Airplanes	385
Helicopters	358
Anti-tank systems	226
Other	158
Vessels	99

Given the overall noisiness of the data and the relatively limited number of observations, it was decided to focus on classifying merely the broad types of vehicles – tanks from infantry fighting vehicles or artillery. Even this approach may not prove to be very accurate, and while there may be enough data from some models (such as T-72 tank variants or BMP-2), it is not clear how that would provide further relevant insights.

Based on general heuristics, only the five most populous classes – Infantry Fighting Vehicles (IFV), Tanks, Transport, Infantry Mobility (IM) and Self-propelled Artillery (SPG) – appeared to contain enough data required by the experiment. The remaining classes were, therefore, excluded from the rest of the analysis.

4.2 Testing Data

4.2.1 Collection

To realistically assess the model's capability to generalise, it needs to be tested – just once, after all the finetuning takes place – on previously unseen data. If there is any contamination of the

model by the testing data, the final metrics will provide little insight into how the system will perform in deployment. Therefore, the initial plan was to train and evaluate the model on the dataset sourced from Warspotting and then collect testing data close to the project's finish date to better simulate deployment in changed conditions (Canty, 2019).

However, as other parts of the project were delayed, the author got to gather the testing data at a point when requesting more data or approval from either the Warspotting or Oryx team was not feasible. It also became apparent that collecting the data manually (as the OSINT teams do) would be incredibly time-consuming.

Therefore, a decision was made to use the Oryx website to source a sample of losses that occurred since the beginning of October 2023. The three-month window between the training cutoff and the start of testing was chosen to rule out contamination due to delays in processing between the two databases. Also, this window increases the distribution shifts between the training and testing data – over the three-month period, the situation on the battlefield has markedly changed.

Given the open presentation of the data and lack of any guidance regarding its reuse, obtaining it from the Oryx team without explicit approval was deemed morally justifiable. Unlike Warspotting, Oryx does not host the images on its server and uses third-party storage instead. Thus, all download links could be scraped from the Oryx webpage in one static load of the page. Any concern that the scraping could negatively impact the website was thus averted. This way, some 950 images were obtained.

4.2.2 Processing

During their processing, several scenarios right out of Sambasivan's paper appeared. First, the Oryx database contains relatively more pictures of multiple vehicles present, sometimes with one of them outlined by a drawn square. Unlike on the Warspotting website, this square is part of the .jpg file, so it is not trivial to perform the automated cropping at scale, and the model is not robustly trained on such images. Also, the Oryx team sometimes combines two images (pre-strike and explosion) into one. Nevertheless, these images may be interesting to analyse as edge cases, investigating whether the model classifies correctly based on some of the vehicles or gets completely confused.

Fig 5 A typical combined image used by Oryx

A second issue that became apparent is that both teams (despite overlaps in their members) group vehicles differently. Combining aeroplanes and helicopters into one category causes no major problems, as these classes are too small. The Warspotting team, including both infantry mobility and mine-resistant ambush-protected vehicles in the larger infantry mobility category, makes no meaningful difference and can be easily remedied by combining the two categories.

It became apparent that the Warspotting team bunches up all *armoured* fighting vehicles (AFVs) except for tanks under the category of *infantry* fighting vehicles (IFVs). First, this is conceptually incorrect. As the debates following the French delivery of AMX-10 RC to Ukraine showed, the discussions of what constitutes a tank, IFV, or other type of equipment can get abstracted.

However, a Ukrainian commander is likely to have a very different reaction to learning that the enemy is concentrating MT-LBs as opposed to BMP-3s. The MT-LB is considered an armoured

Fig 6 A visual comparison of an MT-LB and BMP-3 (both images from the WarSpotting database)

personnel carrier (APC), is lightly armed and armoured, and its main role is to provide infantry mobility with protection against light weapons. A BMP, the typical example of an IFV, can carry fewer troops but is much better armed, has better protection and is able to provide direct combat support. Therefore, including APCs under the IFV category is not merely a matter of semantics.

This is a typical example of a data cascade. At first, handling the data was seen as an engineering task, cropping, merging, filtering, uploading and categorising it so that a first iteration of the model could be trained and evaluated. At the point this issue became visible, going back and separating the categories becomes a daunting task. The model is now learning from arguably incorrect labels due to an "as is" handling of the available dataset. Given the expertise of Warspotting's team, it can be assumed that some similar legacy reasons are responsible for this discrepancy.

4.3 Training Data Split and Data Transforms

In line with common best practices, training data was split into training and evaluation subsets. The 80/20 split was performed using *sklearn's train_test_split* module. The process ensured that the ratio was applied within each class and that images related to the same observation were kept together. The testing data was acquired subsequently, with the total amount and class balance depending on the data available.

Fig 7 A sample of the data being fed into the model, past transforms. Notice that the vehicles are cropped off - this helps reduce the overfitting of the model.

In line with standard practices to improve performance and reduce overfitting, the training data was randomly resized and cropped to the shared size of 224x224 pixels, with eval and test data being resized and cropped along the edges.

4.4 Model and Training

As training neural networks from scratch is resource-intensive, transfer learning helps researchers deploy model faster by taking a model pretrained for a similar task and then only adjusting it to the task and data at hand. The widely-used ResNet18 (He et al., 2016), pre-trained on the ubiquitous ImageNet dataset (Deng et al., 2016), was chosen. This would be the default choice for most researchers running a similar experiment, which speaks to the approach's fit. Also, taking the most common approach helps make the observations more generalisable. The model was trained using the *pytorch* library.

A neuron in a neural network is a basic computational unit that takes inputs, assigns them weights, sums the weighted inputs, passes the sum through an activation function, and outputs the result.

In simple terms, a neural network is made up of interconnected layers of neurons - mathematical functions. Weights (measures of relative importance) are applied to the inputs of each neuron, and the sum of those inputs is then passed through an activation function of the neuron. The resulting output is then passed on to neurons in the next layer. As the network receives feedback on the outputs, the weights are adjusted accordingly and tested again. The weights of a neural network are the most important and valuable asset, closely guarded by for-profit organizations.

In transfer learning, all layers but the final one are usually kept, and their state is frozen. The final layer is replaced by a new one, with the number of neurons equivalent to the number of classes to be predicted. The weights of those neurons are randomly initialised so that they can be trained on the available data.

Describing other steps in the training, such as model configuration and hyperparameters setting, is out of the scope of the thesis. Their values were mostly kept to default or best-practice settings and can be found in the attached jupyter notebook.

4.5 Experiments and Evaluation

Out of curiosity, several ablations were trialled. These refer to removing or disabling parts of the model to see how its performance would be affected. The first one was turning off the learning scheduler. The model training takes place in rounds known as epochs. Usually, the learning rate at each epoch is adjusted by a learning scheduler to optimize performance. For example, early epochs have a high learning rate to speed up the convergence of the model, with later epochs impact being limited to avoid overfitting. The second ablation was removing the weights pre-trained on the ImageNet1k dataset. Thus, the model was trained from scratch, removing the benefits of transfer learning.

Once the most suitable model was found, it was tested with test data, on which it has been neither trained nor evaluated. While preliminary evaluation metrics provide guidance on model configuration and optimisation, test data provides a much better sense of how the model would perform once deployed. In this case, the test data comes from a later period of the war, which will highly likely decrease the performance of the model due to distribution shifts.

The performance of the final model on both validation and test data was then assessed by a range of quantitative and qualitative means. These include the confusion matrix and per-class accuracy. Qualitative means entail the analysis of visualised errors in predictions, as well as speculating on potential causes behind anomalies and the results.

5 Analysis

5.1 Performance on Validation Data

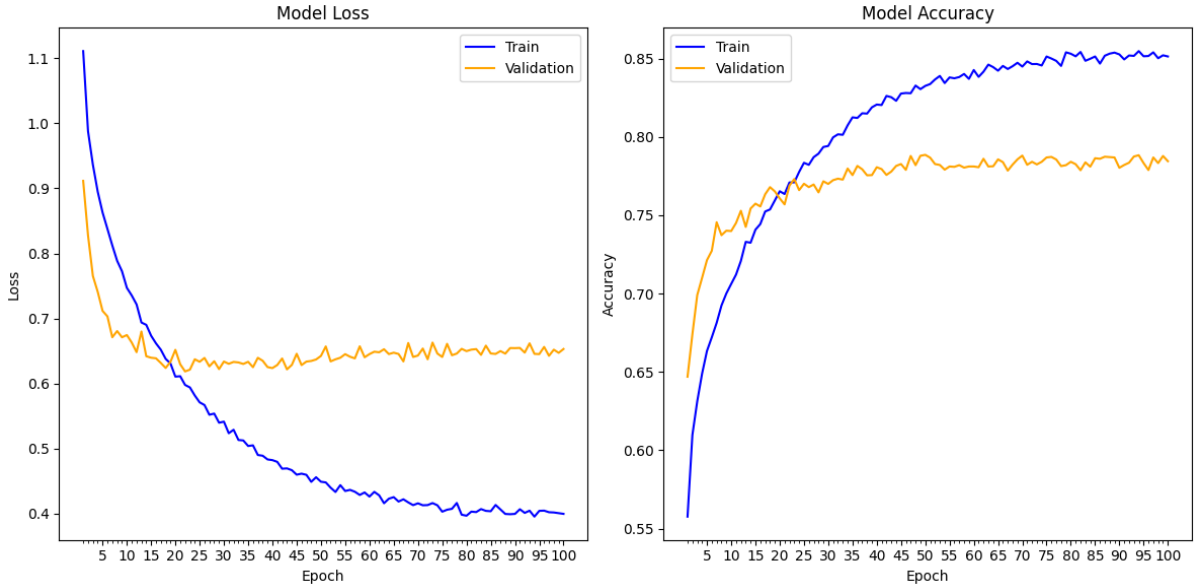


Fig 8 Full model's loss and accuracy on evaluation data

The best-performing model achieved an acceptable accuracy of 78.9% on evaluation data. Turning off the learning scheduler led to a barely noticeable drop in performance while removing the pre-trained weights decreased the accuracy by almost 15%. This demonstrates the importance and usefulness of transfer learning in image classification.

Table 2 Comparing the accuracy of the full model to ablated models

Model	*Best Achieved Accuracy*
Full	0.7886
No Scheduler	0.7832
No Pretrained Weights	0.6430

A closer look at the performance, along with the confusion matrix, can be found in the Disussion chapter.

5.2 Performance on Test Data

The full model correctly identified 61.9% of the 880 test images. That is hardly an impressive performance, and the discussion section will attempt to uncover the potential reasons behind this. On a more positive note, the model at least appears to be aware of its limitations. As Fig 9 shows, the model was more confident when it made predictions that turned out to be right,

and vice versa. This means that the model made few confident mistakes and demonstrated uncertainty in situations in which classification was precarious.

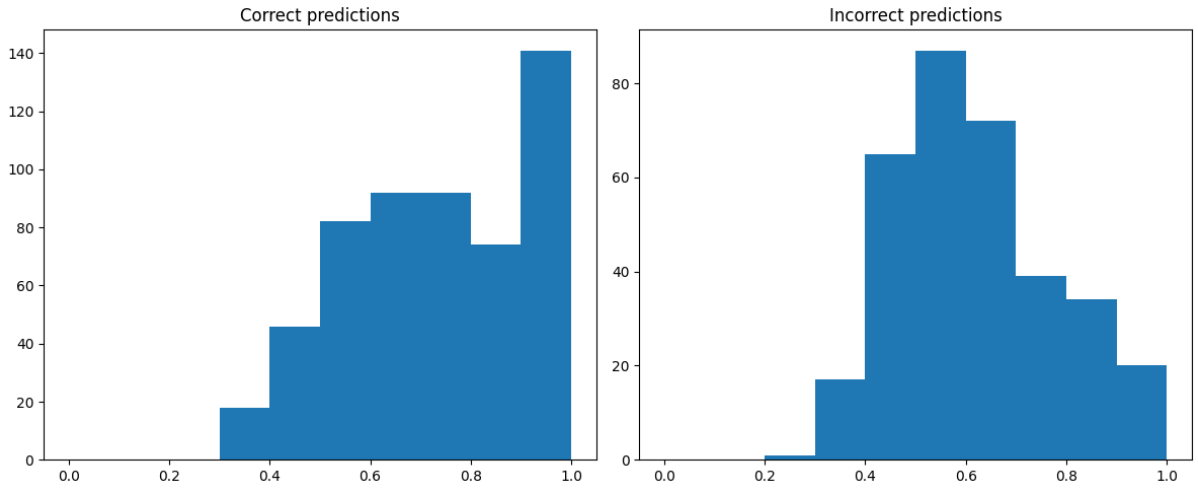


Fig 9 The full model's confidence in correct and incorrect predictions on test data

Table 3 compares the number of samples in both training and test datasets. As is evident from the table, the performance has dropped significantly for all classes, with the lowest decrease for SPGs and by far the largest for Infantry mobility, of which only 19 images were present in the testing data.

Table 3 Comparing class performance on evaluation and test data

Class	*Train samples*	*Test samples*	*Eval accuracy*	*Test accuracy*
IFV	17700	408	0.828	0.696
Tanks	11273	251	0.776	0.673
Transport	7260	129	0.745	0.527
Infantry mobility	3100	19	0.608	0.157
SPG	2768	73	0.370	0.287

5.3 Model's Performance Assessment

In this section, the results will be analysed. While some lessons may appear relevant, it is important to note that post hoc explanations are prone to all kinds of biases. Therefore, it cannot be ruled out that the same facts could be interpreted in an entirely different way if the results require so.

Judging the model's performance is difficult. In isolation, the evaluation performance metrics are not overly impressive. At the same time, a visual examination of the incorrectly classified evaluation images shows that a clear majority of those are rather problematic images.

For example, images containing several vehicles of different classes, showing merely a huge explosion, or where the vehicle is shown so small and blurry that identification could have been done only based on the context or due to the analyst's supreme attention to pixel-level detail. At least on validation data, the algorithm appears to perform well on samples that a follower of the war could easily distinguish. The performance then drops off rapidly in cases where even a trained human would struggle.

Based on this, one might be tempted to stretch the model's limits for more ambitious classification tasks, such as between different tank models. Seeing the performance on train data, however, appears to signal that even the current task proved to be a challenge for the model.

5.4 Performance Drops on Test Data

While significant drops in the classification performance were to be expected, the drop from around 79% to 62% accuracy is a serious issue. Could this have been prevented by implementing more measures that counter overfitting, such as data augmentation or early stopping? Possibly, but these measures alone would be highly unlikely to account for the difference in performance – other factors are clearly at play as well.

It is impossible to tell what proportion of the performance drop is due to changes in the objective reality the data captures – changing battlefield, weapons, image capture methods, etc., or due to the different sources of data. A very interesting follow-up to this experiment would involve collecting test data from the period both from Oryx and WarSpotting and then comparing the model's performance on both test datasets. This could, in theory, control for the variation in the collection and processing of the (largely the same) data and show the significance of reality-based distribution shifts.

This comparison would be especially interesting given that the subset of Oryx data used for testing most likely comes from the analyst Naalsio, who is a part of both the Oryx and WarSpotting teams. This is apparent in the different file naming conventions used by each analyst, which are preserved when uploaded and linked on the Oryx webpage. Naalsio usually

includes the date in the filename of each picture, which allowed for easier scraping while ensuring that losses documented prior to October 2023 were not involved. While Oryx analysts do not appear to split the workload according to any obvious heuristic, it cannot be ruled out that featuring losses processed by a specific analyst may have introduced further biases.

The main difference between the two databases appears to be the way each handles multiple images. While Oryx only features one image per list entry, WarSpotting usually adds multiple. This distinction makes little difference for burn-out wrecks captured from the ground but becomes important for strikes captured in real time. To show both the vehicle and the fact that it has been destroyed, at least two pictures need to be shown, usually *before and after* stills from drone-shot videos. Alternatively, the final pre-impact frame transmitted by an FPV drone is accompanied by an image of the explosion captured by a reconnaissance drone.

As the Oryx team works with one image per list entry, they combine multiple frames of the video into one picture. At times, up to 4 such images are stacked together to create an ultrawide image. The model has seen much fewer combined images of this sort in training.

Another factor that clearly contributed to the drop in performance is the shift from ground-shot photos to drone imagery being dominant. Looking at the training and evaluation data, a rough estimate of 70 - 80% of the images are captured from the ground. This starts with virtually no aerial footage at the loss-heavy beginnings of the war, with its proportion eventually increasing over time. An equally rough look at the test dataset reveals that at least 80% of losses there are shot from a drone. This is hardly surprising, as the battlelines became more static and drones proliferated.

5.5 Analysing a Sample of Misclassifications

To get a better idea of where the model made mistakes, a random sample of 90 misclassified images was manually examined by the author (who claims no expertise but has general familiarity with the topic). Out of these, the author was able to confidently correctly classify 59%, most of which were trivial. The remaining 41% comprised of images where the vehicle was hardly visible or was in such a state of destruction that correct identification required a deep knowledge of the equipment.

Supporting the point of prevalence of aerial footage, 87% of the misclassified images in the sample were shot from a drone. 40% of the sample are amalgams of 2 or more images, with only 3 out of 36 being collages apparently made by the source of the image. In that case, the image would have likely been kept this way in the WarSpotting dataset as well (Fig 10).

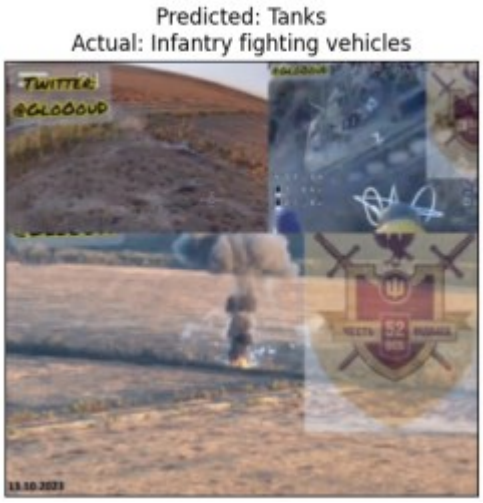


Fig 10 An example of an image provided by the original source.

In four *misclassified* cases, the model correctly identified one of the vehicles in the picture. In two of those, the labelled vehicle was enclosed in a yellow square the model did not recognize. The context of another image (Fig 12), a large portion of which is taken up by a tank, from which a picture of a wrecked IFV is taken, is clear to a human, but the model identified the tank rather than the IFV. In the remaining case, neither the author nor the model could determine which vehicle is supposed to be the centre point of the image (Fig 11).

Fig 11 The model correctly identifies the tank in this picture, but the image is labelled based on the other two IFVs present.

Fig 12 An example of the model lacking the context clear to any human.

For reference, a sample of 32 misclassified images is provided in Appendix 1: A Sample of Misclassified Test Images.

5.6 Class Performance

5.6.1 Validation Data

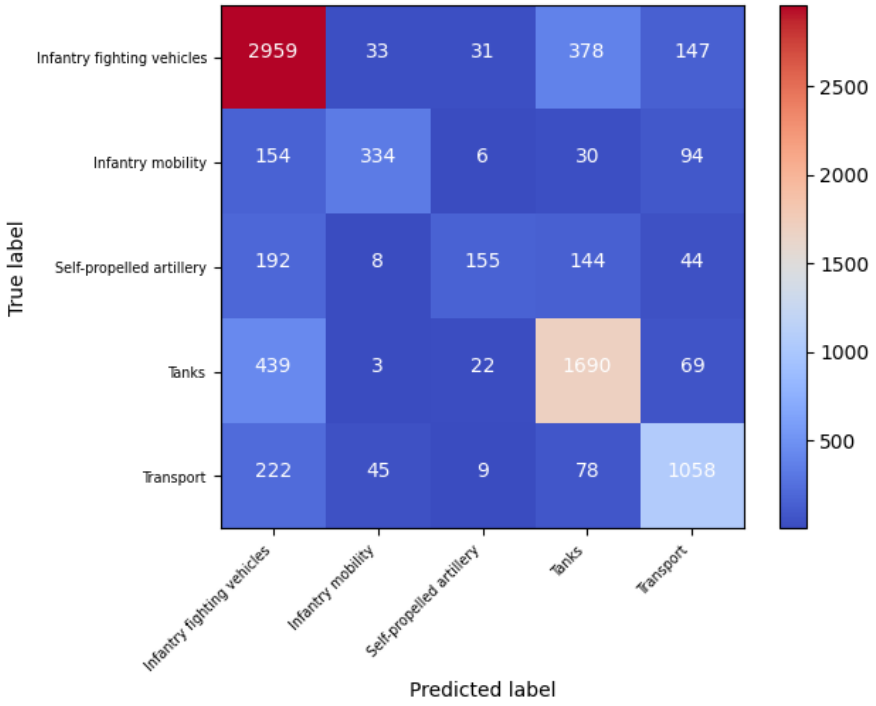


Fig 13 Class confusion matrix for evaluation data

Given the model's poor performance on test data, it appears valuable to take a closer look at the class performance results for the evaluation data, too.

5.6.1.1 Better Performance on Larger Classes

The first observation is that the more images were available for one of the five classes, the better accuracy was achieved. That is hardly surprising, as having more data available allows the model to classify the class better. Also, without special interventions, the model learns to choose the most frequent class, as it is statistically the most likely to be the right answer.

From a more speculative perspective, IFVs having the highest accuracy come with little surprise. Tanks are also well represented in the training data, and their features are intuitively quite different from IFVs. Transport is also distinctive. However, a sharp dropoff comes for Infantry Mobility. The number of available images is less than half of transport, so even its distinctive features may not quite make up for the lack of data. To make matters better, IM is most often confused with IFVs (the largest class) or transport (wheeled, lightly armoured vehicles). Confusions with tanks or SPGs are rare, which makes intuitive sense.

5.6.1.2 Poor Accuracy on SPGs

The very poor performance of SPGs, even when compared to the similarly-sized IM class, warrants a closer investigation. Since SPGs operate further from the frontline than IFVs or tanks, they are most likely to be destroyed in this static phase of the war by either counter-battery fire or loitering munitions. The first case will be likely captured by a surveillance drone from afar, leading to low-quality images. Footage from loitering munitions may be of better quality but is likely to be out of the distribution of older training data.

Another aspect is that all common Soviet SPGs are based on tank or IFV hulls (Fig 14). This class shares many visual features with the rest. This effect is further reinforced by the fact that ammo tends to be stored in the main fighting compartment in Soviet/Russian SPGs. Direct hits can then often lead to a catastrophic explosion, which catapults the gun mount away. According to WarSpotting, 39% of all Msta-S SPG losses were found without a turret (WarSpotting, 2023d). All that then remains in the frame are burnt remains of what is a tank or IFV hull. The T-80 tank, on which the Msta-S is built, has the turret blown off in 52% of documented cases (WarSpotting, 2023b).

Fig 14 A turretless Msta-S, with only the T-80 hull being apparent.

5.6.2 Test Data

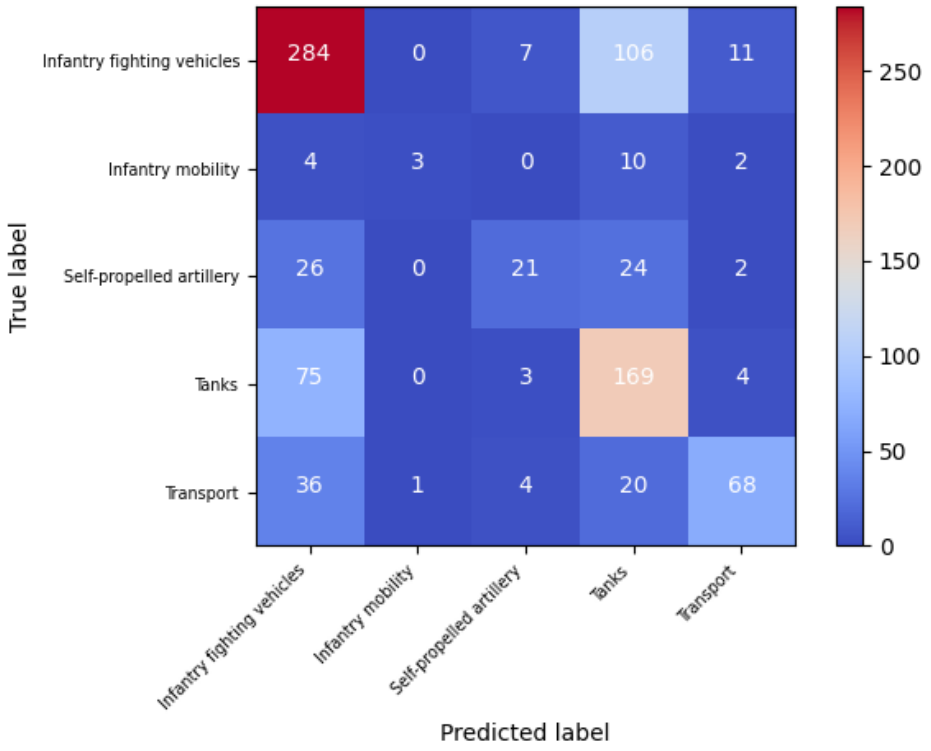


Fig 15. Class confusion matrix on test data

On test data, the top three classes with the most data available were also the ones with the best performance. Unlike on evaluation data, however, SPG classificaiton was more accurate than infantry mobility. While the SPG accuracy dropoff is largely what would be expected with the overall decrease in predictive performance, the 16% accuracy on infantry mobility is abysmal and requires a closer look.

5.6.2.1 Abysmal Accuracy on Infantry Mobility

Transport and infantry mobility were the two most difficult classes for the author to distinguish. Tanks, IFVs, and SPGs can easily be identified by the distinctive shape of their hull or turret, which come in very few shapes (due to manufacturing and design reasons). Transport and infantry mobility, on the other hand, are both often based on civilian designs and come in many shapes and forms.

If anything, the model's acceptable performance at distinguishing these on evaluation data came as a surprise. It is possible that the model may have overfit on this minor class in the heavily imbalanced dataset. That would lead to decent performance on evaluation data but weak performance on test data.

Interestingly, there were only three confusions between Transport and IM on the test data. While that may appear strange given the many intuitive similarities between the vehicles, the class imbalances leading to a bias towards IFVs and Tanks may explain this observation.

The testing dataset involved 19 examples of this class. Out of these, 13 were aerial images, with the remaining 6 being heavily burnt out. Furthermore, 7 of the 19 images contained easily identifiable vehicles of other classes. The author could confidently classify 13 of the vehicles, having to rely on the ability to understand red rectangles surrounding a vehicle in 6 of those (Fig 16). Given the limitations of the data, the low performance is understandable. Also, the small sample reduces the statistical significance of the metrics.

Fig 16 An example of a misclassified image also featuring an IFV.

5.6.2.2 Shift Towards Aerial Footage

Another thing to note is the much higher rate of confusion between IFVs and Tanks than on evaluation data. The obvious culprit would be the model learning to distinguish those based on features mainly visible on ground pictures and thus failing to differentiate between them on aerial imagery. To test this hypothesis, incorrectly classified observations of tanks were manually inspected, and it was found that only 8 out of 82 (9.8%) of the total misclassified images were ground-shot. This rate is lower than the total representation of groud-shot images in the Tank test data (44 out of 251, or 17.5%).

However, the difference is not statistically significant when a z-test is applied. Additionally, lower performance on the generally worse-quality drone shot pictures would be expected, even if both capture methods were equally represented in the training data. It is important not to overanalyse the data. Highly likely we are just seeing a poorly performing model making mistakes in the classification of the two most populous classes, which is to be expected.

5.7 Dataset Imbalance

Of the 17 groups in the training dataset, 12 had to be excluded due to sufficient data not being available. Even if the absolute counts of all classes were higher, the model would still likely perform poorly due to the inherent challenges of heavily imbalanced datasets and the increased challenge of discerning between more classes.

The current paradigm works best on the most common losses. Whether that is a good thing depends both on one's opinion and the system's intended application. On the one hand, it can be argued that an automated system should perform the best on the most common vehicles, as it would help with the bulk of the workload.

On the other hand, it can be assumed that an analyst would have to verify the outputs anyway. In that case, being able to identify ubiquitous vehicles such as BMP-2 is not much help, as the analyst is likely to be very familiar with the system. Human analysts might then prefer a system that helps identify uncommon systems (vessels, electronic warfare, uncommon models) that they may not immediately recognise.

Finally, by definition, the most impactful events in the world are rare – major natural disasters, nuclear wars or large-scale terrorism. Given that the current ML paradigm requires a lot of data to make inferences, this sets a clear limit to how much of our analysis and awareness we can delegate to those systems. The usefulness of ML in many of the relevant tasks cannot be denied, but the notion that ML can solve all our (open source) intelligence problems requires serious scrutiny.

5.8 Excluded Classes

A whole range of important military equipment is missing from the analysis. In addition to the classes excluded due to their insufficient representation in the dataset, it is important not to forget about the materiel that never made it to the list in the first place.

It is good news that even the largest war in Europe since World War 2 does not provide enough data for the models used in this experiment. Accustomed to small-scale insurgency conflicts and expecting a shorter war, the Oryx team was initially tracking losses and captures of smaller drones and anti-tank guided missiles. Seeing the war unfold, the team decided to stop tracking these and has not started tracking loitering munitions after they were introduced on the battlefield.

Another notable category that is not covered by either dataset is civilian vehicles employed by the belligerents. That decision makes sense – how would the analysts decide whether a destroyed SUV was civilian or used for military purposes? Nevertheless, it leaves out another important piece of the battlefield puzzle. The Ukrainians have been vocal about the importance of whatever mobility is available, both to evacuate casualties and to supply the troops.

Many more features of war, such as ammunition, fuel or missiles, are, for obvious reasons, not tracked in the datasets. Unfortunately, this might mean an implicit encouragement of the natural bias towards large weapon systems. Even a casual follower of the war has by now picked up on the fact that the availability of the unexciting 152mm and 155mm artillery ammunition is possibly the most important factor in the war. Securing a robust supply of those shells would almost certainly make the Ukrainian general staff happier than scores of the newest Western tanks or jets.

With the best of intentions, Czech citizens have been collecting money for flashy systems, with the latest effort being a fundraiser of 100 million CZK for a used UH-60 Black Hawk helicopter (Zbraneproukajinu, 2023). However, the cost-benefit analysis of similar investments appears rather unfavourable. Fortunately, the much more sensible effort to fundraise a similar amount for 10,000 FPV drones signals a shift against this tendency (Gruntová, 2023).

5.9 Data Quality

When inspecting the datasets randomly, one is amazed by their size and the care with which they are curated. From that perspective, the data appears perfect, and one would be forgiven to think that these could be the good-quality dataset all ML practitioners crave. However, when the errors are analysed, one can suddenly see all the data issues that may not be readily apparent to a human. An ML algorithm has no way of knowing that a picture was presented in the context of another or that there is an arrow pointing to one of the vehicles. This provides a great lesson in the unpredictability and hidden pitfalls of data.

The performance of the model could likely be improved by better data pre-processing. All pictures could be cropped so that they only show the vehicle in question, leaving out the surrounding environment and other vehicles. Alternatively, aerial photos could be separated from images shot on the ground. Unfortunately, given the volume of the data, these steps would again take a considerable amount of time.

Alternatively, ML models could be trained or repurposed for this task. Even then, some training data would need to be available, and there is no reason to assume these would work perfectly. Even in the best-case scenario, deploying and training these helping models is time and resource-intensive.

Fig 17 Minefields, static battle lines and the distance from which drones capture the battlefield lead to a significant portion of recent images containing more than one vehicle.

5.10 The Author Wanted to do the Model Work

The message of Sambasivan et al. was deeply felt by the author. The data was wrangled, organised and uploaded to Google Drive to be available for the Google Colab platform. After the first version of the model was evaluated, there were multiple paths to take. Given the model's decent performance on intuitively reasonable tasks (images where the vehicle is clearly visible), the reasonable next step might be to try distinguishing between common specific models, for which plenty of data was available. Another option would be to include the smaller classes or manually curate the eval or the eventual test dataset to include only images that a human would deem possible to identify.

However, all these would require parsing the labels, integrating classes from two different datasets, reuploading the images, and potentially even going manually through them. Trying different ablations, modifying the number of epochs, and other fine-tuning takes just a line of code, and one can wake up the next morning to see the quantitative metrics go up a little.

Therefore, the hypothesis that data is still the main issue appears to hold. ML can aid in processing and analysing it, but it cannot solve deep data issues, and it can also abstract these issues away, leading to less robust analysis. Another interesting point is that ML forces the researcher to work with datasets that are too big for manual analysis. One could see this as a

Faustian bargain – we are offered unseen benefits but must give up our ability to understand the data ourselves.

5.11 General Usefulness

As presented, it is difficult to think of how this model could meaningfully contribute to the broader effort of vehicle loss tracking in the war in Ukraine. Even if it performed flawlessly in the experiment, or even some more challenging version thereof (such as distinguishing common tank variants), it would merely match humans at one of the least difficult and laborious parts of the process. An experienced analyst would recognize the vehicle type instantly, along with further details.

The much more challenging steps follow: identifying the specific variant of the vehicle, checking whether the exact vehicle has not already been captured from a different angle or even moved, assessing whether the vehicle is merely damaged or destroyed and making sure it is not a decoy. The model presented helps in none of these and appears very far from being able to do so.

The model's performance is limited by the quality of the available dataset. One of the lessons of this experiment is that using data just because it is available may not always be the best approach. Instead, the suitability of data for a given application should be carefully weighed, comparing the cost of post-processing, wrangling and potentially reduced performance to the benefit of availability. The counterfactual of collecting or curating a new dataset tailored to the use case should be considered.

5.12 Potential Alternative Applications

What this approach lacks in accuracy and attention to detail, however, can be made up for by speed and scale. A different use case can be envisioned, where the model would be trained to analyse large image datasets to identify military vehicles, count them, and output summary statistics. Providing an automated alert that 40 pieces of what are likely IFVs have just disappeared from one of the many Russian military bases could be helpful. This kind of monitoring is also something that would be very time-consuming and unscalable when done by humans.

This capability appears to exist already – the companies Orbital Insight and SpaceKnow claim to be able to automatically identify and track military vehicles from commercially available satellite imagery. Government agencies of major powers are likely to have this capability as

well. Unlike a constellation of random Telegram posts and videos, the images produced by the satellites are broadly consistent, and training on them would lead to better results.

Currently, this approach requires access to a rich archive of past imagery with well-labelled examples of objects of interest. An active subscription to the satellite imagery, as well as considerable computational resources for training and rolling analysis, are also needed, along with decent ML research expertise. Most of these are not currently in the reach of individual volunteers, but it is conceivable this might change in the future as the availability of computing, satellite imagery, and helper models increases.

6 Conclusion

This thesis explored the potential of using machine learning models in OSINT analysis. Data was seen as the uniting factor of the two fields, with a large volume of it being a challenge for human analysts and a requirement for most ML techniques. The example of documenting destroyed military equipment in the Ukraine war was used to explore the topic. Following an extensive literature review on the sources and limitations of OSINT data, a supervised image classification model was trained on a subset of the WarSpotting database compiled by volunteers.

The model achieved decent accuracy in distinguishing between broad vehicle categories like tanks and infantry fighting vehicles on the evaluation dataset. However, performance dropped significantly on testing data that came from a different source and captured a later period of the war. First, the new source often combined multiple images into one, which the model has encountered only rarely in training. Second, the test data, collected months later, reflected the changing realities of the battlefield not represented in the training data. The prominence of aerial drone footage presented unfamiliar perspectives to the model.

While many of these issues have been largely predictable, this experiment serves as a valuable demonstration of the potential issues and limitations of applying ML in dynamic environments. In addition to the challenges posed by the everchanging reality on the battlefield, more universal issues of data were presented. These include the challenge of obtaining suitable data and preprocessing it for use with the ML algorithm. Data cascades, where early sloppiness with the data leads to later more significant issues, have not been avoided by the author, just like the human preference for tweaking the model over going back to the data.

Therefore, the experimental results appear to be in line with the main hypothesis of the thesis. Despite the fact that the specific data and application were shown to be largely unfeasible, ML holds great promise for aiding in OSINT analysis, just as it does for all other parts of our society. The increasing availability of data, computing resources, and helper models will likely lead to ML being more widely applied to OSINT analysis in the future.

Nevertheless, the need to gather, process, and handle the data will always be present, as well as the challenges of bad quality or missing data, data bias, adversarial attacks, and distribution shifts. Merely deploying larger or more advanced models does not solve these issues.

Table of Figures

References

1. Africk, B. (2023, November 8). *Russian field fortifications in Ukraine*. https://read.bradyafrick.com/p/russian-field-fortifications-in-ukraine
2. Alzubi, J., Nayyar, A., & Kumar, A. (2018). Machine Learning from Theory to Algorithms: An Overview. *Journal of Physics: Conference Series*, *1142*(1), 012012. https://doi.org/10.1088/1742-6596/1142/1/012012
3. Armstrong, S., Sotala, K., & Ó hÉigeartaigh, S. S. (2014). The errors, insights and lessons of famous AI predictions – and what they mean for the future. *Journal of Experimental & Theoretical Artificial Intelligence*, *26*(3), 317–342. https://doi.org/10.1080/0952813X.2014.895105
4. Bartošová, K. (2023, May 25). Disney's parks in Florida from Space—SpaceKnowSpaceKnow. *SpaceKnow*. https://spaceknow.com/blog/disneys-parks-in-florida-from-space/
5. BBC News (Director). (2022, February 24). *CCTV shows tanks and Russian military vehicles cross Ukraine border—BBC News*. https://www.youtube.com/watch?v=HaJCwY9YBQ0
6. Bellingcat. (2018, October 9). *Full report: Skripal Poisoning Suspect Dr. Alexander Mishkin, Hero of Russia*. Bellingcat. https://www.bellingcat.com/news/uk-and-europe/2018/10/09/full-report-skripal-poisoning-suspect-dr-alexander-mishkin-hero-russia/
7. Bellingcat Investigation. (2020, December 14). *Hunting the Hunters: How We Identified Navalny's FSB Stalkers*. Bellingcat. https://www.bellingcat.com/resources/2020/12/14/navalny-fsb-methodology/
8. Biljecki, F., & Ito, K. (2021). Street view imagery in urban analytics and GIS: A review. *Landscape and Urban Planning*, *215*, 104217. https://doi.org/10.1016/j.landurbplan.2021.104217
9. Block, L. (2022, April 11). *OSINT - Demystifying the Fog of War?* https://www.leidensecurityandglobalaffairs.nl/articles/osint-demystifying-the-fog-of-war
10. Bump, P. (2021, October 23). Analysis | Here's why the resolution of satellite images never seems to improve. *Washington Post*. https://www.washingtonpost.com/news/politics/wp/2017/04/21/heres-why-the-resolution-of-satellite-images-never-seems-to-improve/
11. Canty, M. J. (2019). *Image Analysis, Classification and Change Detection in Remote Sensing: With Algorithms for Python, Fourth Edition* (4th ed.). CRC Press. https://doi.org/10.1201/9780429464348
12. Christian, B. (2020). *The Alignment Problem: Machine Learning and Human Values*. W. W. Norton & Company.
13. Clausewitz, C. von. (1950). *On War*. Jazzybee Verlag.
14. Coakley, A. (2021, November 15). *Borderline: Tinder profiles of Polish troops appear in Belarus*. The Independent. https://www.independent.co.uk/news/world/europe/belarus-poland-border-tinder-troops-b1957953.html
15. Cremer, D. D. (2020, September 3). What Does Building a Fair AI Really Entail? *Harvard Business Review*. https://hbr.org/2020/09/what-does-building-a-fair-ai-really-entail
16. Cusack, B., & Tian, Z. (2017). Evaluating IP surveillance camera vulnerabilities [PDF]. *Australian Information Security Management Conference*. https://doi.org/10.4225/75/5A84EFBA95B46

17. Deng, X., Liu, Q., Deng, Y., & Mahadevan, S. (2016). An improved method to construct basic probability assignment based on the confusion matrix for classification problem. *Information Sciences*, *340–341*, 250–261. https://doi.org/10.1016/j.ins.2016.01.033
18. Drazdovich, U. (2023). *Words and Actions: Understanding Russia's Information Security Strategy - ProQuest*. https://www.proquest.com/openview/280af9ca3d9a74c3cf0d94285f94244a/1?pq-origsite=gscholar&cbl=18750&diss=y
19. ETI. (2022). *Geolocation Methods: A step by step guide—The Kit 1.0 documentation*. https://kit.exposingtheinvisible.org/en/geolocation.html
20. Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., & Song, D. (2018). Robust Physical-World Attacks on Deep Learning Visual Classification. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1625–1634. https://doi.org/10.1109/CVPR.2018.00175
21. Farberov, S., & feed, G. author R. (2022, April 19). *Ukrainian millionaire Andrey Stavnitser asked military to bomb Russian-occupied mansion*. https://nypost.com/2022/04/19/ukrainian-millionaire-andrey-stavnitser-asked-military-to-bomb-russian-occupied-mansion/
22. Fiorella, G. (2022, November 23). *How to Maintain Mental Hygiene as an Open Source Researcher*. Bellingcat. https://www.bellingcat.com/resources/2022/11/23/how-to-maintain-mental-hygiene-as-an-open-source-researcher/
23. Ganguly, M. (2022). *THE FUTURE OF INVESTIGATIVE JOURNALISM IN THE AGE OF AUTOMATION, OPEN-SOURCE INTELLIGENCE (OSINT) AND ARTIFICIAL INTELLIGENCE (AI)*.
24. Ghioni, R., Taddeo, M., & Floridi, L. (2023). Open source intelligence and AI: A systematic review of the GELSI literature. *AI & SOCIETY*. https://doi.org/10.1007/s00146-023-01628-x
25. Gonzales, C. (2022, October 4). *Scorched Earth: Using NASA Fire Data to Monitor War Zones*. Bellingcat. https://www.bellingcat.com/resources/2022/10/04/scorched-earth-using-nasa-fire-data-to-monitor-war-zones/
26. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). *Explaining and Harnessing Adversarial Examples* (arXiv:1412.6572). arXiv. http://arxiv.org/abs/1412.6572
27. Google. (2023). *Google Earth Help*. https://support.google.com/earth/?hl=en#topic=7364880
28. GOV.UK. (n.d.). *GOV.UK Documentation Example*. Data.Gov.Uk. Retrieved 28 December 2023, from https://guidance.data.gov.uk/guidance.data.gov.uk/
29. Graham-Harrison, E. (2023, September 4). 'A psychological weapon': Inside a Ukrainian factory making decoy kit. *The Guardian*. https://www.theguardian.com/world/2023/sep/04/a-psychological-weapon-inside-a-ukrainian-factory-making-decoy-kit
30. Gruntová, K. (2023, December 18). *Český spolek chce poslat na Ukrajinu deset tisíc dronů. Jeho patronem je šéf armády Řehka*. iROZHLAS. https://www.irozhlas.cz/zpravy-domov/ukrajina-pomoc-dron-armada-karel-rehka-spolek-skupina-d_2312180620_gut
31. Hastie, T., Friedman, J., & Tibshirani, R. (2001). *The Elements of Statistical Learning*. Springer. https://doi.org/10.1007/978-0-387-21606-5
32. He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep Residual Learning for Image Recognition*. 770–778. https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html

33. Hendrycks, D., & Dietterich, T. (2019). *Benchmarking Neural Network Robustness to Common Corruptions and Perturbations* (arXiv:1903.12261). arXiv. https://doi.org/10.48550/arXiv.1903.12261
34. Higgins, E. (2021). *We Are Bellingcat: An Intelligence Agency for the People*. Bloomsbury Publishing.
35. Hofstadter, D. (1999). Gödel, Escher, Bach: An Eternal Golden Braid. *Books*. https://commons.library.stonybrook.edu/library_books/1
36. Humphreys, A., & Wang, R. J.-H. (2018). Automated Text Analysis for Consumer Research. *Journal of Consumer Research*, *44*(6), 1274–1306. https://doi.org/10.1093/jcr/ucx104
37. Jadrný, P. (2022, October 27). *U Chersonu jsou zbytky lepších jednotek. Jejich ztráta by Rusku znemožnila ofenzivu, říká expert Janovský*. iROZHLAS. https://www.irozhlas.cz/zpravy-svet/ukrajina-cherson-tank-ofenziva-oryx-janovsky_2210270700_pj
38. Jakhar, D., & Kaur, I. (2020). Artificial intelligence, machine learning and deep learning: Definitions and differences. *Clinical and Experimental Dermatology*, *45*(1), 131–132. https://doi.org/10.1111/ced.14029
39. JCS, J. C. of S. (2013). *JP 2-0, Joint Intelligence*.
40. Kober, J., Bagnell, A., & Peters, J. (2013). *Reinforcement learning in robotics: A survey*. https://journals.sagepub.com/doi/abs/10.1177/0278364913495721
41. Koch, B., Denton, E., Hanna, A., & Foster, J. G. (2021). *Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research* (arXiv:2112.01716). arXiv. https://doi.org/10.48550/arXiv.2112.01716
42. Lakomy, M. (2022). Assessing the potential of OSINT on the Internet in supporting military operations. *Bezpieczeństwo. Teoria i Praktyka*, *XLVIII*(3), 297–309.
43. Lipton, Z. C., & Steinhardt, J. (2018). *Troubling Trends in Machine Learning Scholarship* (arXiv:1807.03341). arXiv. https://doi.org/10.48550/arXiv.1807.03341
44. Liu, M., Ling, H., & Wu, D. (2021). *Sentinel-2 and Landsat-8 Observations for Harmful Algae Blooms in a Small Eutrophic Lake*. https://www.mdpi.com/2072-4292/13/21/4479
45. Mäkelä, P. (2019, July 14). *Here Is the Ultimate Way to Tell One Russian Tank from Another* [Text]. The National Interest; The Center for the National Interest. https://nationalinterest.org/blog/buzz/here-ultimate-way-tell-one-russian-tank-another-66807
46. Marappan, R., & Bhaskaran, S. (2022). Datasets Finders and Best Public Datasets for Machine Learning and Data Science Applications. *COJ Robotics & Artificial Intelligence*, *2*(1), 1–4.
47. Mareš, M. (2022, November 14). *Český start-up našel na satelitních snímcích z míst výbuchu Nord Streamu dvě tajemné lodě*. Hospodářské noviny (HN.cz). https://hn.cz/c1-67137640-cesky-start-up-nasel-na-satelitnich-snimcich-z-mist-vybuchu-nord-streamu-dve-tajemne-lode
48. Marra, F., Gragnaniello, D., Cozzolino, D., & Verdoliva, L. (2018). Detection of GAN-Generated Fake Images over Social Networks. *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 384–389. https://doi.org/10.1109/MIPR.2018.00084
49. Martin, A. (2023). *Russian naval officer killed near home may have been tracked on Strava app*. https://therecord.media/russia-submarine-captain-killed-strava-app-jogging-route

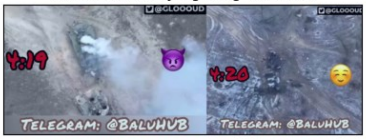
50. Mitzer, S. (2023). *Putting Down The Pen: Reflecting On Oryx's Journey—Oryx*. https://www.oryxspioenkop.com/2023/08/putting-down-pen-reflecting-on-oryxs.html
51. Mostafi, S., & Elgazzar, K. (2021). An Open Source Tool to Extract Traffic Data from Google Maps: Limitations and Challenges. *2021 International Symposium on Networks, Computers and Communications (ISNCC)*, 1–8. https://doi.org/10.1109/ISNCC52172.2021.9615680
52. Muller, M., Lange, I., Wang, D., Piorkowski, D., Tsay, J., Liao, Q. V., Dugan, C., & Erickson, T. (2019). How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3290605.3300356
53. Murauskaite, E. E. (2023). *U.S. Assistance to Ukraine in the Information Space: Intelligence, Cyber, and Signaling*.
54. Murphy, K. P. (2022). *Probabilistic Machine Learning: An Introduction*. MIT Press.
55. Newdick, T. (2023, December 1). *Ukrainian Su-25 Struck By Lancet Drone Was An Elaborate Decoy*. The Drive. https://www.thedrive.com/the-war-zone/ukrainian-su-25-attack-jet-stuck-by-lancet-drone-was-a-decoy
56. Nixon, M., & Aguado, A. (2019). *Feature Extraction and Image Processing for Computer Vision*. Academic Press.
57. NYT. (n.d.). Visual Investigations. *The New York Times*. Retrieved 27 December 2023, from https://www.nytimes.com/spotlight/visual-investigations
58. O'Malley, J. (2018, January 12). *Captcha if you can: How you've been training AI for years without realising it*. TechRadar. https://www.techradar.com/news/captcha-if-you-can-how-youve-been-training-ai-for-years-without-realising-it
59. Omar, S. (2023, May 25). *Twitter plans to axe old accounts may hurt Syria war probes*. Https://Www.Newarab.Com/; The New Arab. https://www.newarab.com/news/twitter-plans-axe-old-accounts-may-hurt-syria-war-probes
60. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). *Training language models to follow instructions with human feedback*.
61. Pastor-Galindo, J., Nespoli, P., Gómez Mármol, F., & Martínez Pérez, G. (2020). The Not Yet Exploited Goldmine of OSINT: Opportunities, Open Challenges and Future Trends. *IEEE Access*, *8*, 10282–10304. https://doi.org/10.1109/ACCESS.2020.2965257
62. Paul, C., & Matthews, M. (2016). *The Russian*. RAND Corporation. https://www.rand.org/pubs/perspectives/PE198.html
63. Perez, L. M., Jones, J., Englert, D. R., & Sachau, D. (2010). Secondary Traumatic Stress and Burnout among Law Enforcement Investigators Exposed to Disturbing Media Images. *Journal of Police and Criminal Psychology*, *25*(2), 113–124. https://doi.org/10.1007/s11896-010-9066-7
64. Pérez-Peña, R., & Rosenberg, M. (2018, January 29). Strava Fitness App Can Reveal Military Sites, Analysts Say. *The New York Times*. https://www.nytimes.com/2018/01/29/world/middleeast/strava-heat-map.html
65. Phillips, D. C., & Burbules, N. C. (2000). *Postpositivism and Educational Research*. Rowman & Littlefield.
66. Postma, F. (2021, May 28). *US Soldiers Expose Nuclear Weapons Secrets Via Flashcard Apps*. Bellingcat. https://www.bellingcat.com/news/2021/05/28/us-soldiers-expose-nuclear-weapons-secrets-via-flashcard-apps/

67. Project, I. U. of T. O. O., & CBC, for C. (2020, July 23). *Unmasking China's invisible fleet in North Korean waters*. https://newsinteractives.cbc.ca/longform/china-at-sea
68. Räsänen, M., & Nyce, J. M. (2013). The Raw is Cooked: Data in Intelligence Practice. *Science, Technology, & Human Values*, *38*(5), 655–677. https://doi.org/10.1177/0162243913480049
69. Rosen, A. (2018, March 7). *Erasing History: YouTube's Deletion Of Syria War Videos Concerns Human Rights Groups*. Fast Company. https://www.fastcompany.com/40540411/erasing-history-youtubes-deletion-of-syria-war-videos-concerns-human-rights-groups
70. Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, *346*(6213), 1063–1064. https://doi.org/10.1126/science.346.6213.1063
71. Ryan, A. B. (2006). Post-Positivist Approaches to Research. In M. Antonesa, H. Fallon, A. B. Ryan, A. Ryan, T. Walsh, & L. Borys (Eds.), *Researching and Writing your Thesis: A guide for postgraduate students* (pp. 12–26). MACE: Maynooth Adult and Community Education. http://mural.maynoothuniversity.ie/874/
72. Salerno-Garthwaite, A. (2022, November 23). *OSINT in Ukraine: Civilians in the kill chain and the information space - Global Defence Technology | Issue 137 | October 2022*. https://defence.nridigital.com/global_defence_technology_oct22/osint_in_ukraine
73. Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021). "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3411764.3445518
74. Samuel, A. (2018, May 15). *Amazon's Mechanical Turk has Reinvented Research*. JSTOR Daily. https://daily.jstor.org/amazons-mechanical-turk-has-reinvented-research/
75. Sanger, D. E., & Miller, C. C. (2014). Obama maintains United States' grip on the data pipeline. *International New York Times*, NA-NA.
76. Scambaiter (Director). (2022, December 20). *These SCAMMERS Panic After Finding Hackers In Their CCTV Cameras!* https://www.youtube.com/watch?v=tyEoOfSECp0
77. Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., & Anderson, R. (2023). *The Curse of Recursion: Training on Generated Data Makes Models Forget* (arXiv:2305.17493). arXiv. https://doi.org/10.48550/arXiv.2305.17493
78. Somers, J. (2021, March 18). The Pastry A.I. That Learned to Fight Cancer. *The New Yorker*. https://www.newyorker.com/tech/annals-of-technology/the-pastry-ai-that-learned-to-fight-cancer
79. Spelda, P., & Stritecky, V. (2022). Human Induction in Machine Learning: A Survey of the Nexus. *ACM Computing Surveys*, *54*(3), 1–18. https://doi.org/10.1145/3444691
80. Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning, second edition: An Introduction*. MIT Press.
81. Teichmann, F. M. J., & Falker, M.-C. (2020). Money laundering – the gold method. *Journal of Money Laundering Control*, *26*(3), 509–522. https://doi.org/10.1108/JMLC-07-2019-0060
82. Tuomi, I. (1999). Data Is More than Knowledge: Implications of the Reversed Knowledge Hierarchy for Knowledge Management and Organizational Memory. *Journal of Management Information Systems*, *16*(3), 103–117. https://doi.org/10.1080/07421222.1999.11518258
83. Turk, K., Pastrana, S., & Collier, B. (2020). A tight scrape: Methodological approaches to cybercrime research data collection in adversarial environments. *2020 IEEE European*

Symposium on Security and Privacy Workshops (EuroS&PW), 428–437. https://doi.org/10.1109/EuroSPW51379.2020.00064
84. Vijayakumar, S., & Sheshadri, K. N. (2019). Applications of Artificial Intelligence in Academic Libraries. *International Journal of Computer Sciences and Engineering*, *7*.
85. WarSpotting. (2023a). *About · WarSpotting*. https://ukr.warspotting.net
86. WarSpotting. (2023b, June 1). *#ThrowbackLoss On 1 Jun 2022* [Tweet]. Twitter. https://twitter.com/WarSpotting/status/1664265604319309824
87. WarSpotting. (2023c, October 18). *With some prominent OSINT accounts suspending their activity lately—@oryxspioenkop, @UAWeapons, anyone remembers @Blue_Sauron btw? - The community is facing seemingly dire outlook. What now? How much steam is left? Here's few thoughts: As usually, some bad and good ones. 1/* [Tweet]. Twitter. https://twitter.com/WarSpotting/status/1714693819780706433
88. WarSpotting. (2023d, December 7). *Interestingly, Msta-S is quite susceptible to loosing its turret* [Tweet]. Twitter. https://twitter.com/WarSpotting/status/1732796221280993629
89. Wu, S., Fei, H., Qu, L., Ji, W., & Chua, T.-S. (2023). *NExT-GPT: Any-to-Any Multimodal LLM* (arXiv:2309.05519). arXiv. https://doi.org/10.48550/arXiv.2309.05519
90. X, (formerly Twitter). (2023, October 19). *Ragnar Gudmundsson ᴉsuA (@ragnarbjartur) / X*. X (Formerly Twitter). https://twitter.com/ragnarbjartur
91. Zafarani, R., Abbasi, M. A., & Liu, H. (2014). *Social Media Mining: An Introduction*. Cambridge University Press.
92. Zaluzhny, V. (2023, November 1). The commander-in-chief of Ukraine's armed forces on how to win the war. *The Economist*. https://www.economist.com/by-invitation/2023/11/01/the-commander-in-chief-of-ukraines-armed-forces-on-how-to-win-the-war
93. Zbraneproukajinu. (2023). *Čestmír—Helikoptéra Black Hawk pro Ukrajinu—Dárek pro Putina*. https://www.zbraneproukrajinu.cz/kampane/cestmir
94. Zuzanna, K., Tomasz, U., Michał, G., & Robert, P. (2022). How High-Tech Solutions Support the Fight Against IUU and Ghost Fishing: A Review of Innovative Approaches, Methods, and Trends. *IEEE Access*, *10*, 112539–112554. https://doi.org/10.1109/ACCESS.2022.3212384

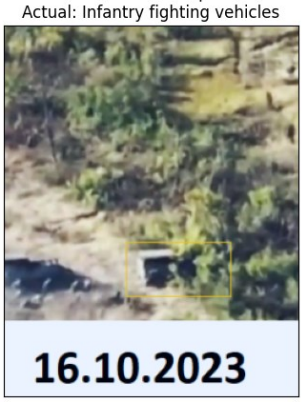
Appendix 1: A Sample of Misclassified Test Images

Predicted: Tanks
Actual: Infantry fighting vehicles



Predicted: Tanks
Actual: Infantry fighting vehicles

Predicted: Transport
Actual: Infantry fighting vehicles



Predicted: Tanks
Actual: Infantry fighting vehicles

Predicted: Tanks
Actual: Infantry fighting vehicles

Predicted: Tanks
Actual: Infantry fighting vehicles

Predicted: Tanks
Actual: Infantry fighting vehicles

Predicted: Tanks
Actual: Infantry fighting vehicles

Predicted: Tanks
Actual: Infantry fighting vehicles

Predicted: Tanks
Actual: Infantry fighting vehicles

Predicted: Tanks
Actual: Infantry fighting vehicles

Predicted: Tanks
Actual: Infantry fighting vehicles

Predicted: Tanks
Actual: Infantry fighting vehicles

Predicted: Tanks
Actual: Infantry fighting vehicles

Predicted: Self-propelled artillery
Actual: Infantry fighting vehicles

Predicted: Transport
Actual: Infantry fighting vehicles