A DEEP LEARNING APPROACH TO PREDICTING POTENTIAL VIRUS SPECIES CROSSOVER USING CONVOLUTIONAL NEURAL NETWORKS AND VIRAL PROTEIN SEQUENCE PATTERNS

by

Rudolph Serage

A thesis submitted in fulfilment of the requirements for the degree of

Magister Scientiae

in the

South African National Bioinformatics Institute

at the

University of the Western Cape

Supervisor: Dr. Dominique Anderson

Co-supervisor: Prof. Alan Christoffels

September 2022

Abstract

Medical science has made substantial progress toward diagnosing, understanding the pathogenesis, and treating various causative agents of infectious disease; however, novel microbial pathogens continue to emerge, and existing pathogens continue to evolve alternative means to thrive in ever-changing environments. Various infectious disease etiological agents originate from animal reservoirs, and many have, over time, acquired the ability to cross the species barrier and alter their host range. The emergence and re-emergence of zoonotic pathogens is reported to be a consequence of changes in several factors, including ecological, behavioural, and socioeconomic variables which are arguably impossible to control. Computational methods with the capacity to evaluate large datasets, are considered invaluable tools for predicting and tracking disease outbreaks and are especially powerful when combined with machine learning techniques. These predictive methods may be integral, not only as early warning systems for outbreak preparedness, but also in the monitoring of intervention effectiveness during epidemics or pandemics. This study aimed to develop a machine learning model which would allow for prediction of potentially zoonotic organisms, by using viral surface proteins which facilitate viral entry into host cells, as the data input for training. Sequence data and metadata was obtained from UniProtKB, transformed into a machine-readable format, using frequency chaos game representation (FCGR). A deep convolution neural network model was developed which identified sequence patterns consistent with viruses which infect humans. The model achieved 96.78% accuracy, 0.97 F1 score and 0.93 MCC on unseen data, outperforming machine learning models found in literature.

Keywords: Convolutional Neural Network, Deep learning, Frequency Chaos Game Representation, Machine learning, Prediction Model, Species Cross-over, Viral Protein Sequences, Viral Zoonosis.

Declaration

I, Rudolph Abel Serage, hereby declare that this thesis entitled *A deep learning approach to predicting potential virus species crossover using convolutional neural networks and viral protein sequence patterns* for which I submit for the degree of Master of Science in Bioinformatics at the University of the Western Cape represents my own work and opinions, has not been submitted at this institution or any institution thereof, and all external sources of information have been clearly acknowledged by citation.

Signature:  Date: 30 July 2022

Acknowledgements

I would first like to acknowledge Ms. Rebotile Lediga, Prof. Katherine Scholtz and Dr. Mushal Ali for their advice and encouragement on joining the South African National Bioinformatics Institute.

I would like to thank my supervisors, Dr. Dominique Anderson and Prof. Alan Christoffels, for their time, commitment and continued support throughout my candidacy. I would like to especially thank Dr. Dominique Anderson for the warm welcome into the university and research group, and mostly appreciate the encouragement to take ownership of the project and facilitating a growth and learning environment.

I would like to thank the SANBI community for facilitating a warm, welcoming, and supportive environment. I would like to acknowledge Peter Van Heusden and Nasr Eshibona for making time to evaluate some of my work and providing valuable feedback. My sincere gratitude goes to the Cubicle 7 research group, Abiola Babajide, Peter Abiodun, Tatenda Mujuru and Dominique Anderson, for their professional and personal support and encouragement throughout.

A warm gratitude goes to my friends, Tshireletso Manaso, Thabo Semenya, and Andrew Mpe, who have been a great support system providing familiarity in an otherwise unfamiliar environment and being available in times of hardship.

A heartfelt appreciation goes to my parents, Joyce and David Serage, for their continued support and encouragement throughout my journey and making it their journey as well.

Lastly, I would like to acknowledge the National Research Foundation (NRF) for their financial support in making this study possible. I would also like to make a general acknowledgement of everyone who has made any financial contribution in pursuit of my studies.

Table of contents

List of figures

List of tables

Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area Under the Curve
CA	Correspondence Analysis
CDC	Centers for Disease Control and prevention
CGR	Chaos Game Representation
CNN	Convolutional Neural Network
CPU	Central Processing Unit
DB	Database
DNA	Deoxyribonucleic Acid
EBI	European Bioinformatics Institute
EID	Emerging Infections Disease
EID2	Enhanced Infectious Diseases Database
EMBL	European Molecular Biology Laboratory
FCGR	Frequency Chaos Game Representation
GO	Gene Ontology
GPU	Graphics Processing Unit
HIV	Human Immunodeficiency Virus
LMICs	Low to Middle Income Countries
MCC	Mathews Correlation Coefficient
ML	Machine learning

NCBI	National Center for Biotechnology Information
OH	One Health
OM	One Medicine
PCA	Principal Component Analysis
PNG	Portable Network Graphics
POC	Proof of Concept
PP	Portal Protein
PPI	Protein-Protein Interaction
RNA	Ribonucleic Acid
ROC	Receiver Under the Operating Characteristic
ROC-AUC	Area Under the Receiver Under the Operating Characteristic Curve
R_0	Basic Reproductive Number
ROS	Random Oversampling
RUS	Random Undersampling
WHO	World Health Organization

Chapter 1: Introduction, aims and objectives

1.1. Introduction

Humans, animals, and pathogens live in a dynamic, interactive, and interconnected environment whereby the health of one affects the other (Calistri et al., 2013). A holistic approach to this ecological interdependence is referred to as One Health and has been defined as a multidisciplinary initiative, at global and national levels, to guarantee relatively optimal health for humans, animals, and the environment (Calistri et al., 2013; Hitziger et al., 2018; Pettan-Brewer et al., 2021). Population growth coupled with alterations in the environment have resulted in very close human and animal (wild and domestic) contact (Brierley and Fowler, 2021; Faburay, 2015; Taylor and Vaisman, 2010). This proximity relationship is postulated to be a driver of increased emergence and re-emergence of infectious diseases through zoonosis, a phenomenon whereby infectious microorganisms from animals cross the species barrier and infect humans (Dallas et al., 2019). Zoonotic diseases have been a major economic burden and public health concern on a global scale (Dallas et al., 2019; Smith et al., 2014). Economic impacts which may result from epidemic or pandemic response measures to prevent spreading of the disease, include, but are not limited to, trade and travel restrictions and increased spending on resources (Cantas and Suer, 2014; Madhav et al., 2017).

Two conceptual frameworks, which guide the dynamics of cross-species events, are the pyramid and the pinhole models. The pyramid model defines spill-over as a gradual process consisting of multiple stages whereby a pathogen migrates from a reservoir host through several intermediate hosts, and ultimately establishes itself as a human specific pathogen (Brierley et al., 2016; Warren and Sawyer, 2019). The pinhole model, however, describes zoonosis as a bottleneck event, where only a very small number of viruses from reservoirs may become zoonotic and as such, spill-over would be relatively difficult to achieve (Warren and Sawyer, 2019). The virus discovery curve indicates that there is a significant number of viruses which are yet to be discovered (Anthony et al., 2013; Woolhouse et al., 2008). However, there is little indication of which, if any, of these newly discovered organisms, could be pathogenic, and potentially result in zoonotic events (Anthony et al., 2013; Carroll et al., 2018; Woolhouse et al., 2008).

Public health research priorities toward emerging infectious diseases are largely focused on the detection and surveillance of emerging infectious diseases (EIDs), as well as the identification of

factors which are drivers of transmission, for public safety and mitigation of disease impacts (Carroll et al., 2018; Temmam et al., 2014). Surveillance methods are also used throughout infectious disease outbreaks to trace the spread of infection, with the intent to limit disease spread, and monitor effectiveness of applied interventions (Baum et al., 2017; Carroll et al., 2018). In addition to traditional surveillance efforts, various models, using statistical and machine learning tools, have been developed to predict cross-species spill-over of novel and re-emerging pathogens, as well as transmission dynamics, once an outbreak has occurred (Han et al., 2015; Royce and Fu, 2020). These models incorporate ecological, demographic, and biogeographic data as covariates for robust algorithm development, to predict potentially zoonotic pathogens or identify existing reservoirs and new potential hosts (Han et al., 2015; Wardeh et al., 2020b). These predictive methods may be integral in early warning systems, allowing nations to prepare for an outbreak, limit human casualties and reduce economic burden, which is of particular importance in low-to-middle income countries (LMICs) (Madhav et al., 2017). More recently, the analysis of pathogen and host protein interaction networks has also been included in AI-based efforts, to predict species cross-over events, at the molecular level (Kösesoy et al., 2021; Yan et al., 2019).

Research of host-pathogen protein-protein interactions (PPIs) often addresses several interactions at once, which is by no means a trivial task, nor without limitations. For example, there may be intracellular protein interactions, but no receptor interactions for a predicted virus-host protein interacting pair. The former may be useful in understanding pathogenesis but may not be indicative of a potential cross-species event mediated by cell receptor binding (Cho and Son, 2019). To circumvent this limitation, virus-receptor PPI models have been developed which are used to predict cross-species events (Cho and Son, 2019; Yan et al., 2019). However, studies are limited by the availability of experimentally derived and validated PPIs, thereby limiting the amount of available input data for machine learning models to produce robust, translatable, and reproducible models (Bae and Son, 2011; Cho and Son, 2019; Yan et al., 2019). Furthermore, PPI studies rely on defined pre-existing interactions which may be unable to predict viral host switching in which a previously unknown host receptor is targeted for entry into host cells (Deng et al., 2021; Kösesoy et al., 2021).

Investigation of virus receptors to identify patterns signifying zoonotic potential have also been explored in studies such as those by Qiang and Kou (2019), on Influenza A proteins, and Qiang et al. (2020) on genomes and spike proteins in coronaviruses. These studies are however, limited to a specific viral species, prompting the need for a predictive model which uses virus surface proteins regardless

of pathogen species. This study aims to fill this gap by developing a machine learning model to predict potential for cross species events using surface protein patterns across a broad range of viral species.

To summarise the thesis, surface protein sequences of viruses which have been reported to infect humans and those which do not infect humans were obtained from UniProtKB and sequence data was transformed into machine readable frequency chaos game representation (FCGR) images in preparation for machine leaning. A convolutional neural network (CNN) model was created and allowed for pattern learning in the training data with respect to the defined positive or negative classes. The trained CNN binary classification model was then tested on unseen data and as a proof of concept, the model was used to predict the zoonotic potential of 4 viruses, two of which are reportedly known to infect humans, and two which have thus far, not been reported to infect humans. This study presents a proof-of-concept approach for building viral zoonosis prediction using FCGR images of protein sequences and CNNs. The findings suggest the presence of protein sequence patterns consistent with viral zoonotic potential, with the model producing accurate predictions on unseen data. The model developed in this study may assist in identifying potentially zoonotic viruses based on sequence data extracted from pathogen surveillance programs, and the results of which may be used to inform public health policy makers and support mitigation of potential epidemics and pandemics.

1.2. Aims and objectives

To achieve the research aims of this study, namely, 1) create an AI based model, trained on viral surface protein sequences, which can predict viruses with the potential to cross the species barrier, and 2) proof-of-concept testing of the model and approach using unseen data, the following objectives were undertaken.

1. Identify viruses reported to have crossed the species barrier and infect humans
2. Identify viruses in other hosts which have not been reported to infect humans
3. Obtain metadata and amino acid sequences of the viruses reported to have crossed the species barrier
4. Obtain metadata and amino acid sequences of the viruses which have not been reported to infect humans
5. Define positive and negative classes
6. Transform the amino acid sequences into machine readable sequences
7. Train a convolutional neural network on the sequence data
8. Evaluate the model on unseen data
9. Proof of concept prediction and evaluation of model using selected unseen sequence data

Chapter 2: Literature Review

Infectious diseases have been an overwhelming burden on civilisations throughout recorded history (Madhav et al., 2017; Wolfe et al., 2007; Woolhouse and Gaunt, 2007). One of the many negative consequences of infectious disease outbreaks is the massive socio-economic impact and imposed disruptions to normalcy. Depending on the scale, impact and region, an outbreak can result in varying scales of economic, social, and political disruptions (Madhav et al., 2017). Aside from the economic devastation of disease outbreaks, immeasurable costs to society, attributed to the loss of human life occurs, with infectious diseases being documented to account for approximately 60 million deaths each year (Languon and Quaye, 2019). Several infectious diseases which have recently appeared as outbreaks, have been termed emerging infectious diseases (EIDs) (Piret and Boivin, 2021). Emerging infectious diseases are defined in literature as those which infect the human population for the first time or have geographic distribution in previously unaffected locations (Bogich et al., 2012; Engering et al., 2013; Funk et al., 2013; NIH, 2007; Taylor et al., 2001). At the time of writing this thesis, two and a half years had passed since the emergence of a new severe acute respiratory syndrome coronavirus 2 (SARS-Cov2) resulting in a persistent pandemic (Haider et al., 2020; Rothan and Byrareddy, 2020).

2.1. Emerging infectious diseases enablers

Emerging (and re-emerging) infectious diseases (EIDs) present an increasing threat to public health and studies have indicated that the majority of these EIDs are zoonotic in origin (Bogich et al., 2012; Engering et al., 2013; Jones et al., 2008). Identifying the exact animal origins of zoonoses is a complex task, despite the use of sophisticated molecular biology techniques (Kerr et al., 2015). In a study by Wolfe et al. (2007), the authors illustrated that animal origins of infectious disease agents varied significantly between temperate and tropical regions. It was observed that in tropical regions, a significant number of zoonotic diseases appeared to originate from wildlife rather than from domesticated animals, and that the opposite phenomenon occurred in temperate regions (Han et al., 2015; Wolfe et al., 2007). The diversity of organisms, including insects and pathogens, increases toward the equator, therefore, deforestation and land encroachment in these regions may cause shifts in wildlife habitats (Brierley et al., 2016; Horby et al., 2014). These changes can subsequently facilitate an increased interaction between wildlife and domesticated animals, thereby allowing wildlife pathogens to adapt to, and inhabit new environmental niches (Horby et al., 2014). This may expand the potential host range of pathogens and indirectly expose human hosts to wildlife-based pathogens, thus resulting in outbreaks (Brierley et al., 2016; Kilpatrick and Randolph, 2012; Loh et al., 2013).

Aside from geographical location, several factors have been proposed as drivers, or enablers, of the emergence and re-emergence of infectious diseases (Brierley et al., 2016; Woolhouse and Gowtage-Sequeria, 2005). Modernisation and globalization activities are commonly associated enablers of EID events. Socio-economic drivers such as environment encroachment, fossil fuel extraction, animal production systems, global trade, wildlife trade, international travel and population expansion have been associated with the emergence and re-emergence of infectious diseases by facilitating increased opportunities for transmission (Brierley et al., 2016; Kilpatrick and Randolph, 2012; Smith and Wang, 2013). Both domestic and wildlife animal production systems are thought to enable food-borne disease outbreaks (Karesh et al., 2012; Kilpatrick and Randolph, 2012; Smith and Wang, 2013) and these are particularly related to practices such as animal slaughter, meat and by-product processing, packaging, transportation, and preparation, prior to consumption (Karesh et al., 2012).

Events related to climate change and pathogen evolution have also been proposed as additional enablers of infectious disease emergence by promoting pathogen richness (French and Holmes, 2020). Alterations in climate conditions impact wildlife and vector ecology, resulting in the formation of new inter-species relationships, which may facilitate species crossover events resulting in zoonoses with pandemic potential (Brierley et al., 2016; Faburay, 2015; French and Holmes, 2020). An example of this is the 1993 and 1997 Hantavirus outbreaks in the Southwestern United States, which were attributed to heavy snow and rainfall leading to changes in the abundance of infected rodents (CDC, 2020). The rodents migrated to areas with adequate vegetation, which was a shared space with humans, thereby enabling Hantavirus spill-over (Horby et al., 2014).

In addition to behavioural and social drivers, the increased interaction between phylogenetically related organisms, or hosts, may serve as an additional enabling factor for species crossover events. Pathogens may require relatively few (or no) mutations to adapt to a different host due to the similarity of immunological and molecular systems in closely related host species (Brierley et al., 2016). Olival et al. (2017) demonstrated that taxonomic relatedness affected the sharing of pathogens among species and, similarly, Wardeh et al. (2020b) indicated that host taxonomy affected which pathogens were shared between hosts.

2.2. Frameworks for investigating emerging infectious diseases

Clearly, a multitude of factors are considered as EID drivers, and as such, research of EID epidemiology spanning multiple disciplines requires collaborative efforts (Salyer et al., 2017; Zinsstag et al., 2012).

The co-dependency of animal, environment and human health requires cooperative frameworks to study and understand the intricacies of their interactions, and overall health (Salyer et al., 2017).

The need for such frameworks which intersect animal and human medicine for public health benefit and species conservation, traces back to the works of Rudolf Virchow, and later Calvin Schwabe (Zinsstag et al., 2012). It was in the mid-80s that Schwabe coined the term "One Medicine" in his book entitled *Veterinary Medicine and Human Health* as a descriptive term for the framework (Lee and Brumme, 2013; Zinsstag et al., 2012). He argued that "the critical needs of man include the combating of diseases, ensuring enough food, adequate environmental quality and a society in which humane values prevail" (Lee and Brumme, 2013). The increased prevalence of zoonosis, coupled to research demonstrating the interconnectivity between driving factors and EID incidence, has motivated for the emergence of the transdisciplinary field of One Health, demonstrated in Figure 1 (Lee and Brumme, 2013; Loh et al., 2013; Zinsstag et al., 2012; Zinsstag et al., 2011).

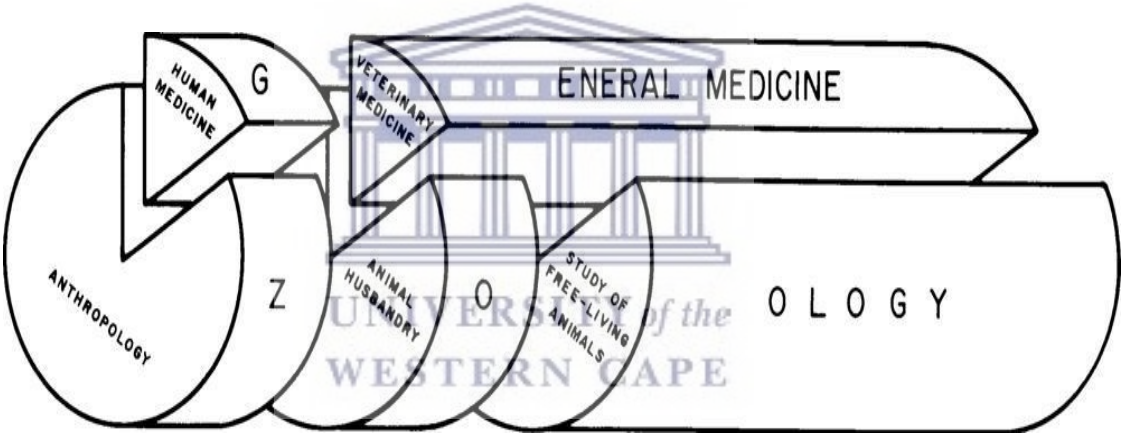

Figure 1: One Medicine as initially described by Schwabe (taken from Zinsstag et al. (2011)).

One Health is defined as a global collaborative and multi-sectorial approach which aims to collate concepts from three pillar disciplines -human health, animal health and environmental health - and prevent or mitigate epidemic or pandemic risks (Baum et al., 2017; Cassidy, 2018; Karesh et al., 2012; Lee and Brumme, 2013; Lerner and Berg, 2017; Loh et al., 2013; Roger et al., 2016). The One Health concept is not limited to zoonotic disease, and involves food safety and health services delivery, amongst others (Baum et al., 2017; Zinsstag et al., 2012). Figure 2 depicts the evolution of One Health,

listing fields and outcomes of the framework, and include research, surveillance, control programs and policy framework development (Baum et al., 2017).

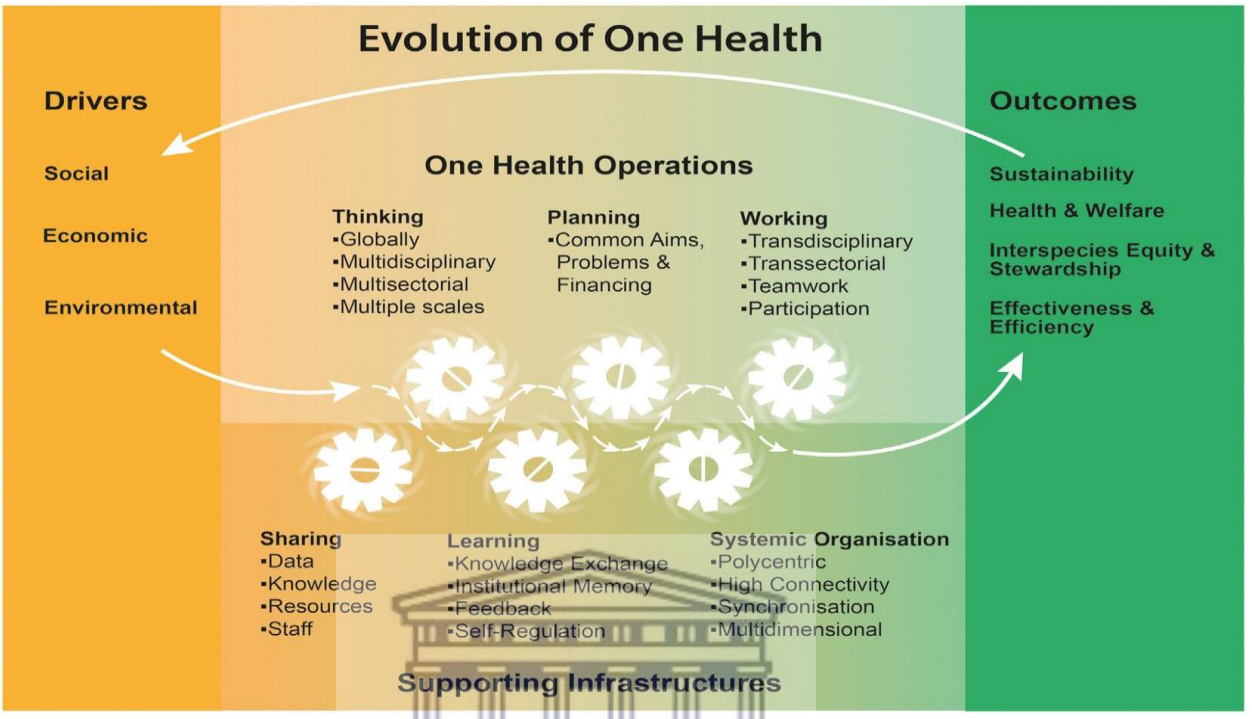

Figure 2: One Health characteristics identified during a workshop in 2015: Network for Evaluation of One Health, http://neoh.onehealthglobal.net (taken from Rüegg et al. (2017)).

A parallel framework termed EcoHealth, formulated by disease ecologists exists, and is an ecosystems approach to health with a focus on environmental and socio-economic issues (Roger et al., 2016). Attempts have been made to collate the EcoHealth and One Health frameworks (Lerner and Berg, 2017; Roger et al., 2016). While EcoHealth is outside of the scope of this study, the reader is directed to studies by Lerner and Berg (2017) and Roger et al. (2016) which provide extensive comparison of the frameworks mentioned above.

2.3. Modelling zoonotic spill-over

Researchers have attempted to model the dynamics of zoonotic spill-over, from the point of exposure to epidemic spread, and human exclusivity (Madhav et al., 2017; Morse et al., 2012; Warren and Sawyer, 2019). The models describe the process of pathogen migration from their primary hosts (reservoir hosts) to humans, however, it is important to note that these models are not ubiquitous, and

follow different patterns depending on the pathogen group (viruses, bacteria, fungi, etc.) (Warren and Sawyer, 2019; Woolhouse et al., 2012). This study primarily focuses on viral pathogens and as such, only spill-over models which are applicable to viral pathogens will be discussed further.

Emerging infectious disease (EID) research frequently demonstrates that viral pathogens are common aetiological agents of infectious disease (Warren and Sawyer, 2019; Woolhouse et al., 2012). Viruses are microscopic parasites capable of infecting single cells and consisting of a simple protein capsid which encases the genomic material (Madhav et al., 2017). Substantial viral diversity has already been characterised from a multitude of environmental niches, with many more still yet to be discovered (Carroll et al., 2018). Viruses are highly diverse in morphology and transmissibility, and coupled with their relatively large host range, viral infectious agents account for a significant number of zoonoses (Olival et al., 2017; Parrish et al., 2008; Siegel, 2018) with approximately 75% of zoonotic infectious diseases estimated to be caused by viruses (Carroll et al., 2018; Haider et al., 2020).

Woolhouse et al. (2012) described a 4-level pyramid zoonosis model, in which the first level represents the initial point of exposure of a new host, to a novel pathogen, from an animal reservoir. This first level is a commonly occurring event, termed 'chatter' (Madhav et al., 2017; Woolhouse et al., 2012). Successful infection of the new host is not guaranteed at this level, however, pathogens which successfully bridge this step, advance to infect the new host – this is represented as the second level of the pyramid. Level 3 represents a subset of pathogens that not only infect the new host but can also be transmitted between individuals of the host. Finally, level 4 of the model represents pathogens that have gained sufficient transmissibility to ignite an infectious disease outbreak with potential epidemic or pandemic consequences. Sufficient transmissibility is further argued to be an epidemiological state of $R_0 > 1$, a state whereby a primary case can potentially generate more than one secondary case (Woolhouse et al., 2012).

A 5-stage zoonosis model by Madhav et al. (2017) defines the degree of zoonotic adaptation as a continuum, spanning initial enzootic infections (stage 1), to end-stage specialized human transmission (stage 5). The intermediate stages of the continuum represent the emergence of pathogens which are not well adapted to human hosts and as such, infections result in a series of localized outbreaks, defined as stuttering chains (Engering et al., 2013; Madhav et al., 2017; Royce and Fu, 2020). Pathogens which progress past stage 3, are arguably the most alarming as they can cause prolonged transmission chains in the human host (Madhav et al., 2017).

The zoonosis emergence framework detailed by Brierley et al. (2016) describes a 4-step process and includes the enabling drivers and additional factors which allow for the pathogen to progress to the next step. The authors apply the model to the emergence of viruses from bat reservoirs to human hosts and the first step represents the exclusive pathogen occurrence in the natural host, bats in this case. Initial pathogen sharing with the new human host is represented in the second step and in the third, which is cyclic in nature, shows the disease propagation cycle moving between endemic and epidemic sub-stages until it is reported as emerging (owing to diagnostic capacity and reporting capacity of the events in the propagation cycle) (Step 4).

Plowright et al. (2015) model the zoonosis model with context to Hendra virus spill-over from bats. It begins with the pathogen being present in the reservoir host and being shed by the reservoir in environments shared by an intermediate host, in the case of Hendra virus, horses (Plowright et al., 2015). The pathogen then survives in the external environment prior to transmission to the intermediate host via a suitable route, such as the inhalation of the virus from grazing fields (Plowright et al., 2015). Contact of horses and humans then allows transfer of the aetiological agent into human hosts. Thus far, human to human transmission of Hendra Virus has not been reported (Plowright et al., 2015).

Several other models (see Figure 3 and Figure 4) have been presented by various authors, (Bogich et al., 2012; Epstein and Field, 2015; Karesh et al., 2012; Morse et al., 2012; Wolfe et al., 2007), however, they are ultimately derivations of two rather simplified and competing models, the pyramid model and the pinhole model, described in Figure 5.

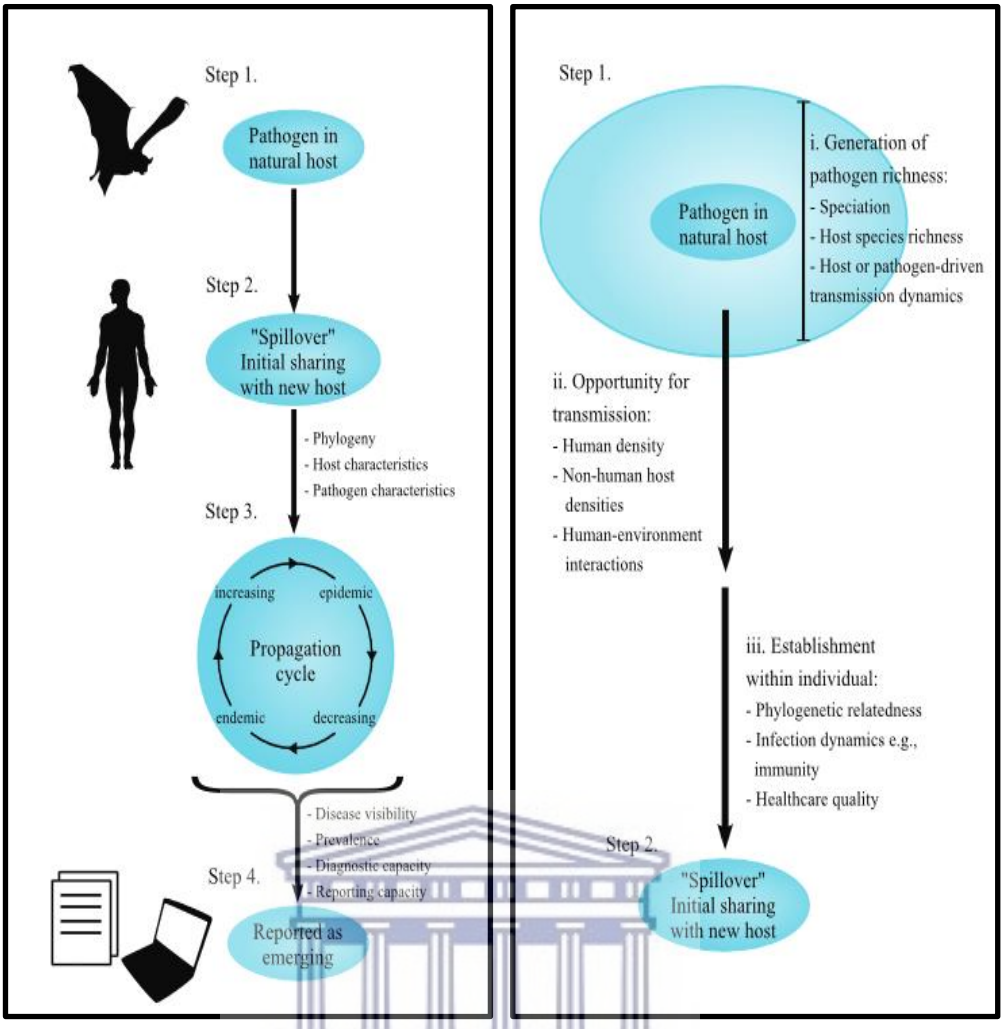

Figure 3: Variants of the pyramid (left) and pinhole (right) model as described by Brierley et al. (2016).

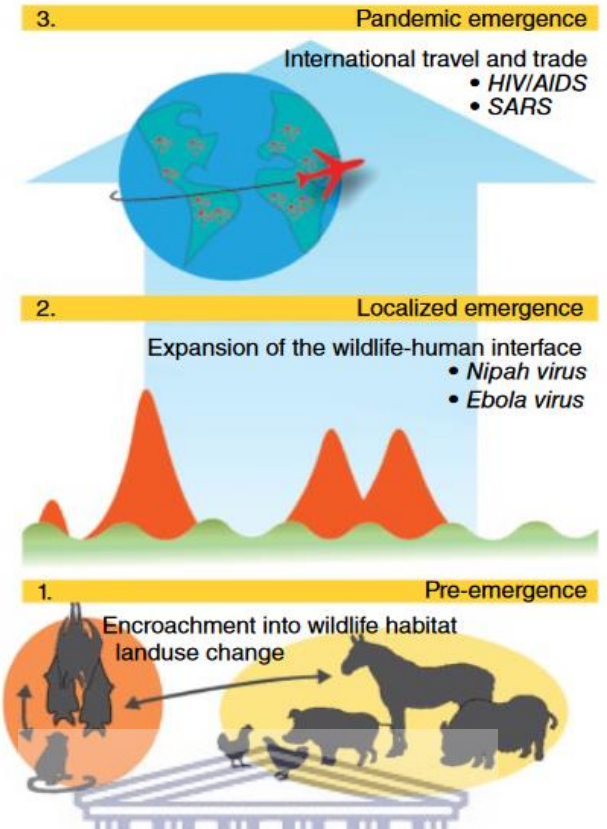

Figure 4: Pinhole model variant illustrated by Epstein and Field (2015).

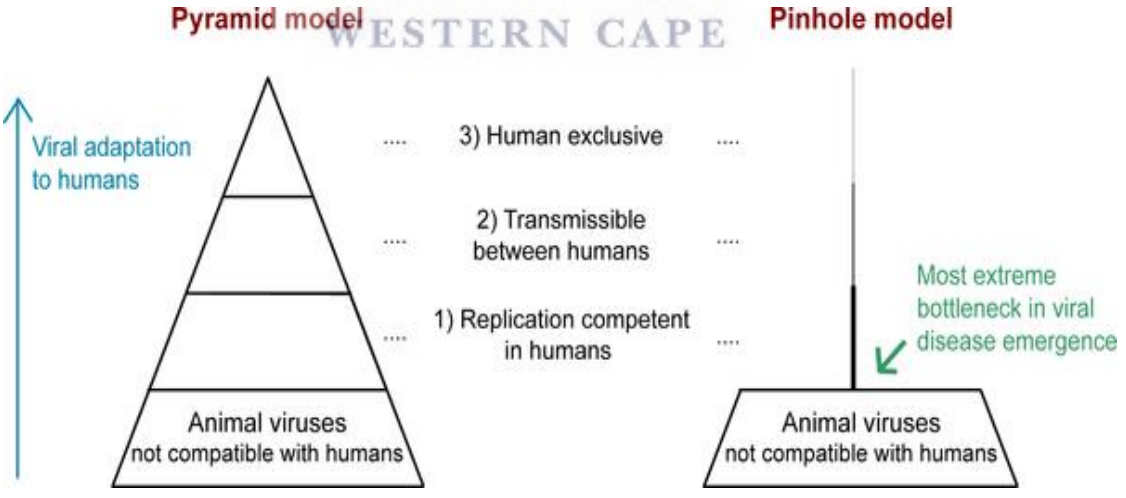

Figure 5: The classic pyramid model and the pinhole model which shows bottlenecks to animal virus progression to sustained inter-human transmission (taken from Warren and Sawyer (2019)).

As the name suggests, the pyramid model of zoonosis is depicted as a triangle with dissections representing stages in the spill-over event. In addition to sections of the spill-over process, it further describes the diversity of pathogens in each section with the base representing the theoretical total diversity of viral pathogens in animal populations and the apex representing the viral diversity which have gained exclusivity in human hosts. The 4 levels described in Woolhouse et al. (2012) and the 5 stages in Madhav et al. (2017) are both derivations of the pyramid model (Warren and Sawyer, 2019). While this model provides an excellent depiction of the steps required by viruses to cross the species barrier and adapt to infect humans, it does not, however, illustrate the complexity and rareness of zoonotic events (Warren and Sawyer, 2019). The pinhole model, a modification of the pyramid model, addresses the infrequency of zoonosis and visually depicts how an exceedingly small fraction of animal viruses can gain the ability to adapt to, and replicate in human hosts (Warren and Sawyer, 2019).

There is a plethora of animal, human and environmental microorganisms currently known, and an even greater number which have not yet been discovered. These organisms are, in fact, integral components in the ecosystem (Enard et al., 2016; French and Holmes, 2020). Several relationships exist which have the potential to facilitate the interchange of microorganisms; however, only a small proportion are capable of species or niche crossover (French and Holmes, 2020; Warren and Sawyer, 2019; Woolhouse and Gowtage-Sequeria, 2005). Successful zoonosis events require numerous favourable conditions which include pathogen evolvability, frequent gene reassortment or recombination, and quasi-species formation, amongst others (Engering et al., 2013).

Even then, the above conditions require supplemental factors such as a viable transmission route and innate host factors, such as phylogenetic relatedness and tissue tropism, for successful zoonosis (Olival et al., 2017; Warren and Sawyer, 2019). It has been postulated that evolutionary trade-offs exist, in which host natural selection for resistance to one viral species, could in fact result in susceptibility to another viral species, supporting re-emergence, or the breakthrough of novel virus strains which previously could not infect humans (Brierley et al., 2016; Daugherty and Malik, 2012; Enard et al., 2016; Kerr et al., 2015; McBee et al., 2015; Woolhouse et al., 2013).

Despite these multiple barriers, zoonotic pathogens continue their emergence, and considering the discovery curve, described by Woolhouse et al. (2012), the existence of novel pathogens which are yet to be discovered, increases the probability of zoonosis and disease outbreak events (Carroll et al., 2018).

2.4. Emerging infectious disease surveillance and modelling

Public health research priorities toward emerging infectious diseases are largely focused on the detection and surveillance of EIDs, as well as the identification of factors driving transmission, to intervene for public safety and mitigate the effects of disease (Carroll et al., 2018; Temmam et al., 2014). Detection efforts are focused on deployment of analytical, laboratory-based methods for identification of microorganisms, ranging from traditional culturing to modern molecular and "-omics" techniques (Carroll et al., 2018; Sandle, 2016; Temmam et al., 2014). Surveillance efforts include the screening of microbes from various sources with the aim to identify potential infectious agents prior to human host transmission, either directly or indirectly through intermediate animal hosts (Carroll et al., 2018; Cuervo-Soto et al., 2018; Temmam et al., 2014). Surveillance methods are also used throughout infectious disease outbreaks to trace the spread of infection and with the intent to mitigate spread of the disease.

In addition to traditional surveillance efforts, various statistical and machine learning (ML) models which make use of different covariates and prediction targets, have been developed to predict cross-species spill-over of novel and re-emerging infectious agents, as well as transmission dynamics once an outbreak has occurred (Eid et al., 2016; Han et al., 2015; Royce and Fu, 2020). These models incorporate ecological, demographic, and biogeographic data as covariates for robust algorithm designs to predict potential zoonotic pathogens or identify reservoirs and potential hosts. Analysis of pathogen and host interaction networks have also been included in machine learning based efforts. Wardeh et al. (2020b) used mammalian viral traits and network features in machine learning algorithm development to predict potential mammalian hosts of known viruses. The same authors also used shared pathogen networks and machine learning to predict reservoirs of zoonotic pathogens (Wardeh et al., 2020b). In their efforts, they demonstrate the importance of host phylogeny in pathogen sharing and quantify the extent of pathogen sharing between humans and other mammals.

Each year, several mathematical models are used to predict strains of concern for influenza, to manufacture vaccines against the predicted strains (Ray and Reich, 2018). Studies such as the one by Eng et al. (2017) have built computational models with machine learning tools to predict zoonotic influenza strains, by using host tropism signatures of avian zoonotic and human influenza strains. More recently, a study by Qiang and Kou (2019) used ML approaches and protein sequences to predict avian influenza interspecies transmission.

Bats and rodents are considered one of the most species-rich mammalian taxa and are known to harbour several potentially zoonotic diseases (Han et al., 2015; Olival et al., 2017). Work undertaken by Han et al. (2015) resulted in the creation of a model to predict rodent species which may be reservoir hosts of undiscovered zoonoses, using the biogeographic and ecological data of rodents. They obtained prediction accuracies in the 90th percentile and in addition, identified over 150 new hyper-reservoir rodent species. Notably, the analyses further indicated that viral taxa represented the majority of zoonotic agents in rodents, followed by protozoans.

While ecological models examine species crossover events at a broader scale, they lack information on a molecular scale. In addition, they also tend to focus on a few host or pathogen species, which may not translate well for more general application and prediction (Han et al., 2015; Olival et al., 2017; Wardeh et al., 2020b). Royce and Fu (2020) used the knowledge that intermediate hosts allow passage of otherwise rare diseases (allowing for greater adaptability to human hosts), to develop a mathematical model of the disease dynamics between reservoir and human hosts. The authors used epidemiology modelling and ecology, as well as the basic reproductive number of the pathogen, as the model parameters (Royce and Fu, 2020). Their findings indicate that even pathogens which have an $R_0 < 1$ in the intermediate host, may still have greater capacity to establish in a human host (Royce and Fu, 2020). While this model may be more generalisable, only a limited number of host and pathogen traits, such as biogeography, were used (Royce and Fu, 2020). Additional approaches to surveillance and development of EID models at the molecular scale, have leveraged interspecies protein interactions between viruses and hosts.

2.5. Modelling zoonosis at the molecular scale: protein-protein interactions

Viruses do not genetically encode the necessary machinery for replication and as such, require ‘help’ from a living host (NIH, 2007; Madhav et al., 2017). Viral invasion into host cells differs at the molecular level depending on the virus family in question (Letko et al., 2020; Madhav et al., 2017), however, the initial step of invasion is similar, albeit with differences in macromolecular component interactions (Liang and Bushman, 2021; Madhav et al., 2017; Sanjuán and Domingo-Calap, 2016). As a first step toward cellular entry, viral surface proteins require compatible host receptors for attachment and entry (Alguwaizani et al., 2018; Driscoll et al., 2009; Dyer et al., 2010). Post-entry, a myriad of protein-protein interactions between host and viral agent, may defend the host from infection, or facilitate viral infection (Alguwaizani et al., 2018; Doolittle and Gomez, 2011; Dyer et al., 2010).

Pathogen-host protein-protein interactions (PPIs) have been investigated in single viral species, and such studies have been valuable in the elucidation of intracellular pathway interactions in viral infections, aiding a greater understanding of cellular inter-communication and interactions in diseased states (Eid et al., 2016; Kösesoy et al., 2021; Yan et al., 2019). Pathogen-host protein-protein interactions research additionally provides a fundamental basis which can be applied to elucidate mechanisms of pathogenesis (Driscoll et al., 2009). Historically, PPI studies have been conducted experimentally in the laboratory (Deng et al., 2021; Qi et al., 2010) and these experimental studies tend to be laborious and expensive (Alguwaizani et al., 2018; Deng et al., 2021). As a result, computational methods to study PPIs have been developed, to increase throughput.

Computational methods used for protein-protein interaction studies make use of machine learning and deep learning to examine existing interactions and relationships, with a view to predict if a viral pathogen has the potential to interact with human proteins, and possibly, infect human hosts. Computational methods may be sequence based (Alguwaizani et al., 2018; Cui et al., 2012), structure based (de Chassey et al., 2013; Doolittle and Gomez, 2011), domain based (Dyer et al., 2007; Zheng et al., 2014) or motif-based (Evans et al., 2009; Zhang et al., 2017). With earlier studies in this domain being limited to a single species (Doolittle and Gomez, 2011; Kim et al., 2017; Mei and Zhao, 2018; Wuchty, 2011), recent studies have attempted to create models which have applicability across multiple species (Deng et al., 2021; Eid et al., 2016; Kösesoy et al., 2021).

While PPI studies model out the transition from level 1 to level 2 of the pyramid model, addressing several interactions at once is not a trivial task, and is not without limitations. For example, many interactions may be positive for intracellular proteins, but not for surface receptors, which is arguably the most important initial interaction in viral infection. Several computational methods applied to the study of PPI have been reviewed by Soyemi et al. (2018), and the below sections will focus on sequence-based methodologies.

Gussow et al. (2020) conducted an in-depth molecular analysis of coronaviruses to assess enhanced pathogenicity. Using comparative genomics and machine learning, the authors identified signatures present in key genomic regions, such as the nucleocapsid protein and the spike glycoprotein, which appear to be associated with higher case fatality rate and host switching. A recent study by Brierley and Fowler (2021) sought to predict the animal hosts (reservoir and intermediate) of coronaviruses. Here, machine learning techniques were applied to analyse whole genome sequences and compositional biases of the viral spike glycoprotein. The study demonstrated how evolutionary signals

in spike glycoproteins were as informative as whole genome sequences. Similarly, coronavirus spike protein sequences were used by Qiang et al. (2020) to aid prediction of species cross-over from non-human hosts of this viral taxa, and results from this work suggested that SARS-CoV-2 taxonomic relatives may indeed be of concern and should potentially be monitored. Cross species zoonotic transmission is a serious concern for public health as they may result in emerging and re-emerging infectious diseases.

2.6. Literature review summary

Efforts to predict, and ultimately inform public health policy, have been made at the ecological scale. Additional molecular methods in the form of PPI studies have been included, however, these appear to be limited to specific viral species, such as Influenza viruses in Ray and Reich (2018) and Qiang and Kou (2019) and are generated from small datasets which impacts applicability and translatability to a broader scope of pathogens. Furthermore, PPI studies focusing on receptor analysis make use of existing receptor-pathogen interactions, and as such may not be able to detect emerging pathogens which use different host receptors. Therefore, there is a need to create robust predictive models for species crossover events which identifies patterns consistent with zoonosis and may, ultimately, be used to inform public health policy to mitigate potential epidemics and pandemics.

Chapter 3: Methods and Materials

A summary of the workflow used in this study is shown in Figure 6. Simply, data was obtained from online databases and pre-processed. Thereafter the resultant data was transformed into a machine-readable format, and machine learning activities commenced, and in addition, a reproducible Nextflow pipeline was created (see Appendix I). Detailed methods are provided below, and all software dependencies have been specified in a singularity image definition file included in the supporting material (see Appendix II).

3.1. Data acquisition

Data obtained from the UniProtKB Knowledge Database (The UniProt Consortium et al., 2021) was accessed through the UniProtKB website (https://www.uniprot.org/uniprot/) on the 21st of September 2021. Using the "View by" section on the website navigation panel, "Keywords" was selected - which produced a list of dropdown items such as molecular function, domain, and biological process, to mention a few. The "biological process" dropdown was expanded to show additional dropdown options, of which "Virus entry into host cell" was selected, which can directly be accessed through https://www.uniprot.org/keywords/KW-1160. A panel on the lefthand side of the page provided a key map indicating the number of items in total for the keyword alongside the number of reviewed and unreviewed items pertaining to the keyword. The UniProtKB key map was used to obtain a data table containing the protein entries. Relevant data table fields were selected, prior to downloading in tabular format, and included; *Entry*, *Entry name*, *Status*, *Protein names*, *Organism*, *Length*, *Taxonomic lineage IDs*, *Taxonomic lineage*, and *Virus hosts*. The dataset is referred to as KW-1160 through this study. The corresponding protein sequences were also obtained in FASTA format.

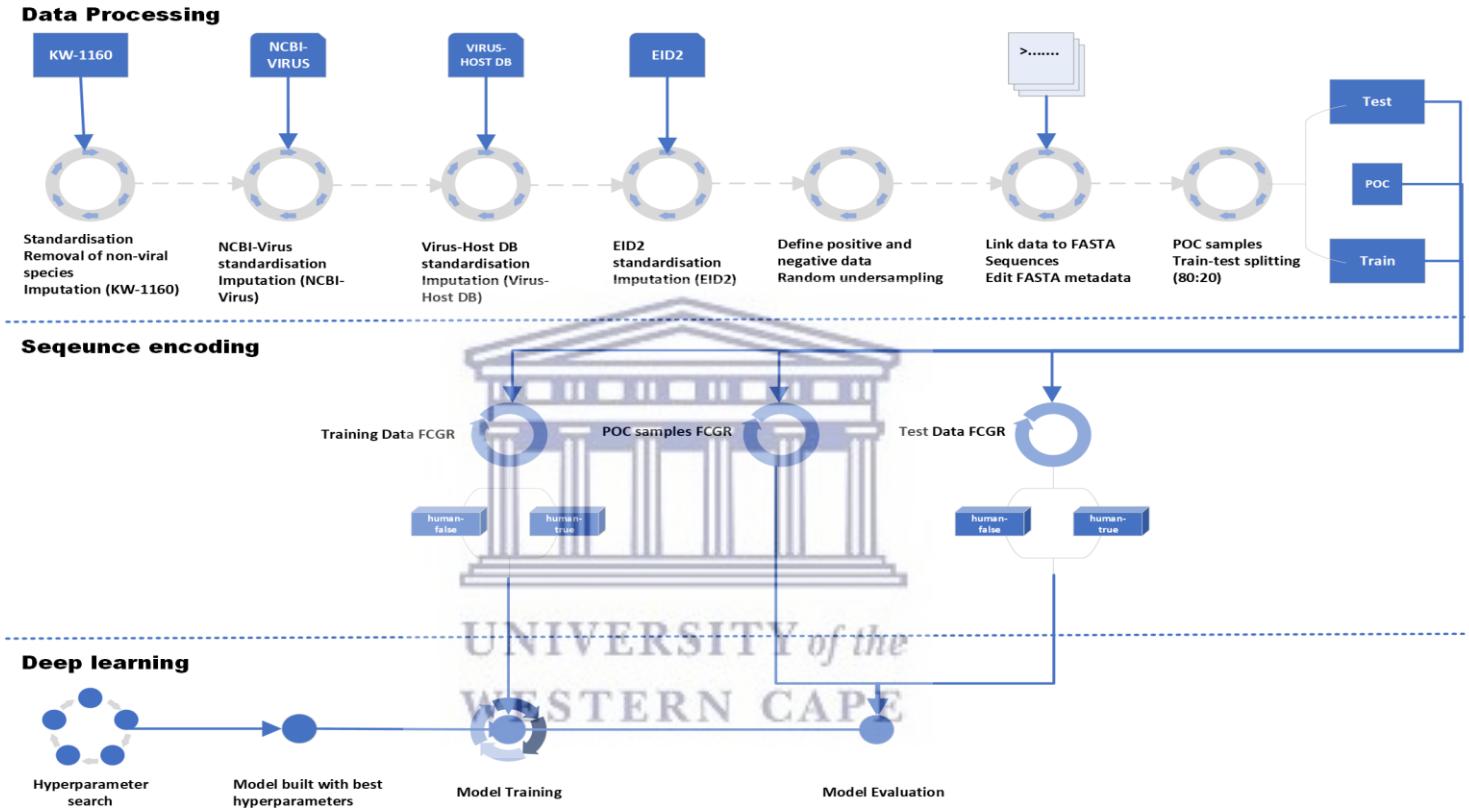

Figure 6: A summary outline of the workflow used in this study.

3.2. Data cleaning and imputation

The KW-1160 dataset was examined to determine the need for pre-processing and cleaning. A script was written using the Python programming language to automate the cleaning step (see Appendix IV), and 16 CPUs and 32GB of RAM were required for this process. The KW-1160 dataset was standardised using the *ete3 toolkit* Python package (Huerta-Cepas et al., 2016), a phylogenetic analysis package with access to the NCBI taxonomy database. During the standardisation process, the *Taxonomic lineage IDs* column of the KW-1160 dataset was standardised to the taxonomic IDs in the NCBI database. The virus species names, contained in the *Organism* column, were also standardised to the names used in the NCBI taxonomy database.

The virus organism taxonomic super kingdom and family was obtained from the NCBI database, using *ete3 toolkit*, and was added to the existing dataset in newly created columns. Microorganisms other than viruses were removed from the KW-1160 dataset. Furthermore, the dataset was intermittently grouped by virus species names and missing host information. A copy of the data was kept with only viral organism taxonomic identifiers and UniProtKB entry fields.

Additional data was obtained from NCBI Virus (Hatcher et al., 2017), Enhanced Infectious Disease Database (EID2) (Wardeh et al., 2015) and Virus-Host database (Mihara et al., 2016). NCBI Virus was accessed through https://www.ncbi.nlm.nih.gov/labs/virus/vssi/ on the 21st of September 2021. The "All proteins" option was used to obtain the data table and the *species*, *host* and *molecule type* columns were selected. The data table was then downloaded in comma-separated-value (CSV) format. Data from EID2 was downloaded as accompanying data from the study by Wardeh et al. (2015). To obtain the CSV file with species interactions, datasets were pre-processed to only include viruses, through filtering on the cargo classification column. Data from Virus-Host DB was accessed through https://www.genome.jp/virushostdb/ on the 21st of September 2021 and selected data was downloaded in tab-separated-value (tsv) file format.

Host information for viruses without this metadata was first imputed from other records in the KW-1160 dataset. Data from external sources was also used for imputation of missing host information in the KW-1160 dataset. The data from the external sources was first standardised to use the NCBI taxonomy names, followed by extraction of corresponding taxonomic IDs. The host data was also standardised to match the nomenclature in the KW-1160 dataset, in the format *[host name TaxID:ID]*. Following standardisation of the data, each of the datasets were merged with the KW-1160 dataset,

using a left-inner join, such that only records with a matching taxonomic ID would be imputed. Records which still contained missing host information following imputation, were removed from the data.

A column named *Infects human* was added to the dataset. The column contained binary data indicating whether the taxonomic ID for *Homo sapiens (9606)* was present in the list of host names in the virus hosts column. The rows matching the parameter were to be the positive data, labelled 'human true' in the *Infects human* column, and the rows which do not match the parameter were to be the negative data, labelled 'human false' in the *Infects human* column. Additional fields were added to the KW-1160 dataset which contained virus host taxonomic Superkingdom and Kingdom.

The FASTA file containing the protein sequences was then linked to the filtered UniProtKB data. The sequences were mapped to their corresponding data, based on the protein UniProtKB entry identifier in the *Entry* column of the dataset. The grouping of the KW-1160 dataset was then reversed using the copy created initially, as a reference.

The FASTA sequence headers contained the protein entry, protein name and virus species name. The protein names in the KW-1160 dataset were replaced with the protein names in the FASTA headers as a more simplified nomenclature. The headers were then modified using the information in the KW-1160 tabular data. Following modification, the header now contained the infection status, human-true or human-false, based on the information from the *Infects human* field. It also contained the unique entry, protein name and the name of the virus.

3.3. Sample size determination and train-test data splitting

Two organisms known to have crossed the species barrier and 2 others known to have not crossed the species barrier were removed from the initial dataset for later use in the model as proof of concept and their sequences were saved, these are referred to as the POC data. The majority class, based on the *Infects human* field, was randomly down-sampled to be 67% more than the minority class using the *imbalanced-learn* Python package (Lemaître and Nogueira, 2017). The KW-1160 dataset was then saved as a compressed CSV file. Thereafter, the KW-1160 data was randomly split into training and test data at a ratio of 80:20, respectively. The training and test data was saved as sequences, in the FASTA file format, in separate train and test directories.

3.4. Sequence encoding

To convert the FASTA sequences into machine readable input, frequency chaos game representation (FCGR) was utilised. A script was written in the R programming language to automate the process (see Appendix IV). The sequence headers were first edited to remove unwanted meta-characters such as a forward slash (/) and the pipe character (|) as some of the header information was to be used in naming the outputs. Frequency chaos game representation was then performed on the POC, training and testing sequences using the *kaos* R package (Löchel et al., 2020). The resolution parameter was set to 100, the mode parameter was set to 'matrix' to produce a frequency matrix, and the 'labels' parameter was set to false. The resulting plots were saved as portable network graphics (PNG) images of 224x224 pixels (width x height). The images were saved in human-true and human-false sub-directories, in each of the train and test directories, as this is the format required by the machine learning Python package (*keras*) used in this study. The POC data was not saved in directories. To increase efficiency, asynchronous and parallel programming was implemented making use of *later* and *parallel* R packages, respectively (Zhao, 2016). A computational cluster compute node with 32 CPUs and 40GB of RAM was used for this process.

3.5. Deep Learning, hyperparameter searching and model architecture

Deep learning using convolutional neural networks (CNNs) was used to build the classification model. To deduce the number, and types, of layers to include in the model, hyperparameter tuning was performed. The hyperparameters chosen for building the classification model were; number of 2D convolution layers (between 1 and 3 layers), number of units in the convolution layer (between 48 and 128 units), threshold of evaluation metrics (between 0.5 and 0.9), optimizers (RMSProp and Adam), and optimizer learning rate (0.001 to 0.019 with an incremental value of 0.002). The *Keras-tuner* Python package was used to implement Bayesian hyperparameter search (O'Malley et al., 2019).

The model was created using the *keras* package (Chollet et al., 2015) and the *TensorFlow* (Tensorflow Developers, 2022) package as a backend in the Python programming language (see Appendix IV). The hyperparameters obtained from the hyperparameter search were used to construct and compile the model.

3.5.1. Model training and validation

The model was trained using the training data FCGR images. Twenty percent of the training data was used to validate the model on every epoch. The model was set to train for 50 epochs in data batches of 64 images. Furthermore, the images were shuffled on each iteration. Model checkpoints were created only saving the best overall model through all the epochs. The model was trained on the Ilifu cluster using a computational node with 16 CPUs, 32GB of RAM and a 12GB GPU. Memory growth for the GPU was enabled to prevent training failure due to memory. The training was further adapted to use multiple GPUs with *TensorFlow* mirrored distribute strategy, should more GPUs be needed.

3.5.2. Model evaluation and proof of concept

The model was tested using the test data. The performance metrics of accuracy, precision, recall, f1 score, area under the receiver operating characteristic curve (ROC-AUC) and the Matthews correlation coefficient (MCC), true positive values, false positive values, true negative values, and false negative values, were captured. Furthermore, the model was used for proof-of-concept prediction using the POC samples.

Chapter 4: Results and Discussion

Several epidemics and pandemics are linked to host switching by viral pathogens, originally established in an animal host or reservoir (Parrish et al., 2008). Epizootic and zoonotic disease are driven by spill-over of a pathogen to a previously unexposed, non-susceptible host and when these events occur, the resultant outbreaks can have devastating consequences (Parrish et al., 2008). Despite the clear threats to public health and biosecurity which are caused by emergence and re-emergence of zoonotic diseases, many host crossover events are not detected, or reported, and the modelling of infectious disease to predict spill-over remains constrained by several challenges (Glennon et al., 2019; Roberts et al., 2021). Computational approaches have been used to predict zoonotic potential of pathogenic species from a biological perspective (Cho and Son, 2019; Qiang and Kou, 2019) and in this study, a machine learning approach was used to develop a model capable of learning protein sequence patterns of viral pathogens. Considering that host specificity is critically dependent on viral interaction with host cells, receptor binding (and changes thereof) inevitably plays a vital role (Parrish et al., 2008), and as such, viral proteins involved in pathways of host cell entry were used to train, validate, and evaluate the model. The positive dataset used in this study consisted of viral pathogens known to infect human hosts, while those documented to not have a human host formed the negative dataset. The trained model could then be used to predict if an unknown virus would be capable of infecting a human host cell, based on consistency of protein sequence patterns, learned during model training.

4.1. Dataset description and exploratory data analysis

Machine learning is a sub-field of artificial intelligence primarily focused on enabling machines to learn patterns from large scale empirical data and convert the knowledge into usable models, without any explicit programming (Edgar and Manz, 2017; Thessen, 2016; Woolf, 2009). High quality data is important in data analytics as it often determines the quality of the subsequent analysis (Chu et al., 2016). Furthermore, machine learning classification models are sensitive to data quality, therefore high-quality data is a priority for generation of robust models (Klie et al., 2022). In this study, sequence data and accompanying metadata was obtained from a comprehensive and trusted database resource for protein sequences and annotation data (The UniProt Consortium et al., 2021). The data consisted of a total of 358333 data entries and 9 fields. The metadata fields selected for ML activities in this study are shown in Table 1.

Table 1: Details of the metadata fields of the data obtained from UniProt.

Field	Description
Entry	UniProt unique entry identifiers.
Entry name	UniProt unique entry identifiers and an abbreviation of the virus name separated by an underscore
Status	Records of whether the data was reviewed and annotated by UniProtKB curators (reviewed) or Computer-annotated (unreviewed).
Protein names	The names of the protein associated with the entry.
Organism	Refers to the organism of which the protein was extracted. Includes isolates and location of which it was identified.
Length	The amino acid length of the protein sequence associated with the entry.
Taxonomic lineage IDs	Refers to the taxonomic ID of the species in question.
Taxonomic lineage (SPECIES)	The organism genus and species name without the any additional information.
Virus hosts	A list of hosts of the virus in question separated by a semi-colon. At the end of each name, the host taxonomic ID is given in the format [Tax: *ID*].

The *Organism* field was selected to identify the host organism from which the virus was isolated or identified; however, inconsistencies and missing labels in the data points were observed. For example, a West Nile virus entry did not contain strain, isolate, and date information, which was present in an Influenza A virus entry. The observed inconsistencies may be due to the metadata simply not being collected, individual reporting errors when uploading the information to the database, and errors from the UniProtKB automated annotation, a phenomenon which is often observed even if gold standards are used (Klie et al., 2022). Additionally, there may be protein entries which have been computationally predicted and therefore lacking the additional information which is otherwise captured in experimentally derived entries. Generally, with consistent data, the information can be extracted by matching consistent patterns. However, with the observed inconsistencies, it would require special data mining algorithms to efficiently extract the required data, which is by no means a trivial task. Due to time limitations, this was not performed in this study, and is noted to be a limitation, as greater data consistency could have resulted in generation of a more specific model of better quality.

Two thousand, five hundred and nineteen (2519) protein names were represented in the data. However, numerous entries contained long and ambiguous nomenclature, for example, entry L7WIQ8 for HIV1 contained the protein named ‘Envelope glycoprotein gp160 (Env polyprotein) [Cleaved into: Surface

protein gp120 (SU) (Glycoprotein 120) (gp120); Transmembrane protein gp41 (TM) (Glycoprotein 41) (gp41)]'. To resolve the ambiguity with respect to nomenclature, the names defined in the FASTA headers of the sequences were used. This was necessary because the protein names were used in downstream processes. Additionally, entries in the KW-1160 dataset with the same protein name were observed to have different amino acid lengths. This may be due to the submission of partial sequences by different researchers. For example, entries A0A7G4JM42 and Q67278 for Influenza A virus Nucleoprotein had 426 and 468 amino acids, respectively. This protein is known to have a length of 468 amino acids (Reid et al., 2004). These data occurrences were not, however, excluded from the dataset as they may introduce sequence variation, in a similar manner to strain level genetic variations, thereby lessening model bias.

In order to classify the positive and negative dataset, the *Virus hosts* field was used to indicate whether a viral pathogen was documented to have a human host. The prime objective of the model was to predict pathogens with the potential to cross the species barrier and infect humans, and as such, viruses which are reported to successfully infect humans would be classified as positive (with the assumption that they did not originate in the human host) and others, classified as negative. It should be noted that the holistic OneHealth approach was undertaken in this study, with plant and non-eukaryotic pathogens included in the negative training dataset, to enrich the data used for training. Explorative analysis was performed to ensure consistency in the data prior to downstream use, however, this revealed incomplete data, shown in Figure 7. A total of 237573 entries with missing data was observed, of which 53 were from manually annotated records, i.e., those which have been extracted from literature and curator evaluated (The UniProt Consortium et al., 2021). The majority of records (237520) with incomplete metadata was observed in unreviewed records i.e., those which await full manual annotation (The UniProt Consortium et al., 2021). This was an important observation as the existence of incomplete data would have a negative impact on classification and, ultimately, the model (Li et al., 2021). Li et al. (2021) investigated the impact of data cleaning on machine learning classification and recommend that machine learning researchers and engineers prioritize evaluation of the data fed into models. Furthermore, Frenay and Verleysen (2014) also identified that erroneous labels negatively impact machine learning models.

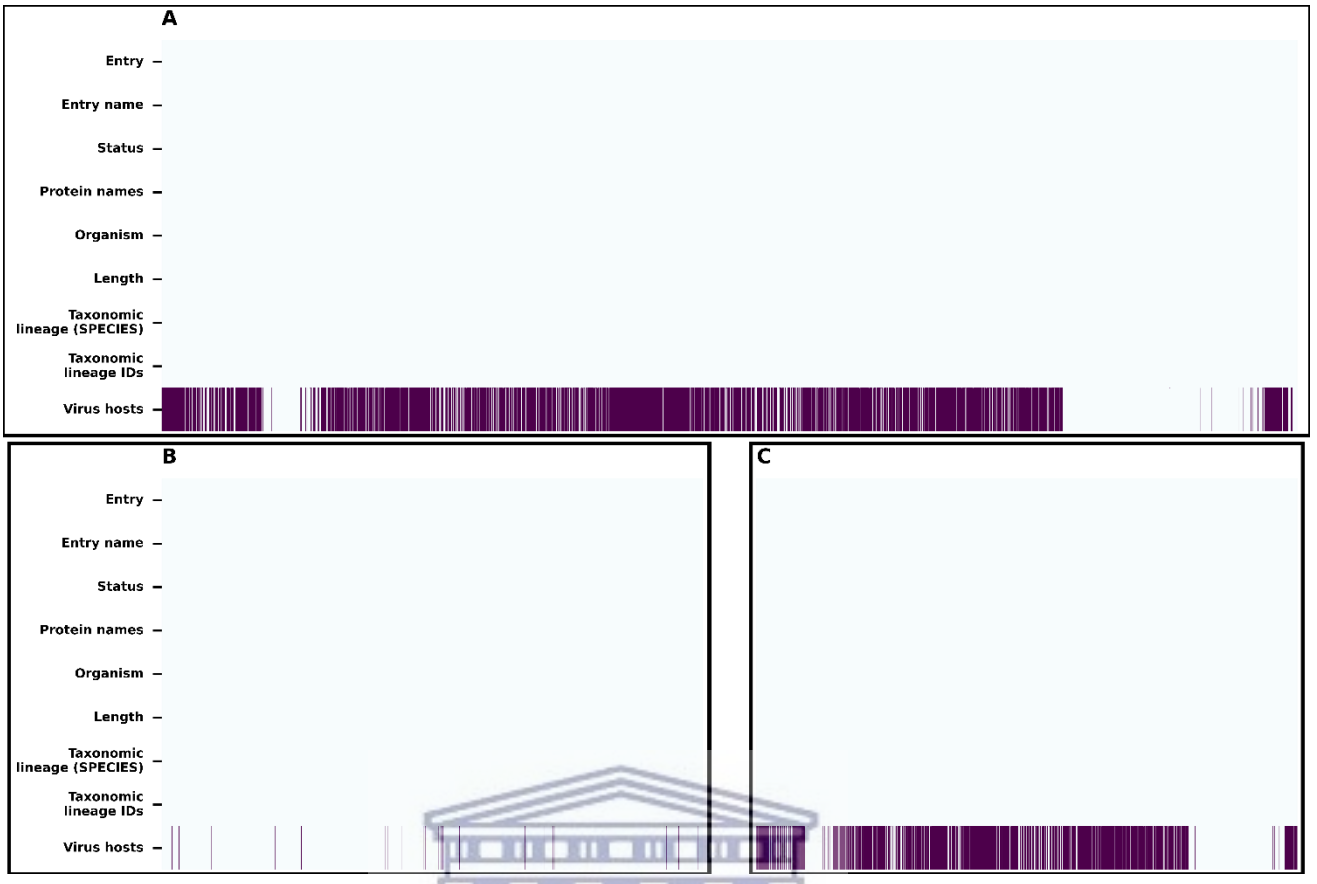

Figure 7: Visual representation of missing data in the KW-1160 dataset prior to data cleaning. Purple bands indicate records in the dataset which have data missing. A: The total number of records with missing information in the KW-1160 dataset (237573 missing records). B: Missing data from the KW-1160 dataset classified as reviewed (53 missing records). C: Missing data in the KW-1160 dataset classified as unreviewed (237520 missing records).

Exploratory data analysis performed in this study also revealed that the data did not exclusively contain entries from viral pathogens. This was a relatively surprising observation as the dataset was expected to contain only viral data, based on the search criteria used to obtain the data. This is a noted ontology weakness in EMBL-EBI gene ontology (GO) Annotation blacklist (https://www.ebi.ac.uk/QuickGO/term/GO:0046718) which is used by the UniProtKB database for annotation (The UniProt Consortium et al., 2021).

Interestingly, Nucleoprotein from Influenza viruses, was classified along with other proteins which facilitate viral entry into host cells, in the UniProt database (e.g., entries A0A7G4JM42 and Q67278). This protein encapsidates the viral RNA but has been reported as a determinant of host switching and host specificity (Long et al., 2019; Selman et al., 2012). While this entry was included in the study, this

highlights that filtering data to simply exclude non-viral sequences, may still result in erroneous inclusion of incorrectly curated and annotated entries.

There is an enormous amount of data in databases, and it continues to grow; however, manual curation remains slow and automated annotation, and curation, is error prone (Klie et al., 2022). Therefore, it is recommended that data checking be implemented in the initial stages of data mining prior to use, even when data is obtained from highly trusted resources.

Figure 8 shows the distribution of all micro-organisms in the dataset. There was a total of 7903 Virus species from 182 families and 6833 Bacterial species from 799 families. Additionally, there were 115 organisms from the *Eukaryota* (85 families), 7 from *Archaea* (7 families), 15 metagenomes, 2 uncultured organisms and 2 plasmids. This study specifically examined viral pathogens and as such, any entries which were not classified in the *Virus* superkingdom were removed prior to commencing downstream activities.

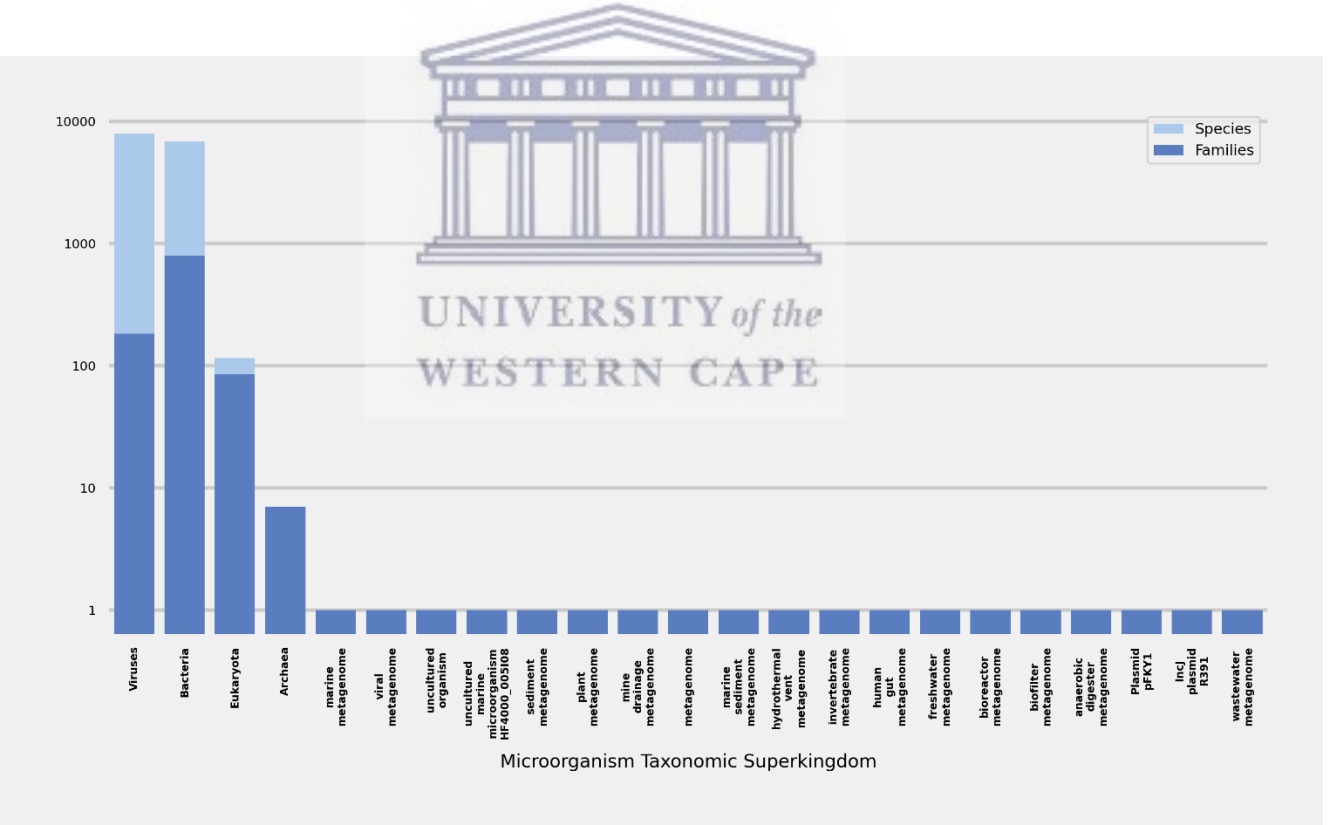

Figure 8: The taxonomic super kingdoms of the microorganisms observed in the KW-1160 dataset and the number of families and species of each super kingdom. On the x-axis is the super kingdom of the parasitic organism and on the y-axis is the number of species or families.

4.2. Data imputation and cleaning

Missing data is reported to be a commonly observed challenge in statistical analysis and machine learning which introduces bias into analyses, and as such, several techniques have been developed in an attempt to minimise the impact of missing data (Johnson and Khoshgoftaar, 2019; Liu, 2016). The most common of these approaches includes; 1) continuation of analysis with missing data without any adjustment, 2) removal of missing data, and 3) imputation of missing data with suitable estimates (Piquero and Carmichael, 2005). The deep learning approach to machine learning used in this study is negatively impacted by missing data, and as such data imputation with suitable data from external databases was used to complete the data (Li et al., 2021). Data size and data variation have been shown to have a significant impact in performance for image classification tasks (Keshari et al., 2020). Smaller datasets often result in models with poor generalization and high overfitting (Keshari et al., 2020) and as such, removal of missing data from the current dataset would have resulted in a significant reduction of the dataset size, from 358333 to 2373 records, and was not a feasible option in this study. Instead, a strategy of data imputation with suitable estimates was employed in this work.

Data obtained from NCBI Virus (Hatcher et al., 2017), Virus-Host DB (Mihara et al., 2016) and EID2 (Wardeh et al., 2015) was used for imputation. Tables 2, 3 and 4 detail the fields selected for use from each of the databases. These data sources are considered secondary databases because they obtain data from primary, archival, databases such as GeneBank which contains experimental results, directly submitted by researchers. Additionally, these databases also use computational and manual analysis to derive knowledge from the primary databases (The UniProt Consortium, 2021).

Table 2: Information contained in the data fields selected from the data obtained from NCBI

Field	Description
Species	The genus and species name of the virus.
Molecule_type	The type of nucleic acid material, DNA or RNA, contained in the virus. The data goes to further classify them based on their strand types, single or double stranded. Viruses with unknown molecule types are labelled as unknown.
Host	The genus and species name of the host, each row has only one host. However, viruses with multiple hosts are repeated, each time with a different host.

Table 3: Details of the data fields selected from the data obtained from the Virus-Host

Field	Description
virus tax id	The taxonomic identifier of the virus.
virus name	The genus and species name of the virus.
host tax id	The taxonomic identifier of the host, each row has only one host and viruses with multiple hosts are repeated, each time with a different host.
host name	The genus and species name of the host, each row has only one host. However, viruses with multiple hosts are repeated, each time with a different host. If the species name is unknown only the genus name is given followed by *sp*.

Table 4: Detailed information on the data fields used from the data obtained from EID2

Field	**Description**
Cargo	The genus and species name of the virus. Viruses with multiple hosts are repeated, each time with a different host.
Carrier	The genus and species name of the host, each row has only one host. Similar to the Virus Host DB dataset, if the species name is unknown only the genus name is given followed by *sp.*

Due to the use of multiple data sources, there was a need for standardization across the datasets to maintain consistency and interoperability (Bhalla et al., 2017). Taxonomic IDs are regularly updated in the NCBI taxonomic database, and as such, data records created at different time periods may have different identifiers. However, the NCBI taxonomic database archives all taxonomic IDs when updated (Schoch et al., 2020). The *ete3 toolkit* used in this study downloads the NCBI taxonomy database, enabling local storage of a fixed taxonomic database. All taxonomic records were translated to the version downloaded in the *ete3 toolkit* to allow interoperability between the datasets used.

The NCBI Virus database is a secondary database which obtains RefSeq records as soon as they are updated (Hatcher et al., 2017). In addition, the database contains records from literature, is subjected to automated and manually curation, and was the first data source used for imputation activities in this study. Virus host DB, which manually curates data from multiple databases, such as RefSeq and ViralZone, was used for the second round of imputation (Mihara et al., 2016). The EID2 was used for a third imputation pass, and contained data from primary databases such as RefSeq, with additional records extracted from PubMed (Wardeh et al., 2015).

Figure 9 depicts the decrease in missing data at the end of each imputation pass from the various databases. The attempted imputation from other records in the KW-1160 data had no impact and this was to be expected, under the assumption that the automated curation algorithm used by UniProtKB conducts a similar operation to the one performed in this study. The NCBI Virus dataset was expected to contribute significantly, however, a small impact was observed. The expected impact of imputation, versus what was observed, may be explained by database development and redundancy, especially given that primary databases are known to contain data redundancy (Chen et al., 2017) and that NCBI Virus curates data directly from RefSeq. The Virus-Host DB provided the most notable contribution as seen in Figure 9. This observation may be due to the manual curation activities of this database, as the data is also collected from primary databases such as NCBI and EBI (Mihara et al., 2016), thereby containing higher quality data compared to the other databases used. Following imputation from the 2 databases it was expected that the last database used, EID2, would only have a small impact, and this was indeed the case.

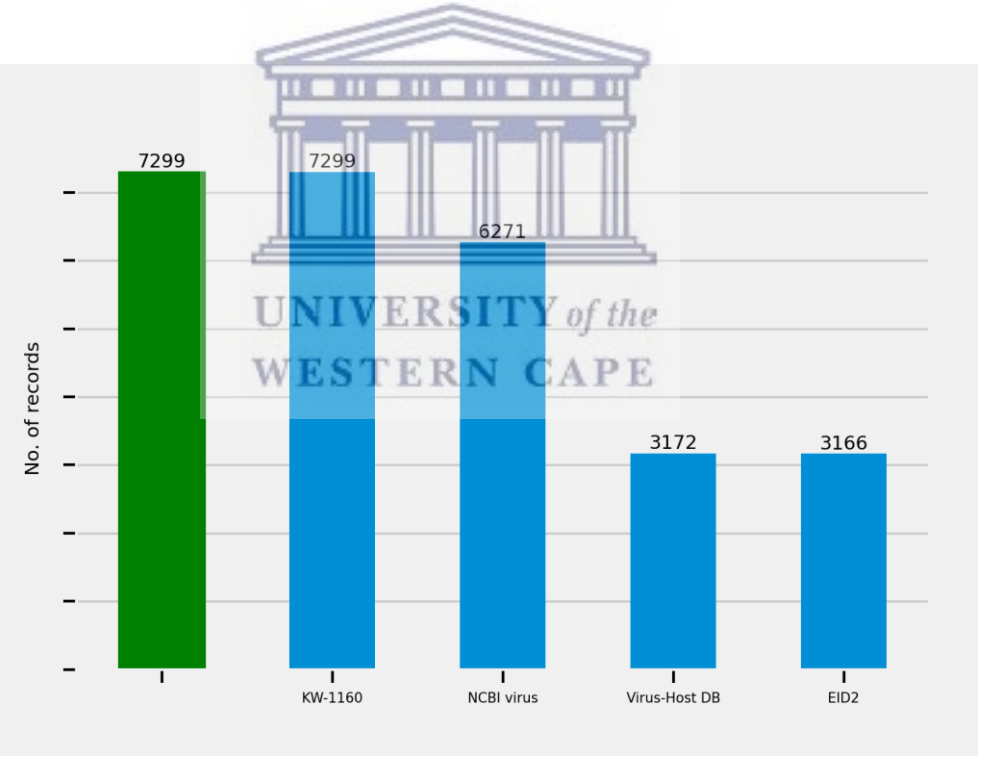

Figure 9: Visual representation of the change in data dimension through the varying imputation steps. The figure depicts the reduction of missing values following each data imputation pass. On the y-axis is the record counts of missing values and on the x-axis is the dataset used for imputation. The first bar represents missing data in the KW-1160 dataset prior to imputation. The remaining bars are labelled on the x-axis. The labels represent the database used for imputation.

Explorative data analysis and data cleaning was initially performed on a local machine prior to obtaining access to a high-performance computational environment. When performing imputation on a local machine, high resource usage was observed, often resulting in malfunction and shutdown. This was due to memory (RAM) limitations which resulted from performing the operations on over 300000 records. Dimensionality reduction (DR) is a technique which reduces data dimensions but retains significant patterns within the data (Nguyen and Holmes, 2019). Feature extraction and feature selection are 2 common types of DR, each with different algorithms (Nguyen and Holmes, 2019). Feature selection methods include filter, wrapper and hybrid algorithms while feature extraction methods include simple data grouping, principal component analysis (PCA) and correspondence analysis (CA) (Nguyen and Holmes, 2019). The reader is referred to Nguyen and Holmes, 2019 for in-depth reviews of DR methods. To circumvent the computational limitations, basic DR was used, and the data was grouped by virus species taxonomic identifiers. This reduced the records from 358333 to approximately 7000. A copy of the dataset, was made prior to dimensionality reduction, retaining only the entry and virus species taxonomic identifier fields. This copy was later used to revert the data back to the original dimensionality.

Table 5 details the contents of KW-1160 following pre-processing, with 40772 records which still contained incomplete metadata, being removed. The discarding of this proportion of records was considered to have a relatively small impact when compared to the option of complete removal of data records with incomplete information prior to imputation and cleaning (Section 4.1, page 24).

Table 5: Details of the data fields in the KW-1160 dataset after data cleaning

Field	Description
Entry	UniProtKB unique entry identifiers.
Species name	The genus and species name of the virus.
Species taxonomic ID	The taxonomic identifier of the virus.
Species family	The name of the taxonomic family virus.
Virus hosts	A list of hosts of the virus in question separated by a semi-colon. At the end of each name, the host taxonomic ID is given in the format *[Tax: ID]*.
Virus hosts ID	A list of hosts of the virus in question separated by a semi-colon.
Host kingdom	The taxonomic kingdom of the virus hosts. If there are 2 viruses hosts of different kingdoms they are included as a semi-colon separated list.
Host superkingdom	The taxonomic super kingdom of the virus hosts. If there are 2 viruses hosts of different kingdoms they are included as a semi-colon separated list.
Molecule type	The type of nucleic acid material, DNA or RNA, contained in the virus. The data goes to further classify them based on their strand types, single or double stranded. Viruses with unknown molecule types are labelled as unknown.
Infects human	The infection status of the virus in question. Records with a human-true label correspond to viruses which have a human host and records with human-false label correspond to viruses which do not have a human host.

The study aimed to produce a predictive model which informs if a given viral sequence entry infects, or could potentially infect, a human host. Such a model would have either a positive, or negative, class and required strict labelling of the training data (Alaeddine and Jihene, 2020). To define the classes, all viral entries reported to have a human host were considered as the positive data class (labelled human-true), even if the viral pathogen infected additional non-human hosts. All entries with no reported human host, irrespective of the scope of hosts, were considered as the negative data class (labelled human false).

4.3. Undersampling and splitting

A class imbalance was observed in the data, whereby the positive class had substantially more data points (278791 entries) when compared to the negative class (38770 entries). While the imbalance was remarkably high, the observation was consistent with existing literature which demonstrates that non-human infection events are understudied and are often not reported, and as such, until such issues are addressed, a bias is unavoidable (Glennon et al., 2019; Parvez and Parveen, 2017). Class imbalance is

a common problem in machine learning and random undersampling (RUS) of the majority class, and random oversampling (ROS) of the minority class, are techniques often used to approach data imbalance in statistics and machine learning (Hasanin et al., 2019; Johnson and Khoshgoftaar, 2020). However, there are no specific values to employ for either ROS or RUS. Random undersampling (RUS) was used in this study to reduce redundancy while maintaining variation in the dataset, and due to randomized selection, all proteins would be represented even if the majority class was down sampled.

A small-scale study was conducted, and the results obtained were used to ascertain the undersampling ratio for the main study. Three models were trained using data of varying proportions derived from the KW-1160 dataset, and following evaluation, the most optimal model, in terms of performance, was used to inform the undersampling ratio. One dataset was an undersampled derivative with the majority class being only two-thirds (67%) greater than the minority class (*ZoonosisTwoThirds* - 57862 entries for the positive and 38768 entries for the negative class). The second undersampled derivative dataset represented equal proportions of the majority and minority class, containing 38768 data points for each. The third dataset represented the complete KW-1160 dataset with no modifications (*ZoonosisFull* - 278789 entries for the positive and 38768 entries for the negative class).

Table 6 shows the performance of the models on test data. As indicated by Chicco and Jurman (2020), if a high accuracy model is observed and is accompanied by a low MCC, it indicates the likelihood of the model being trained on imbalanced data, and this was indeed observed for the complete KW-1160 dataset with no modifications (*ZoonosisFull*). The 67% RUS method used in *ZoonosisTwoThirds* model was selected as the favourable sampling method as it allowed for the presence of an adequate volume of data that exhibits the least amount of dataset imbalance, and which is not too low to introduce overfitting of the model. Overfitting, or model variance, in statistics and machine learning is defined as a phenomenon whereby a machine learning model perfectly fits training data, achieving high training accuracy, but fails to generalise on unseen data, and subsequently achieves low validation or test accuracy (Edgar and Manz, 2017).

Table 6: Evaluation metrics of the models obtained from training at different sample sizes.

Name	Accuracy	True Positive	False Positive	True Negative	False Negative	MCC	F1 Score	ROC-AUC
ZoonosisFull	93.57	94,772	1,767	12,092	5,575	0.74	0.96	0.97
ZoonosisTwoThirds	95.37	20,274	775	13,061	844	0.90	0.96	0.99
ZoonosisOne2One	97.14	13,211	129	13,917	671	0.94	0.97	1.00

4.4. Training, validation, and testing datasets

Machine learning models 'learn' on larger training data and are evaluated on smaller test data (Schilling, 2016). In some cases, evaluation during model training is implemented to detect poor performance, using a separate, smaller dataset, referred to as the validation dataset, (Schilling, 2016). Model training can be a time and resource consuming process, and as such, early detection of poor performance allows early termination of model training, followed by adjustment of the model architecture (Schilling, 2016). In this study, the train-to-test ratio of 80:20, a ratio that is commonly used (Alaeddine and Jihene, 2020; Wu et al., 2018), resulted in 77304 randomly selected entries being selected for use in training and 19326, for testing. Twenty percent (20%) of the training data was later defined as validation data at the beginning of training.

4.5. Sequence feature encoding

Machine learning algorithms often require machine readable, numeric or byte data, as input (Modarresi and Munir, 2018) and efforts have been made in bioinformatics and protein engineering to represent text-based FASTA protein sequences as machine readable input, while maintaining sequence integrity (Jing et al., 2020). Chaos game representation (CGR) is a sequence representation scheme inspired by chaos theory in physics, originally proposed by Jeffrey (1990), as a visual representation scheme for DNA sequences. Frequency chaos game representation (FCGR) is an adaptation of the original CGR method and has been modified to accommodate protein sequences (Löchel et al., 2020). Another variant of CGR proposed by Mu et al. (2019), called DCGR, incorporates amino acid physiochemical attributes, which are important determinants of protein structure, interaction, and function. However, the latter is implemented in MATLAB, a proprietary software, and requires privileged access. As such, FCGR was used in this study as it is an open-source source software which is easy to implement.

Frequency chaos game representation, employed in this study, generated greyscale images of 224x224 pixels. Examples of the generated features for 4 entries are shown in Figure 10. The FCGR image is a large icosagon, which contains twenty edges and twenty icosagons (Löchel et al., 2020), with the edges representing each of the 20 standard amino acids.

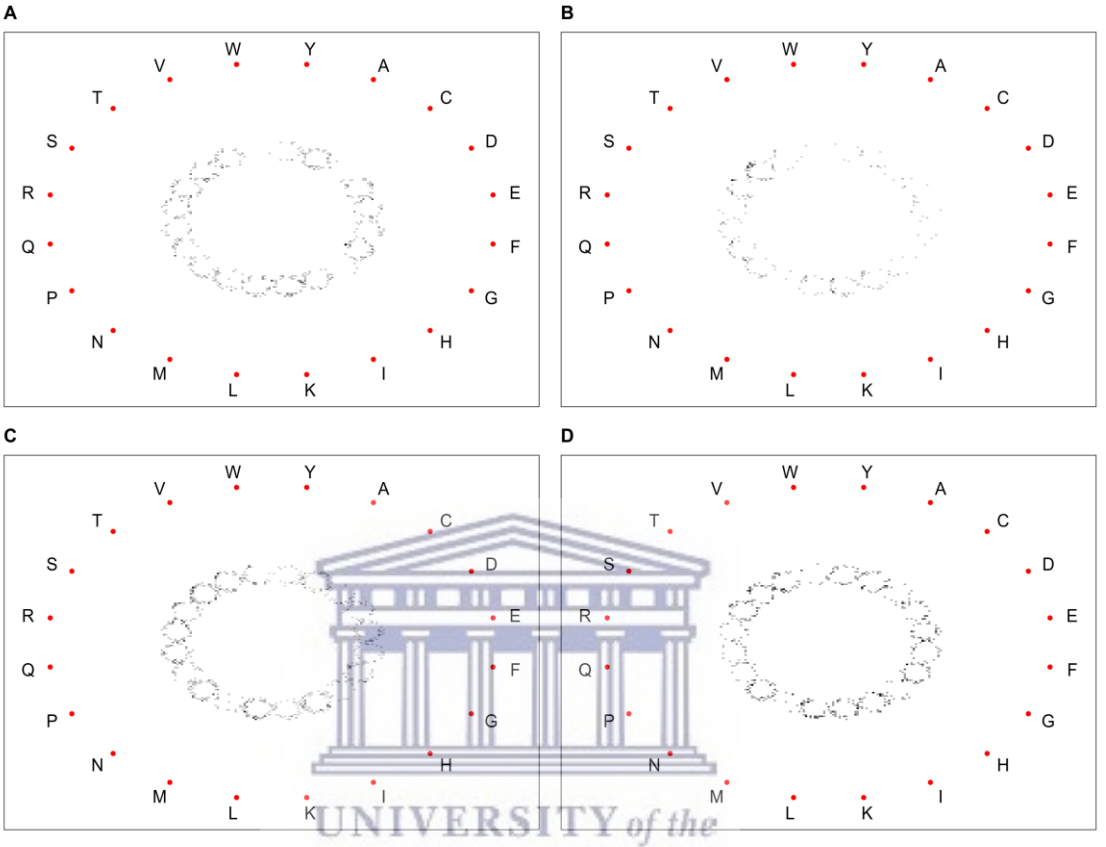

Figure 10: The frequency chaos game representation (FCGR) of 4 virus surface proteins. A: Influenza B virus Nucleoprotein (560aa). B: Human orthopneumovirus Major surface glycoprotein G (315aa). C: Simian immunodeficiency virus Envelope glycoprotein gp160 (865aa). D: Influenza A virus Hemagglutinin (566aa).

The FCGR implementation automatically detects the sequence type from a given input and the generated image only includes 20 letters corresponding to the standard proteinogenic amino acids (Steward, 2019). However, sequencing errors occasionally occur in the sequencing of protein isolates such that an amino acid cannot be clearly identified (Pietrzyk et al., 2013; Searle et al., 2004; Vyatkina et al., 2015). A precise distinction between aspartic acid or asparagine, or glutamic acid or glutamine is often an issue when sequencing by chromatography, mass spectrometry, and X-ray crystallography (Pietrzyk et al., 2013). Such errors may result in the presence of different letter representations - **B** (aspartic acid [A] or asparagine [N]), **J** (leucine [U] or isoleucine [W]), **X** (unknown amino acid), **Z**

(glutamic acid [G] or glutamine [Q]) - in the sequenced proteins. There is no record of how the software handles these cases and it is assumed that these letters are omitted by the FCGR software. However, this may not be a significant issue as a majority of proteins in databases are translated from nucleic acid sequences (The UniProt Consortium, 2021). Although the software supports other letters when the alphabet parameter is set to 'LETTER', for upper case letters, or 'letter', for lower case letters, this would require experimentation with the scaling factor to prevent the production of unexpected or erroneous results (Löchel et al., 2020).

There are two additional, recently discovered amino acids; selenocysteine, and pyrrolysine, represented by the letters U and O, respectively (Lopez and Mohiuddin, 2022). Pyrrolysine has only been found in proteins from several methanogenic organisms such as archaea and bacteria (Rother and Krzycki, 2010), and selenocysteines are present in proteins which facilitate redox reactions (Mariotti et al., 2018; Rother and Krzycki, 2010). It is not clear how FCGR would deal with these amino acids should they occur in a given protein sequence, and while this was not of high concern in this study, which was focussed on viral proteins, it is an important consideration when FCGR is applied to studies involving proteins containing these unique amino acids. Additional consideration with FCGR is related to the protein sequence length. The FCGR software has not been tested on peptides and smaller protein sequences (Löchel et al., 2020), and as such, the effects of smaller sequence lengths are unclear and further research into FCGR performance in relation to sequence length should be investigated. Considering that the data used in this study contained sequences of varying lengths, it is possible that shorter sequences may have had poor representation.

The implementation of FCGR with the *kaos* package was slow, due to the large number of sequences, as well as the existence of sequences with long lengths, in the dataset. The low performance was due to a bottleneck effect, created when the FCGR images were saved (Goranova et al., 2015). Saving images is considered a part of input-output operations and increasing the speed at which this operation occurs, requires the use of multiple threads in a central processing unit (CPU) (Goranova et al., 2015). The R programming language does not support multiple thread operations, however, processing on multiple CPUs through additional packages such as *parallel* and *foreach* is possible (Zhao, 2016). Multiprocessing was implemented using *parallel*, demonstrating a significant increase in performance. Studies have indicated that the use of asynchronous programming in single thread programming languages, such as the R programming language (Goranova et al., 2015; Zhao, 2016), may offer significant performance improvement especially with input-output operations. In synchronous

programming, tasks are run in a sequential manner such that a task cannot begin before the prior task has run to completion. In asynchronous programming, a new task is initiated while awaiting completion of the current task in operation. When all the tasks are completed, the remaining sequential tasks are executed, making the entire process more efficient (Abadi et al., 2016; Goranova et al., 2015). The coupling of multiprocessing and asynchronous programming in this study, resulted in significant performance improvements to the FCGR (see Appendix VII for benchmark results).

4.6. Model architecture

Deep learning applies non-linear transformations through layered networks, termed artificial neural networks (ANNs) (Valueva et al., 2020; Vargas et al., 2018). A convolution neural network (CNN, or ConvNet) is a type of ANN often used in computer vision to analyse images (Valueva et al., 2020). Convolutional neural networks were used in this study due to their exceptional image classification capability, particularly for the FCGR images. The FCGR image features output in this study were greyscale and as such 2D CNNs were used for training. There is no convention for building CNN models due to varying performances of different model architectures, and as such, it is the researcher's or engineer's task to find an optimal model architecture for a given dataset (Lu et al., 2019). This is often complicated by the presence of a multitude of parameters which require adjustment (Sarawagi and Ganguli, 2021; Klein et al., 2017) and can include the number of layers to use, number of nodes within each layer, and in some instances, the parameter itself may have values which require adjustments, such as the learning rate of an optimizer (Sarawagi and Ganguli, 2021). Hyperparameter optimization is an approach used to identify the best parameters from a defined set of values (Klein et al., 2017; Lu et al., 2019).

Bayesian hyperparameter optimization was the optimisation approach used in this study due to efficiency, and lower resource consumption (Klein et al., 2017). The search space, defined as the set of all possible values (Lu et al., 2019), in this study is summarised in Table 7. Deep learning models produce probability values when given certain input, and as such a threshold hyperparameter value was included to help classify output values, such that entries with a probability value above the threshold would be classified as positive and entries with a probability value lower than the threshold would be classified as negative. The hyperparameter search space size was 19440, composed of 5 node values, 3 kernel sizes, 3 pool sizes, 8 threshold values, 9 learning rate values, 3 layers, and 2 optimisers. The search was implemented for 500 trials, 500 combinations of hyperparameters, and training for 2 epochs per trial, with validation at the end of each epoch. One seventh (5538 samples) of the data was used to

search for hyperparameters. Five-hundred (500) trials were executed on this reduced data, to prevent the cessation of trails prior to completion due to limited GPU memory, even when distributed across multiple GPUs. While RUS evaluation determined that the ZoonosisTwoThird dataset was the optimal sampling methodology, the ZoonosisOne2One dataset was used for the hyperparameter optimisations as it was assumed that a balanced dataset would give equal importance for both the positive and negative class because the trials ran for a limited number of epochs.

Table 7: The hyperparameter search space summary. Minimum and Maximum refer to the lowest and highest hyperparameter values, respectively, for numeric hyperparameters. Step refers to the increment value from the minimum to the maximum value for numeric input.

Hyperparameters	**Minimum**	**Maximum**	**Step**		
Nodes	48	128	16		
Kernel size	1	3	1		
Pool size	1	3	1		
Optimizer				Adam	RMSprop
Threshold	0.5	0.9	0.05		
Learning rate	0.001	0.019	0.002		
Layers			1 Conv layer	2 Conv layers	3 Conv layers

The trials were rated based on the MCC score obtained from the validation at each training iteration such that the combination of hyperparameters with the highest MCC would have a high rating. Bayesian hyperparameter optimization employs Bayes theorem, using prior values to inform the next choice of parameters (Dewancker et al., 2016). Only 5 of the 8 threshold values were used, 3 of the 9 learning rate values, and the hyperparameter values for nodes, kernel size, pool size and layers were all exhausted. The best values for each hyperparameter were often repeated indicating that their inclusion consistently produced high performing models. The top 10 hyperparameter sets are shown in Table 8, with the best hyperparameters for our model being: 1 Conv layer with 128 units, kernel size of 2, a max pool with a pool size of 2 and an Adam optimizer with 0.019 learning rate and metrics would be measured at a threshold of 0.5.

Table 8: The top 10 combinations of hyperparameters. The results for all 500 trials can be found in Appendix V

Nodes	Kernel size	Pool size	Optimizer	Threshold	Learning rate	Layers	Validation MCC
128	2	2	Adam	0.5	0.0190	1 Conv layer	0.8145
128	2	2	Adam	0.5	0.0190	1 Conv layer	0.8133
48	3	3	RMSprop	0.5	0.0190	1 Conv layer	0.8124
48	3	3	RMSprop	0.5	0.0190	1 Conv layer	0.8118
128	2	2	Adam	0.5	0.0190	1 Conv layer	0.8058
48	3	3	RMSprop	0.5	0.0190	1 Conv layer	0.8056
48	3	3	RMSprop	0.5	0.0190	1 Conv layer	0.8050
128	2	2	Adam	0.5	0.0190	1 Conv layer	0.8046
48	3	3	RMSprop	0.5	0.0190	1 Conv layer	0.8043
48	3	3	RMSprop	0.5	0.0190	1 Conv layer	0.8032

Figure 11 visually represents the final model architecture. L2 regularisation is a regularisation method which prevents a model from overfitting when trained with highly correlated and high-dimensionality data (Alaeddine and Jihene, 2020; Ghojogh and Crowley, 2019; Humayoo and Cheng, 2019). The model in this study was trained on proteins of the same broad classification (surface proteins) and as such some highly correlated regions were expected in relation to the conserved regions of the proteins (Rudd et al., 2017; Shiliaev et al., 2016). The rectified linear unit (ReLu) activation function was used in the neural network to improve model efficiency by addressing inherent neural network problems such as the vanishing gradient problem (Alaeddine and Jihene, 2020; Lin and Shen, 2018). The binary classification implemented in this study required the use of a sigmoid activation in the output layer (Alaeddine and Jihene, 2020; Korotcov et al., 2017; Thakkar et al., 2018).

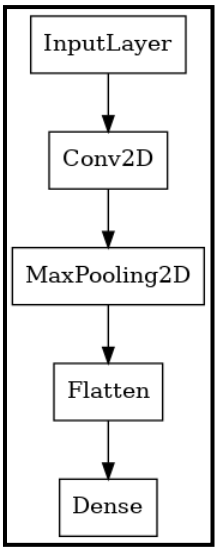

Figure 11: Visual representation of the model architecture. Each box represents a layer in the model architecture. The text in the box is the type of layer.

4.7. Model training

In this study, deep learning was implemented using the *keras* package on a *TensorFlow* backend to allow for distributed operations, using multiple graphics processing units (GPUs) to amplify the model training process. Deep learning models use an iterative learning approach whereby patterns are identified and refined during a subsequent training cycle, and this continues for a predefined number of iterations (Sarawagi and Ganguli, 2021). The model was trained in batches of 64 images for 50 epochs with validation at each epoch. The loss function used for the model was a binary cross-entropy loss which is recommended for binary classification (Ho and Wookey, 2020; Shrestha and Mahmood, 2019).

Hinton et al. (2012) showed that models can adapt to data, resulting in overfitting. As a result, El Korchi and Ghanou (2019) proposed a data centred approach to prevent overfitting, whereby samples of the training data are randomly drawn at the start of each epoch. This can also be referred to as data shuffling and is included as a parameter in the *keras* package used in this study. As such, data shuffling was implemented at the start of each training iteration to allow for rigorous training of the model.

When training a model, a plateau is reached, after which, continuation of the learning process may result in further overfitting. This commonly occurs when the model is allowed to train for too many epochs (Sarawagi and Ganguli, 2021) and in this study, model checkpoints were generated with every epoch to monitor the best model across all epochs. For example, if the model performance was high at

the end of the 12th epoch, the checkpoint would not be updated until an improved model was observed, and if no improvement was observed, at the end of training, the best model would be the one generated on the 12th epoch. The best model in this study, based on the Mathews correlation coefficient (MCC) obtained from model validation, was selected for further performance evaluation, discussed in Section 4.8 below.

4.8. Model evaluation

The performance of a model is commonly measured by its overall accuracy - the ratio of correct classifications to the total number of samples - when tested on previously unseen data. However, accuracy may be prone to bias and is therefore not always considered the best metric (Chicco and Jurman, 2020). In cases of data imbalance, accuracy may appear highly sensitive because a model may simply 'guess' outcomes that favour the majority class and provides the assumption that a model is performing well. As an example, if there are 100 test inputs with 85 from the positive class, and 15 from the negative class, a 'guessing' model which classified all points as positive, would achieve 85% accuracy (Alaeddine and Jihene, 2020; Chicco and Jurman, 2020). Additional metrics such as the F1 score and MCC, which consider the rate of the positive and negative classes with minimal bias, are therefore recommended metrics for evaluation, in addition to accuracy (Chicco et al., 2021; Chicco and Jurman, 2020). These metrics use true positive values, false positive values, true negative values, and false negative values to inform their results (Chicco and Jurman, 2020). As such, F1 score and MCC metrics were used to evaluate the performance of the model in this study. Furthermore, the receiver operating characteristics (ROC) is another metric often used to visually compare several models, where the area under the receiver operating curve (ROC-AUC) is a single value which summarises the ROC metric (Fawcett, 2006; Tharwat, 2021). The ROC-AUC metric was also included in this study to allow comparison with future studies.

Figure 12 depicts the accuracy and loss (error rate), observed for our model during the training phase. Notably, poor performance was observed during the first 12 epochs of training, which may be due to the model first learning the distribution of the data. There was a significant improvement in performance which plateaued in subsequent epochs past the 12th epoch. The accuracy and loss of the training and validation data were comparable from the 13th training epoch indicating consistent performance with no overfitting. The best model was obtained on the 48th epoch which achieved 96.80% accuracy and 0.92 MCC on validation data.

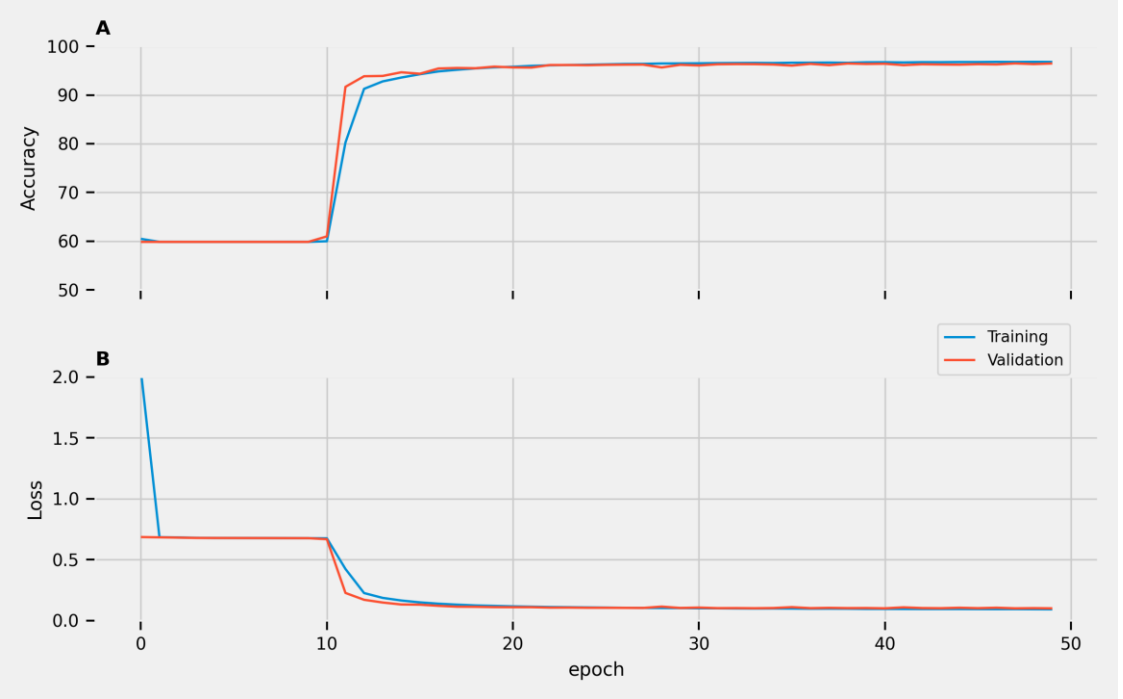

Figure 12: Accuracy and loss of the model during training. A: The accuracy of the model during the training epochs. On the x-axis is the epoch and on the y-axis is the percentage accuracy. B: The loss (error rate) of model during training during the training epochs. On the x-axis is the epoch and on the y-axis is the loss.

A small difference in the accuracy and MCC obtained during training on the training data (99.38%) and on test data (96.78%) was observed, which further indicated model consistency with no overfitting. The high accuracy obtained with the very simplistic single convolution layer model used in this study shows the excellent capability of convolution neural networks, coupled with the FCGR features. Furthermore, we believe the model offers high quality performance because of the significantly large dataset (total of 96630) used in this study, when compared to the quantity of data used in previous studies such as the 10 host receptor protein sequences in Bae and Son (2011), 211 interaction pairs in Yan et al. (2019), and 277 host receptor protein sequences in Cho and Son (2019). Additionally, the training data used in this study consisted of a highly diverse dataset, which included viruses reported to infect plants as well as those reported to infect non-eukaryote organisms. Our study also demonstrates that improving data cleaning methods can significantly help improve the analysis and build better performing models.

Taken together, these results suggest the presence of consistent patterns in surface proteins of viruses reported to infect humans which differ from surface proteins of viruses which do not infect humans. From a biological perspective, this is expected, as host range is determined by successful infection

(Carlson et al., 2019; Wardeh et al., 2020a; Wells et al., 2020), and virus-host cellular protein-protein interactions are a key mechanism (Kerr et al., 2015; Kösesoy et al., 2021; Parvez and Parveen, 2017). The question we ask is, if mutation-based evolution of the viral surface proteins is a key predictor of potential species cross-over events, would a model, such as presented in this study, be a suitable early-stage surveillance tool to monitor such events? Furthermore, we surmise that it is possible that a similar approach can be followed to design a model which predicts epizootic events for hosts other than humans, and particularly for animals of domestic and agricultural importance. In addition, the approach is flexible enough to support multi-category classification, with a simple modification of the final layer in the model architecture, such that a single model could potentially predict cross-species likelihood for several hosts rather than for a single host. Such a model would be a valuable application of machine learning into the One Health initiative, moving the focus from solely humans to other host organisms.

Efforts are continually being made by the deep learning community to identify points of high activation in deep neural networks to inform the use of the patterns identified by the deep learning model (Kindermans et al., 2017; Liang et al., 2021; Linardatos et al., 2020). These approaches include guided backpropagation (Liang et al., 2021), Deep Taylor (Alber et al., 2018), PatternNet (Kindermans et al., 2017), Pattern Attribution (Kindermans et al., 2017) amongst others. These approaches could not, however, be used in our study as they have been designed for multiple classification models (Alber et al., 2018), whereas this study focused on a binary classification model. The ability to understanding the patterns identified by the neural network would be invaluable as they may prompt refined studies in drug discovery and vaccine development, for example.

4.9. Analysis pipeline and optimization

For reproducible research and the adoption of our approach in future studies, a Nextflow pipeline was developed (see Appendix I). Nextflow allows for seamless integration of multiple analysis tools, e.g., Python scripts, R scripts, Bash scripts and others, into a single pipeline (Di Tommaso et al., 2017; Song et al., 2021). The pipeline is designed in such a way that multiple entry points are available, e.g., with the 'EncodeTrain' entry point, feature encoding and model training can be initiated, avoiding the data cleaning steps, and with the 'TestOnly' entry point, an existing model is tested on defined test data. Moreover, if a user has different data pre-processing steps, as well as different encoding steps from the ones used in this study, user defined steps can be integrated into the pipeline. Additionally, custom configuration has been added to support the user's computational environment.

4.10. Proof of concept

When tested on unseen data, the model obtained an accuracy score of 96.78% from a test dataset of 19326 protein sequences. There were 11245 true positive predictions, 259 false positive predictions, 7469 true negative predictions, and 353 false negative predictions. The F1 metric and the MCC were 0.97 and 0.93, respectively, and the ROC-AUC was 0.99. This indicates that our model is robust and reliable, and performed remarkably well on previously unseen data.

Interestingly, five notable phage portal protein (PP) entries from bacterial- and plant- hosts were observed in the false positive predictions, namely, A0A0K2FHA1 (Achromobacter phage phiAxp-2), A7TWJ1 (Staphylococcus virus tp310-2), I7HHN4 (Helicobacter virus KHP30), I7KR94 (Yersinia virus R1RT), and M4QNQ7 (Tetraselmis viridis virus S20). Portal proteins have a low sequence similarity, but highly conserved functionality, playing a role in bi-directional viral DNA passage (Lokareddy et al., 2017). These phage portal proteins are being considered as potential antiviral drug targets in herpes simplex virus infections (Dedeo et al., 2019). The 'plasticity' of phage PP may explain the erroneous classification by the model, due to the presence of signatures consistent with proteins involved in viral entry into human host cells. This observation may indeed be of interest for further investigation, as false positives in this dataset may contain entries which could be considered for therapeutic experimentation as in Dedeo et al. (2019). The other false positives may be as a result of protein similarity; however, this does not eliminate the possibility that some of the false positives may be of future concern, having the capability to bind to human host cells, but still lacking machinery for sustained infection and replication.

A surprising observation in the false negative class was erroneous classification of 31 Human Immunodeficiency Virus (HIV) entries. This was unexpected as HIV is an established, long-term endemic virus with characteristic signatures of viruses with reported human hosts. Investigation of some of the HIV entries such as A0A2P1DQ38, Q7SPP5, and A0A2P1DR91 showed the warning "Lacks conserved residue(s) required for the propagation of feature annotation", according to UniProtKB. Computationally derived feature annotation is reliant on existing knowledge and annotations based on sequence homology, result in errors which are propagated in databases and give rise to contradictory interpretations of the data (Holliday et al., 2017; Zaru et al., 2020). Despite the annotation artifacts, the classification of the HIV entries as false negatives does demonstrate high specificity of the model and may indicate that the FCGR features elucidate important details present, or absent, in protein sequences.

Due to time constraints, we could not thoroughly investigate each erroneous entry, however, we hypothesize other possible explanations which may contribute to incorrect predictions, such as inherent model bias, and labelling errors resulting from pre-processing and imputation. Another possible explanation for incorrect predictions was thought to be small sequence lengths as they may have possibly had poor FCGR representations as discussed in Section 4.5. However, entries with small sequence lengths in the test data were correctly predicted suggesting that sequence length did not contribute to erroneous classifications. This observation warrants further investigation which may explain erroneous classification, so that adjustments can be made in future.

Four entries from A0A1W5YKT3 (Bat coronavirus, Spike glycoprotein), A0A0P0KH07 (Human coronavirus 229E, Spike glycoprotein), Q5EED8 (Human immunodeficiency virus 1, Envelope glycoprotein), and A0A0M4Q8U3 (Influenza D virus, Nucleoprotein) were tested on the model. The Bat coronavirus and Influenza D virus were correctly predicted as non-human infecting viruses with probability scores of 0.0010 and 0.00042, respectively, which are below the threshold of 0.5. The significantly low probability scores may indicate that the proteins do not have signatures associated with sequences from viruses which have been reported to infect humans and may also indicate that these viruses potentially require substantial sequence evolution to permit future species barrier cross-over. The Bat coronavirus is indicated to have 101 hosts in the KW-1160 dataset and the low probability score obtained from the model, coupled with the wide host range of this virus, illustrates the complexity and rareness of zoonotic events, thus possibly supporting the pinhole model (Warren and Sawyer, 2019). However, the selected entry for the POC may be a strain which has not undergone mutations to allow species crossover. This represents one notable limitation of the study which pertains to metadata availability and consistency in the databases. For example, a specific strain may have capacity to infect several hosts, but in a database, may appear to infect fewer host species. This phenomenon may be due to research priority, based on perceived host 'value' (human vs. horse). In this way, even if the virus can infect additional hosts, systemic bias in data representation and data priority exists. It is hoped that with the increased research in One Health, that research priority will become less skewed. It is further recommended that standardised minimal metadata be developed, and database submissions include information such as isolate and strain. This would produce a more comprehensive dataset, and subsequently, a significantly better model capable of strain level prediction.

Compared to other models, such as those developed by Bae and Son (2011), Cho and Son (2019), and Yan et al. (2019), our model performed significantly better. The previous models often focus on virus-

host interactions (Yan et al., 2019) and analysis of host receptor similarity (Bae and Son, 2011; Cho and Son, 2019) which may limit the utility of the latter model, particularly if a virus emerges and uses a different host receptor to those already known. Our model learns patterns present in viral surface proteins such that even if a new virus emerges, targeting an uncommon host receptor, the viral protein patterns will still be detected. To our knowledge, no other study found in our literature searches has used FCGR of virus surface proteins and CNNs to produce a machine learning model with a view to potentially predict viral species cross-over events.

Chapter 5: Conclusion and future research

5.1. Conclusion

Below, key research findings, limitations and future directions pertaining to this study are summarised. The observed increase of epidemic and pandemic events has prompted the need to understand emerging infectious disease outbreaks, with the view to predict and mitigate future incidents. This study used machine learning tools to build a predictive model for zoonosis, by examining the existence of patterns in viral protein sequences linked to entry into host cells. Sequence data and associated metadata was collected from the UniProtKB public database. However, additional processing was required to enrich the data, particularly with respect to observed missing values in the dataset. The study has further demonstrated the inherent data bias which exists across databases, and as such, it is recommended that researchers carefully evaluate data inputs, even when attained from trusted, curated resources. It was also demonstrated in this study that data cleaning and imputation can reduce the inherent bias and improve inputs for the creation of robust models.

This study presents a proof-of-concept approach for building viral zoonosis prediction using FCGR images of protein sequences and convolutional neural networks. To our knowledge, this is the first study which uses FCGR images of viral proteins and CNNs for predictive modelling of species crossover events. Using this approach, we demonstrate the capability of generating extremely robust models with outstanding performance metrics. However, the model has not been compared to other models using common test data and may have biases which have not been observed and documented herein. The model developed in this study suggests the existence of patterns in sequences of virus surface proteins which interact with host cells at the initial stage of infection, and which may be indicative of zoonotic potential. This model could possibly aid in identifying zoonotic viruses, using sequence data extracted from pathogen surveillance programs, as input into the model. Although not explored in this study, the approach has the capability of allowing visualisation of the patterns identified by the model. While this study has promising findings, there are several limitations, which have been noted.

5.2. Future research

Future research would be strengthened through the incorporation of data with clear evidence of zoonosis to generating an improved model. This study developed a binary classification model which limits cross-species prediction to human hosts, and as such, we suggest that future studies include other

host organisms by building a multi-categorical model representing the varying host species, in line with a holistic One Health approach. Furthermore, the studies may be extended to include other pathogenic microorganisms to broaden the zoonotic scope. Several areas of research are investigating FCGR images and pattern analysis for biological inference, and as such, a more refined methodology could strengthen the current model, and increase the reliability and subsequent use of ML technologies in public health related research and pathogen surveillance.

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2016. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. arXiv:1603.04467 [cs].

2. Alaeddine, H., Jihene, M., 2020. A Comparative Study of Popular CNN Topologies Used for Imagenet Classification, in: Deep Neural Networks for Multimodal Imaging and Biomedical Applications. IGI Global, pp. 89–103.

3. Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K.T., Montavon, G., Samek, W., Müller, K., Dähne, S., Kindermans, P., 2018. iNNvestigate neural networks! arXiv:1808.04260 [cs, stat].

4. Alguwaizani, S., Park, B., Zhou, X., Huang, D., Han, K., 2018. Predicting Interactions between Virus and Host Proteins Using Repeat Patterns and Composition of Amino Acids. Journal of Healthcare Engineering 2018. https://doi.org/10.1155/2018/1391265

5. Anthony, S.J., Epstein, J.H., Murray, K.A., Navarrete-Macias, I., Zambrana-Torrelio, C.M., Solovyov, A., Ojeda-Flores, R., Arrigo, N.C., Islam, A., Khan, S.A., Hosseini, P., Bogich, T.L., Olival, K.J., Sanchez-Leon, M.D., Karesh, W.B., Goldstein, T., Luby, S.P., Morse, S.S., Mazet, J.A.K., Daszak, P., Lipkin, W.I., 2013. A Strategy To Estimate Unknown Viral Diversity in Mammals. mBio 4. https://doi.org/10.1128/mBio.00598-13

6. Bae, S., Son, H.S., 2011. Classification of viral zoonosis through receptor pattern analysis. BMC Bioinformatics 12. https://doi.org/10.1186/1471-2105-12-96

7. Baum, S.E., Machalaba, C., Daszak, P., Salerno, R.H., Karesh, W.B., 2017. Evaluating one health: Are we demonstrating effectiveness? One Health 3, 5–10. https://doi.org/10.1016/j.onehlt.2016.10.004

8. Bhalla, S., Sachdeva, S., Batra, S., 2017. Semantic Interoperability in Electronic Health Record Databases: Standards, Architecture and e-Health Systems, in: Reddy, P.K., Sureka, A., Chakravarthy, S., Bhalla, S. (Eds.), Big Data Analytics, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 235–242. https://doi.org/10.1007/978-3-319-72413-3_16

9. Bogich, T.L., Olival, K.J., Hosseini, P.R., Zambrana-Torrelio, C., Loh, E., Funk, S., Brito, I.L., Epstein, J.H., Brownstein, J.S., Joly, D.O., Levy, M.A., Jones, K.E., Morse, S.S., Aguirre, A.A., 2012. Using Mathematical Models In A Unified Approach To Predicting The Next Emerging Infectious Disease, in: Aguirre A.A, Ostfeld R.S, Daszak P. (Eds.), New Directions in Conservation Medicine: Applied Cases of Ecological Health. Oxford University Press, New York, USA, pp 607-618.

10. Brierley, L., Fowler, A., 2021. Predicting the animal hosts of coronaviruses from compositional biases of spike protein and whole genome sequences through machine learning. PLoS Pathogens 17. https://doi.org/10.1371/journal.ppat.1009149

11. Brierley, L., Vonhof, M.J., Olival, K.J., Daszak, P., Jones, K.E., 2016. Quantifying Global Drivers of Zoonotic Bat Viruses: A Process-Based Perspective. The American Naturalist 187, E53–E64. https://doi.org/10.1086/684391

12. Calistri, P., Iannetti, S., Danzetta, M.L., Narcisi, V., Cito, F., Sabatino, D.D., Bruno, R., Sauro, F., Atzeni, M., Carvelli, A., Giovannini, A., 2013. The components of 'One World - One Health' approach. Transboundary and Emerging Diseases 60, 4–13. https://doi.org/10.1111/tbed.12145

13. Cantas, L., Suer, K., 2014. Review: The Important Bacterial Zoonoses in "One Health" Concept. Frontiers in Public Health 2. https://doi.org/10.3389/fpubh.2014.00144

14. Carlson, C.J., Zipfel, C.M., Garnier, R., Bansal, S., 2019. Global estimates of mammalian viral diversity accounting for host sharing. Nature Ecology & Evolution 3, 1070–1075. https://doi.org/10.1038/s41559-019-0910-6

15. Carroll, D., Daszak, P., Wolfe, N.D., Gao, G.F., Morel, C.M., Morzaria, S., Pablos-Méndez, A., Tomori, O., Mazet, J.A.K., 2018. The Global Virome Project. Science 359, 872–874. https://doi.org/10.1126/science.aap7463

16. Cassidy, A., 2018. Humans, Other Animals and "One Health" in the Early Twenty-First Century, in: Animals and the Shaping of Modern Medicine. Springer International Publishing, Cham, pp. 193–236. https://doi.org/10.1007/978-3-319-64337-3_6

17. Centers for Disease Control and Prevention, 2020. CDC - History of Hantavirus Pulmonary Syndrome (HPS) - Hantavirus. URL https://www.cdc.gov/hantavirus/outbreaks/history.html (accessed 11.04.2022).

18. Chen, Q., Zobel, J., Verspoor, K.M., 2017. Duplicates, redundancies and inconsistencies in the primary nucleotide databases: A descriptive study. Database J. Biol. Databases Curation. https://doi.org/10.1093/database/baw163

19. Chicco, D., Jurman, G., 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 21, 6. https://doi.org/10.1186/s12864-019-6413-7

20. Chicco, D., Tötsch, N., Jurman, G., 2021. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Mining 14, 13. https://doi.org/10.1186/s13040-021-00244-z

21. Cho, M., Son, H.S., 2019. Prediction of cross-species infection propensities of viruses with receptor similarity. Infection, Genetics and Evolution 73, 71–80. https://doi.org/10.1016/j.meegid.2019.04.016

22. Chollet, F., others, 2015. Keras. https://github.com/fchollet/keras (accessed 16.03.2022)

23. Chu, X., Ilyas, I.F., Krishnan, S., Wang, J., 2016. Data Cleaning: Overview and Emerging Challenges, in: Proceedings of the 2016 International Conference on Management of Data, SIGMOD '16. Association for Computing Machinery, New York, USA, pp. 2201–2206. https://doi.org/10.1145/2882903.2912574

24. Cuervo-Soto, L.I., López-Pazos, S.A., Batista-García, R.A., 2018. Metagenomics and Diagnosis of Zoonotic Diseases. Farm Animals Diseases, Recent Omic Trends and New Strategies of Treatment. https://doi.org/10.5772/intechopen.72634

25. Cui, G., Fang, C., Han, K., 2012. Prediction of protein-protein interactions between viruses and human by an SVM model. BMC Bioinformatics 13, S5. https://doi.org/10.1186/1471-2105-13-S7-S5

26. Dallas, T.A., Carlson, C.J., Poisot, T., 2019. Testing predictability of disease outbreaks with a simple model of pathogen biogeography. Royal Society Open Science 6, 190883. https://doi.org/10.1098/rsos.190883

27. Daugherty, M.D., Malik, H.S., 2012. Rules of Engagement: Molecular Insights from Host-Virus Arms Races. Annual Review of Genetics 46, 677–700. https://doi.org/10.1146/annurev-genet-110711-155522

28. de Chassey, B., Meyniel-Schicklin, L., Aublin-Gex, A., Navratil, V., Chantier, T., André, P., Lotteau, V., 2013. Structure homology and interaction redundancy for discovering virushost protein interactions. EMBO Reports 14, 938–944. https://doi.org/10.1038/embor.2013.130

29. Dedeo, C.L., Cingolani, G., Teschke, C.M., 2019. Portal Protein: The Orchestrator of Capsid Assembly for the dsDNA Tailed Bacteriophages and Herpesviruses. Annual review of virology 6, 141–160. https://doi.org/10.1146/annurev-virology-092818-015819

30. Deng, L., Nie, W., Zhao, J., Zhang, J., 2021. A hybrid deep learning framework for predicting the protein-protein interaction between virus and host (Preprint). In Review. https://doi.org/10.21203/rs.3.rs-506156/v1

31. Dewancker, I., McCourt, M., Clark, S.C., 2016. Bayesian Optimization for Machine Learning: A Practical Guidebook. ArXiv.

32. Di Tommaso, P., Chatzou, M., Floden, E.W., Barja, P.P., Palumbo, E., Notredame, C., 2017. Nextflow enables reproducible computational workflows. Nature Biotechnology 35, 316–319. https://doi.org/10.1038/nbt.3820

33. Doolittle, J.M., Gomez, S.M., 2011. Mapping Protein Interactions between Dengue Virus and Its Human and Insect Hosts. PLoS Neglected Tropical Diseases 5, e954. https://doi.org/10.1371/journal.pntd.0000954

34. Driscoll, T., Dyer, M.D., Murali, T.M., Sobral, B.W., 2009. PIG–the pathogen interaction gateway. Nucleic Acids Research 37, D647–D650. https://doi.org/10.1093/nar/gkn799

35. Dyer, M.D., Murali, T.M., Sobral, B.W., 2007. Computational prediction of host-pathogen protein-protein interactions. Bioinformatics (Oxford, England) 23, i159–166. https://doi.org/10.1093/bioinformatics/btm208

36. Dyer, M.D., Neff, C., Dufford, M., Rivera, C.G., Shattuck, D., Bassaganya-Riera, J., Murali, T.M., Sobral, B.W., 2010. The Human-Bacterial Pathogen Protein Interaction Networks of Bacillus anthracis, Francisella tularensis, and Yersinia pestis. PLoS ONE 5, e12089. https://doi.org/10.1371/journal.pone.0012089

37. Edgar, T.W., Manz, D.O., 2017. Chapter 6 - Machine Learning, in: Edgar, T.W., Manz, D.O. (Eds.), Research Methods for Cyber Security. Syngress, pp. 153–173. https://doi.org/10.1016/B978-0-12-805349-2.00006-6

38. Eid, F., ElHefnawi, M., Heath, L.S., 2016. DeNovo: Virus-host sequence-based proteinprotein interaction prediction. Bioinformatics 32, 1144–1150. https://doi.org/10.1093/bioinformatics/btv737

39. El Korchi, A., Ghanou, Y., 2019. Unrestricted Random Sampling of data batch to improve the efficiency of neural networks, in: Proceedings of the New Challenges in Data Sciences: Acts of the Second Conference of the Moroccan Classification Society, SMC '19. Association for Computing Machinery, New York, USA, pp. 1–6. https://doi.org/10.1145/3314074.3314102

40. Enard, D., Cai, L., Gwennap, C., Petrov, D.A., 2016. Viruses are a dominant driver of protein adaptation in mammals. eLife 5, e12469. https://doi.org/10.7554/eLife.12469

41. Eng, C.L.P., Tong, J.C., Tan, T.W., 2017. Predicting Zoonotic Risk of Influenza A Viruses from Host Tropism Protein Signature Using Random Forest. International Journal of Molecular Sciences 18. https://doi.org/10.3390/ijms18061135

42. Engering, A., Hogerwerf, L., Slingenbergh, J., 2013. Pathogenhostenvironment interplay and disease emergence. Emerging Microbes & Infections 2, 1–7. https://doi.org/10.1038/emi.2013.5

43. Epstein, J.H., Field, H.E., 2015. Anthropogenic Epidemics: The Ecology of Bat-Borne Viruses and Our Role in their Emergence, in: Wang, L, Cowled, C. (Eds.), Bats and Viruses. John Wiley & Sons, Inc, Hoboken, NJ, pp. 249–279. https://doi.org/10.1002/9781118818824.ch10

44. Evans, P., Dampier, W., Ungar, L., Tozeren, A., 2009. Prediction of HIV-1 virus-host protein interactions using virus and host sequence motifs. BMC Medical Genomics 2, 27. https://doi.org/10.1186/1755-8794-2-27

45. Faburay, B., 2015. The case for a "one health" approach to combating vector-borne diseases. Infection Ecology & Epidemiology. https://doi.org/10.3402/iee.v5.28132

46. Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognition Letters 27, 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

47. Frenay, B., Verleysen, M., 2014. Classification in the Presence of Label Noise: A Survey. IEEE Transactions on Neural Networks and Learning Systems 25, 845–869. https://doi.org/10.1109/TNNLS.2013.2292894

48. French, R.K., Holmes, E.C., 2020. An Ecosystems Perspective on Virus Evolution and Emergence. Trends in Microbiology 28, 165–175. https://doi.org/10.1016/j.tim.2019.10.010

49. Funk, S., Bogich, T.L., Jones, K.E., Kilpatrick, A.M., Daszak, P., 2013. Quantifying Trends in Disease Impact to Produce a Consistent and Reproducible Definition of an Emerging Infectious Disease. PLoS ONE 8, e69951. https://doi.org/10.1371/journal.pone.0069951

50. Ghojogh, B., Crowley, M., 2019. The Theory Behind Overfitting, Cross Validation, Regularization, Bagging, and Boosting: Tutorial. arXiv:1905.12787

51. Glennon, E.E., Jephcott, F.L., Restif, O., Wood, J.L.N., 2019. Estimating undetected Ebola spillovers. PLOS Neglected Tropical Diseases 13. https://doi.org/10.1371/journal.pntd.0007428

52. Goranova, M., Kalcheva-Yovkova, E., Penkov, S., 2015. Task-based Asynchronous Pattern with async and await, in: International Scientific Conference Computer Science, pp. 150-155.

53. Gussow, A.B., Auslander, N., Faure, G., Wolf, Y.I., Zhang, F., Koonin, E.V., 2020. Genomic determinants of pathogenicity in SARS-CoV-2 and other human coronaviruses. Proceedings of the National Academy of Sciences of the United States of America 117, 15193–15199. https://doi.org/10.1073/pnas.2008176117

54. Haider, N., Rothman-Ostrow, P., Osman, A.Y., Arruda, L.B., Macfarlane-Berry, L., Elton, L., Thomason, M.J., Yeboah-Manu, D., Ansumana, R., Kapata, N., Mboera, L., Rushton, J., McHugh, T.D., Heymann, D.L., Zumla, A., Kock, R.A., 2020. COVID-19 or Emerging Infectious Disease? Frontiers in Public Health 8, 763. https://doi.org/10.3389/fpubh.2020.596944

55. Han, B.A., Schmidt, J.P., Bowden, S.E., Drake, J.M., 2015. Rodent reservoirs of future zoonotic diseases. Proceedings of the National Academy of Sciences 112, 7039–7044. https://doi.org/10.1073/pnas.1501598112

56. Hasanin, T., Khoshgoftaar, T.M., Leevy, J.L., Seliya, N., 2019. Examining characteristics of predictive models with imbalanced big data. Journal of Big Data 6, 69. https://doi.org/10.1186/s40537-019-0231-2

57. Hatcher, E.L., Zhdanov, S.A., Bao, Y., Blinkova, O., Nawrocki, E.P., Ostapchuck, Y., Schäffer, A.A., Brister, J.R., 2017. Virus Variation Resource - improved response to emergent viral outbreaks. Nucleic Acids Research 45, D482–D490. https://doi.org/10.1093/nar/gkw1065

58. National Institutes of Health (US), 2007. Understanding Emerging and Re-emerging Infectious Diseases, NIH Curriculum Supplement Series [Internet]. National Institutes of Health (US).

59. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2012. Improving neural networks by preventing co-adaptation of feature detectors. arXiv:1207.0580 [cs.NE]

60. Hitziger, M., Esposito, R., Canali, M., Aragrande, M., Häsler, B., Rüegg, S.R., 2018. Knowledge integration in One Health policy formulation, implementation and evaluation. Bulletin of the World Health Organization 96, 211–218. https://doi.org/10.2471/BLT.17.202705

61. Ho, Y., Wookey, S., 2020. The Real-World-Weight Cross-Entropy Loss Function: Modeling the Costs of Mislabeling. IEEE Access. https://doi.org/10.1109/ACCESS.2019.2962617

62. Holliday, G.L., Davidson, R., Akiva, E., Babbitt, P.C., 2017. Evaluating Functional Annotations of Enzymes Using the Gene Ontology. Methods in molecular biology (Clifton, N.J.) 1446, 111–132. https://doi.org/10.1007/978-1-4939-3743-1_9

63. Horby, P.W., Hoa, N.T., Pfeiffer, D.U., Wertheim, H.F.L., 2014. Drivers of Emerging Zoonotic Infectious Diseases, in: Yamada, A., Kahn, L.H., Kaplan, B., Monath, T.P., Woodall, J., Conti, L. (Eds.), Confronting Emerging Zoonoses: The One Health Paradigm. Springer Japan, Tokyo, pp. 13–26. https://doi.org/10.1007/978-4-431-55120-1_2

64. Huerta-Cepas, J., Serra, F., Bork, P., 2016. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Molecular Biology and Evolution 33, 1635–1638. https://doi.org/10.1093/molbev/msw046

65. Humayoo, M., Cheng, X., 2019. Parameter Estimation with the Ordered $\ell 2$ Regularization via an Alternating Direction Method of Multipliers. Applied Sciences 9, 4291. https://doi.org/10.3390/app9204291

66. Jeffrey, H.J., 1990. Chaos game representation of gene structure. Nucleic Acids Research 18, 2163–2170. https://doi.org/10.1093/nar/18.8.2163

67. Jing, X., Dong, Q., Hong, D., Lu, R., 2020. Amino Acid Encoding Methods for Protein Sequences: A Comprehensive Review and Assessment. IEEE/ACM Transactions on Computational Biology and Bioinformatics 17, 1918–1931. https://doi.org/10.1109/TCBB.2019.2911677

68. Johnson, J.M., Khoshgoftaar, T., 2019. Deep Learning and Data Sampling with Imbalanced Big Data. 2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI). https://doi.org/10.1109/IRI.2019.00038

69. Johnson, J.M., Khoshgoftaar, T.M., 2020. The Effects of Data Sampling with Deep Learning and Highly Imbalanced Big Data. Information Systems Frontiers 22, 1113–1131. https://doi.org/10.1007/s10796-020-10022-7

70. Jones, K.E., Patel, N.G., Levy, M.A., Storeygard, A., Balk, D., Gittleman, J.L., Daszak, P., 2008. Global trends in emerging infectious diseases. Nature 451, 990–993. https://doi.org/10.1038/nature06536

71. Karesh, W.B., Dobson, A., Lloyd-Smith, J.O., Lubroth, J., Dixon, M.A., Bennett, M., Aldrich, S., Harrington, T., Formenty, P., Loh, E.H., Machalaba, C.C., Thomas, M.J., Heymann, D.L., 2012. Ecology of zoonoses: Natural and unnatural histories. The Lancet 380, 1946–1955. https://doi.org/10.1016/S0140-6736(12)61678-X

72. Kerr, S.A., Jackson, E.L., Lungu, O.I., Meyer, A.G., Demogines, A., Ellington, A.D., Georgiou, G., Wilke, C.O., Sawyer, S.L., 2015. Computational and Functional Analysis of the Virus-Receptor Interface Reveals Host Range Trade-Offs in New World Arenaviruses. Journal of Virology 89, 11643–11653. https://doi.org/10.1128/JVI.01408-15

73. Keshari, R., Ghosh, S., Chhabra, S., Vatsa, M., Singh, R., 2020. Unravelling Small Sample Size Problems in the Deep Learning World. IEEE Sixth International Conference on Multimedia Big Data (BigMM), 2020, 134-143, https://doi.org/10.1109/BigMM50055.2020.00028

74. Kilpatrick, A.M., Randolph, S.E., 2012. Drivers, dynamics, and control of emerging vector-borne zoonotic diseases. The Lancet 380, 1946–1955. https://doi.org/10.1016/S0140-6736(12)61151-9

75. Kim, B., Alguwaizani, S., Zhou, X., Huang, D., Park, B., Han, K., 2017. An improved method for predicting interactions between virus and human proteins. Journal of Bioinformatics and Computational Biology 15, 1650024. https://doi.org/10.1142/S0219720016500244

76. Kindermans, P.., Schütt, K.T., Alber, M., Müller, K., Dähne, S., 2017. PatternNet and PatternLRP - Improving the interpretability of neural networks. arXiv:1705.05598 [stat.ML].

77. Klein, A., Falkner, S., Bartels, S., Hennig, P., Hutter, F., 2017. Fast Bayesian hyperparameter optimization on large datasets. Electronic Journal of Statistics 11. https://doi.org/10.1214/17-EJS1335SI

78. Klie, J., Webber, B., Gurevych, I., 2022. Annotation Error Detection: Analyzing the Past and Present for a More Coherent Future.

79. Korotcov, A., Tkachenko, V., Russo, D.P., Ekins, S., 2017. Comparison of Deep Learning With Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Datasets. Molecular pharmaceutics 14, 4462–4475. https://doi.org/10.1021/acs.molpharmaceut.7b00578

80. Kösesoy, İ., Gök, M., Kahveci, T., 2021. Prediction of host-pathogen protein interactions by extended network model. Turkish Journal of Biology 45, 138–148. https://doi.org/10.3906/biy-2009-4

81. Languon, S., Quaye, O., 2019. Filovirus Disease Outbreaks: A Chronological Overview. Virology: Research and Treatment 10. https://doi.org/10.1177/1178122X19849927

82. Lee, K., Brumme, Z.L., 2013. Operationalizing the One Health approach: The global governance challenges. Health Policy and Planning 28, 778–785. https://doi.org/10.1093/heapol/czs127

83. Lemaître, G., Nogueira, F., 2017. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. Journal of Machine Learning Research 18, 559-563.

84. Lerner, H., Berg, C., 2017. A Comparison of Three Holistic Approaches to Health: One Health, EcoHealth, and Planetary Health. Frontiers in Veterinary Science 4, 163. https://doi.org/10.3389/fvets.2017.00163

85. Letko, M., Seifert, S.N., Olival, K.J., Plowright, R.K., Munster, V.J., 2020. Bat-borne virus diversity, spillover and emergence. Nature Reviews Microbiology 18, 461–471. https://doi.org/10.1038/s41579-020-0394-z

86. Li, P., Rao, X., Blase, J., Zhang, Y., Chu, X., Zhang, C., 2021. CleanML: A Study for Evaluating the Impact of Data Cleaning on ML Classification Tasks, in: 2021 IEEE 37th International

Conference on Data Engineering (ICDE). pp. 13–24. https://doi.org/10.1109/ICDE51399.2021.00009

87. Liang, G., Bushman, F.D., 2021. The human virome: Assembly, composition and host interactions. Nature Reviews Microbiology 19, 514–527. https://doi.org/10.1038/s41579-021-00536-5

88. Liang, Y., Li, S., Yan, C., Li, M., Jiang, C., 2021. Explaining the black-box model: A survey of local interpretation methods for deep neural networks. Neurocomputing 419, 168–182. https://doi.org/10.1016/j.neucom.2020.08.011

89. Lin, G., Shen, W., 2018. Research on convolutional neural network based on improved Relu piecewise activation function. Procedia Computer Science, Recent Advancement in Information and Communication Technology: 131, 977–984. https://doi.org/10.1016/j.procs.2018.04.239

90. Linardatos, P., Papastefanopoulos, V., Kotsiantis, S., 2020. Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy 23, 18. https://doi.org/10.3390/e23010018

91. Liu, X., 2016. Chapter 14 - Methods for handling missing data, in: Liu, X. (Ed.), Methods and Applications of Longitudinal Data Analysis. Academic Press, Oxford, pp. 441–473. https://doi.org/10.1016/B978-0-12-801342-7.00014-9

92. Löchel, H.F., Eger, D., Sperlea, T., Heider, D., 2020. Deep Learning on Chaos Game Representation for Proteins. Bioinformatics, 36, 272-279, https://doi.org/10.1093/bioinformatics/btz493

93. Loh, E.H., Murray, K.A., Zambrana-Torrelio, C., Hosseini, P.R., Rostal, M.K., Karesh, W.B., Daszak, P., 2013. Ecological Approaches to Studying Zoonoses. Microbiology Spectrum 1. https://doi.org/10.1128/microbiolspec.OH-0009-2012

94. Lokareddy, R.K., Sankhala, R.S., Roy, A., Afonine, P.V., Motwani, T., Teschke, C.M., Parent, K.N., Cingolani, G., 2017. Portal protein functions akin to a DNA-sensor that couples genome-packaging to icosahedral capsid maturation. Nature Communications 8, 14310. c10.1038/ncomms14310

95. Long, J.S., Mistry, B., Haslam, S.M., Barclay, W.S., 2019. Host and viral determinants of influenza A virus species specificity. Nature Reviews Microbiology 17, 67–81. https://doi.org/10.1038/s41579-018-0115-z

96. Lopez, M.J., Mohiuddin, S.S., 2022. Biochemistry, Essential Amino Acids, in: StatPearls. StatPearls Publishing, Treasure Island (FL).

97. Lu, Z., Chiang, C., Sha, F., 2019. Hyper-parameter Tuning under a Budget Constraint, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, 5744-5750. https://doi.org/10.24963/ijcai.2019/796

98. Madhav, N., Oppenheim, B., Gallivan, M., Mulembakani, P., Rubin, E., Wolfe, N., 2017. Pandemics: Risks, Impacts, and Mitigation, in: Jamison, D.T., Gelband, H., Horton, S., Jha, P., Laxminarayan, R., Mock, C.N., Nugent, R. (Eds.), Disease Control Priorities: Improving Health

and Reducing Poverty. The International Bank for Reconstruction and Development / The World Bank, Washington (DC).

99. Mariotti, M., Salinas, G., Gabaldón, T., Gladyshev, V., 2018. Use of selenocysteine, the 21st amino acid, in the fungal kingdom. bioRxiv. https://doi.org/10.1101/314781

100. McBee, R.M., Rozmiarek, S.A., Meyerson, N.R., Rowley, P.A., Sawyer, S.L., 2015. The Effect of Species Representation on the Detection of Positive Selection in Primate Gene Data Sets. Molecular Biology and Evolution 32, 1091–1096. https://doi.org/10.1093/molbev/msu399

101. Mei, J., Zhao, J., 2018. Prediction of HIV-1 and HIV-2 proteins by using Chou's pseudo amino acid compositions and different classifiers. Scientific Reports 8. https://doi.org/10.1038/s41598-018-20819-x

102. Mihara, T., Nishimura, Y., Shimizu, Y., Nishiyama, H., Yoshikawa, G., Uehara, H., Hingamp, P., Goto, S., Ogata, H., 2016. Linking Virus Genomes with Host Taxonomy. Viruses 8, 66. https://doi.org/10.3390/v8030066

103. Modarresi, K., Munir, A., 2018. Standardization of Featureless Variables for Machine Learning Models Using Natural Language Processing, in: Shi, Y., Fu, H., Tian, Y., Krzhizhanovskaya, V.V., Lees, M.H., Dongarra, J., Sloot, P.M.A. (Eds.), Computational Science ICCS 2018. Springer International Publishing, Cham, pp. 234–246. https://doi.org/10.1007/978-3-319-93701-4_18

104. Morse, S.S., Mazet, J.A.K., Woolhouse, M., Parrish, C.R., Carroll, D., Karesh, W.B., Zambrana-Torrelio, C., Lipkin, W.I., Daszak, P., 2012. Prediction and prevention of the next pandemic zoonosis. Lancet 380, 1956–1965. https://doi.org/10.1016/S0140-6736(12)61684-5

105. Mu, Z., Yu, T., Qi, E., Liu, J., Li, G., 2019. DCGR: Feature extractions from protein sequences based on CGR via remodeling multiple information. BMC Bioinformatics 20, 351. https://doi.org/10.1186/s12859-019-2943-x

106. Nguyen, L.H., Holmes, S., 2019. Ten quick tips for effective dimensionality reduction. PLOS Computational Biology 15, e1006907. https://doi.org/10.1371/journal.pcbi.1006907

107. O'Malley, T., Bursztein, E., Long, J., Chollet, F., Jin, H., Invernizzi, L., others, 2019. KerasTuner. https://github.com/keras-team/keras-tuner

108. Olival, K.J., Hosseini, P.R., Zambrana-Torrelio, C., Ross, N., Bogich, T.L., Daszak, P., 2017. Host and viral traits predict zoonotic spillover from mammals. Nature 546, 646–650. https://doi.org/10.1038/nature22975

109. Parrish, C.R., Holmes, E.C., Morens, D.M., Park, E., Burke, D.S., Calisher, C.H., Laughlin, C.A., Saif, L.J., Daszak, P., 2008. Cross-Species Virus Transmission and the Emergence of New Epidemic Diseases. Microbiology and Molecular Biology Reviews : MMBR 72, 457–470. https://doi.org/10.1128/MMBR.00004-08

110. Parvez, M.K., Parveen, S., 2017. Evolution and emergence of pathogenic viruses: Past, present, and future. Intervirology 60, 1–7. https://doi.org/10.1159%2F000478729

111. Pettan-Brewer, C., Pettan-Brewer, C., Martins, A.F., Abreu, D.P.B. de, Brandão, A.P.D., Barbosa, D.S., Figueroa, D.P., Cediel, N., Kahn, L.H., Brandespim, D.F., Velásquez, J.C.C., Carvalho, A.A.B., Takayanagui, A.M.M., Galhardo, J.A., Maia-Filho, L.F.A., Pimpão, C.T., Vicente, C.R., Biondo, A.W., Biondo, A.W., 2021. From the Approach to the Concept: One Health in Latin America-Experiences and Perspectives in Brazil, Chile, and Colombia. Frontiers in Public Health 9. https://doi.org/10.3389/fpubh.2021.687110

112. Pietrzyk, A.J., Bujacz, A., Jaskolski, M., Bujacz, G., 2013. Identification of amino acid sequences via X-ray crystallography: A mini review of case studies. BioTechnologia 1, 9–14. https://doi.org/10.5114/bta.2013.46427

113. Piquero, A.R., Carmichael, S., 2005. Attrition, Mortality, and Exposure Time, in: Kempf-Leonard, K. (Ed.), Encyclopedia of Social Measurement. Elsevier, New York, USA, pp. 97–101. https://doi.org/10.1016/B0-12-369398-5/00048-7

114. Piret, J., Boivin, G., 2021. Pandemics Throughout History. Frontiers in Microbiology 11. https://doi.org/10.3389/fmicb.2020.631

115. Plowright, R.K., Eby, P., Hudson, P.J., Smith, I.L., Westcott, D., Bryden, W.L., Middleton, D., Reid, P.A., McFarlane, R.A., Martin, G., Tabor, G.M., Skerratt, L.F., Anderson, D.L., Crameri, G., Quammen, D., Jordan, D., Freeman, P., Wang, L.., Epstein, J.H., Marsh, G.A., Kung, N.Y., McCallum, H., 2015. Ecological dynamics of emerging bat virus spillover. Proceedings of the Royal Society B: Biological Sciences 282, 20142124. https://doi.org/10.1098/rspb.2014.2124

116. Qi, Y., Tastan, O., Carbonell, J.G., Klein-Seetharaman, J., Weston, J., 2010. Semi-supervised multi-task learning for predicting interactions between HIV-1 and human proteins. Bioinformatics 26, i645–i652. https://doi.org/10.1093/bioinformatics/btq394

117. Qiang, X., Kou, Z., 2019. Predicting interspecies transmission of avian influenza virus based on wavelet packet decomposition. Computational Biology and Chemistry 78, 455–459. https://doi.org/10.1016/j.compbiolchem.2018.11.029

118. Qiang, X., Xu, P., Fang, G., Liu, W., Kou, Z., 2020. Using the spike protein feature to predict infection risk and monitor the evolutionary dynamic of coronavirus. Infectious Diseases of Poverty 9. https://doi.org/10.1186/s40249-020-00649-8

119. Ray, E.L., Reich, N.G., 2018. Prediction of infectious disease epidemics via weighted density ensembles. PLOS Computational Biology 14, e1005910. https://doi.org/10.1371/journal.pcbi.1005910

120. Reid, A.H., Fanning, T.G., Janczewski, T.A., Lourens, R.M., Taubenberger, J.K., 2004. Novel Origin of the 1918 Pandemic Influenza Virus Nucleoprotein Gene. Journal of Virology 78, 12462–12470. https://doi.org/10.1128/JVI.78.22.12462-12470.2004

121. Roberts, M., Dobson, A., Restif, O., Wells, K., 2021. Challenges in modelling the dynamics of infectious diseases at the wildlifehuman interface. Epidemics 37, 100523. https://doi.org/10.1016/j.epidem.2021.100523

122. Roger, F., Caron, A., Morand, S., Pedrono, M., Garine-Wichatitsky, M. de, Chevalier, V., Tran, A., Gaidet, N., Figuié, M., de Visscher, M., Binot, A., 2016. One Health and EcoHealth: The same wine in different bottles? Infection Ecology & Epidemiology 6, 30978. https://doi.org/10.3402/iee.v6.30978

123. Rothan, H.A., Byrareddy, S.N., 2020. The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak. Journal of Autoimmunity 109, 102433. https://doi.org/10.1016/j.jaut.2020.102433

124. Rother, M., Krzycki, J., 2010. Selenocysteine, Pyrrolysine, and the Unique Energy Metabolism of Methanogenic Archaea. Archaea. https://doi.org/10.1155/2010/453642

125. Royce, K., Fu, F., 2020. Mathematically modeling spillovers of an emerging infectious zoonosis with an intermediate host. PLOS ONE 15, e0237780. https://doi.org/10.1371/journal.pone.0237780

126. Rudd, T.R., Preston, M.D., Yates, E.A., 2017. The nature of the conserved basic amino acid sequences found among 437 heparin binding proteins determined by network analysis. Molecular BioSystems 13, 852–865. https://doi.org/10.1039/C6MB00857G

127. Rüegg, S.R., McMahon, B.J., Häsler, B., Esposito, R., Nielsen, L.R., Ifejika Speranza, C., Ehlinger, T., Peyre, M., Aragrande, M., Zinsstag, J., Davies, P., Mihalca, A.D., Buttigieg, S.C., Rushton, J., Carmo, L.P., De Meneghi, D., Canali, M., Filippitzi, M.E., Goutard, F.L., Ilieski, V., Milićević, D., O'Shea, H., Radeski, M., Kock, R., Staines, A., Lindberg, A., 2017. A Blueprint to Evaluate One Health. Frontiers in Public Health 5, 20. https://doi.org/10.3389/fpubh.2017.00020

128. Salyer, S.J., Silver, R., Simone, K., Barton Behravesh, C., 2017. Prioritizing Zoonoses for Global Health Capacity Building from One Health Zoonotic Disease Workshops in 7 Countries, 2014. Emerging Infectious Diseases 23. https://doi.org/10.3201/eid2313.170418

129. Sandle, T., 2016. 9 - Microbial identification, in: Sandle, T. (Ed.), Pharmaceutical Microbiology. Woodhead Publishing, Oxford, pp. 103–113. https://doi.org/10.1016/B978-0-08-100022-9.00009-8

130. Sanjuán, R., Domingo-Calap, P., 2016. Mechanisms of viral mutation. Cellular and Molecular Life Sciences 73, 4433–4448. https://doi.org/10.1007/s00018-016-2299-6

131. Sarawagi, S., Ganguli, R., 2021. Deep Neural Network Surrogates for Optimal Design of Helicopter Rotor. Transactions of the Indian National Academy of Engineering 6, 653–664. https://doi.org/10.1007/s41403-021-00227-w

132. Schilling, F., 2016. The Effect of Batch Normalization on Deep Convolutional Neural Networks. Masters Thesis, KTH Royal Institute of Technology, Stockholm.

133. Schoch, C.L., Ciufo, S., Domrachev, M., Hotton, C.L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., Sharma, S., Soussov, V., Sullivan, J.P., Sun, L., Turner, S., Karsch-Mizrachi, I., 2020. NCBI Taxonomy: A comprehensive update on curation, resources and tools. Database: The Journal of Biological Databases and Curation 2020, baaa062. https://doi.org/10.1093/database/baaa062

134. Searle, B.C., Dasari, S., Turner, M., Reddy, A.P., Choi, D., Wilmarth, P.A., McCormack, A.L., David, L.L., Nagalla, S.R., 2004. High-Throughput Identification of Proteins and Unanticipated Sequence Modifications Using a Mass-Based Alignment Algorithm for MS/MS de Novo Sequencing Results. Analytical Chemistry 76, 2220–2230. https://doi.org/10.1021/ac035258x

135. Selman, M., Dankar, S.K., Forbes, N.E., Jia, J., Brown, E.G., 2012. Adaptive mutation in influenza A virus non-structural gene is linked to host switching and induces a novel protein by alternative splicing. Emerging Microbes & Infections 1, 1-10. https://doi.org/10.1038/emi.2012.38

136. Shiliaev, N.G., Selivanova, O.M., Galzitskaya, O.V., 2016. Search for conserved amino acid residues of the α-crystallin proteins of vertebrates. Journal of Bioinformatics and Computational Biology 14, 1641004. https://doi.org/10.1142/S0219720016410043

137. Shrestha, A., Mahmood, A., 2019. Review of Deep Learning Algorithms and Architectures. IEEE Access. https://doi.org/10.1109/ACCESS.2019.2912200

138. Siegel, R.D., 2018. Classification of Human Viruses. Principles and Practice of Pediatric Infectious Diseases 1044–1048.e1. https://doi.org/10.1016/B978-0-323-40181-4.00201-2

139. Smith, I., Wang, L., 2013. Bats and their virome: An important source of emerging viruses capable of infecting humans. Current Opinion in Virology 3, 84–91. https://doi.org/10.1016/j.coviro.2012.11.006

140. Smith, K.F., Goldberg, M., Rosenthal, S., Carlson, L., Chen, J., Chen, C., Ramachandran, S., 2014. Global rise in human infectious disease outbreaks. Journal of the Royal Society, Interface 11, 20140950. https://doi.org/10.1098/rsif.2014.0950

141. Song, Z., Gurinovich, A., Federico, A., Monti, S., Sebastiani, P., 2021. Nf-gwas-pipeline: A Nextflow Genome-Wide Association Study Pipeline. J. Open Source Softw. https://doi.org/10.21105/joss.02957

142. Soyemi, J., Isewon, I., Oyelade, J., Adebiyi, E., 2018. Inter-Species/Host-Parasite Protein Interaction Predictions Reviewed. Current Bioinformatics 13, 396–406. https://doi.org/10.2174/1574893613666180108155851

143. Steward, K., 2019. Essential Amino Acids: Chart, Abbreviations and Structure. Applied Sciences from Technology Networks. http://www.technologynetworks.com/applied-sciences/articles/essential-amino-acids-chart-abbreviations-and-structure-324357

144. Taylor, L.H., Latham, S.M., Woolhouse, M.E., 2001. Risk factors for human disease emergence. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 356, 983–989. https://doi.org/10.1098/rstb.2001.0888

145. Taylor, T.J., Vaisman, I.I., 2010. Discrimination of thermophilic and mesophilic proteins. BMC Structural Biology 10, S5. https://doi.org/10.1186/1472-6807-10-S1-S5

146. Tensorflow Developers, 2022. TensorFlow. https://doi.org/10.5281/zenodo.5949169

147. Temmam, S., Davoust, B., Berenger, J., Raoult, D., Desnues, C., 2014. Viral Metagenomics on Animals as a Tool for the Detection of Zoonoses Prior to Human Infection? International Journal of Molecular Sciences 15, 10377–10397. https://doi.org/10.3390/ijms150610377

148. Thakkar, V., Tewary, S., Chakraborty, C., 2018. Batch Normalization in Convolutional Neural Networks: A comparative study with CIFAR-10 data. 2018 Fifth International Conference on Emerging Applications of Information Technology (EAIT). https://doi.org/10.1109/EAIT.2018.8470438

149. Tharwat, A., 2021. Classification assessment methods. Applied Computing and Informatics 17, 168–192. https://doi.org/10.1016/j.aci.2018.08.003

150. The UniProt Consortium, 2021. UniProt: The universal protein knowledgebase in 2021. Nucleic Acids Research 49, D480–D489. https://doi.org/10.1093/nar/gkaa1100

151. Thessen, A., 2016. Adoption of Machine Learning Techniques in Ecology and Earth Science. One Ecosystem 1, e8621. https://doi.org/10.3897/oneeco.1.e8621

152. Valueva, M.V., Nagornov, N.N., Lyakhov, P.A., Valuev, G.V., Chervyakov, N.I., 2020. Application of the residue number system to reduce hardware costs of the convolutional neural network implementation. Mathematics and Computers in Simulation 177, 232–243. https://doi.org/10.1016/j.matcom.2020.04.031

153. Vargas, R., Mosavi, A., Ruiz, R., 2018. Deep Learning: A Review. Preprints 2018. https://doi.org/10.20944/PREPRINTS201810.0218.V1

154. Vyatkina, K., Wu, S., Dekker, L.J.M., VanDuijn, M.M., Liu, X., Tolić, N., Dvorkin, M., Alexandrova, S., Luider, T.M., Paša-Tolić, L., Pevzner, P.A., 2015. De Novo Sequencing of Peptides from Top-Down Tandem Mass Spectra. Journal of Proteome Research 14, 4450–4462. https://doi.org/10.1021/pr501244v

155. Wardeh, M., Blagrove, M.S.C., Sharkey, K.J., Baylis, M., 2020a. Divide and conquer machine-learning integrates mammalian, viral, and network traits to predict unknown virus-mammal associations. bioRxiv 2020.06.13.150003. https://doi.org/10.1101/2020.06.13.150003

156. Wardeh, M., Risley, C., McIntyre, M.K., Setzkorn, C., Baylis, M., 2015. Database of host-pathogen and related species interactions, and their global distribution. Scientific Data 2, 150049. https://doi.org/10.1038/sdata.2015.49

157. Wardeh, M., Sharkey, K.J., Baylis, M., 2020b. Integration of shared-pathogen networks and machine learning reveals the key aspects of zoonoses and predicts mammalian reservoirs. Proceedings of the Royal Society B: Biological Sciences 287, 20192882. https://doi.org/10.1098/rspb.2019.2882

158. Warren, C.J., Sawyer, S.L., 2019. How host genetics dictates successful viral zoonosis. PLOS Biology 17, e3000217. https://doi.org/10.1371/journal.pbio.3000217

159. Wells, K., Morand, S., Wardeh, M., Baylis, M., 2020. Data from: Distinct spread of DNA and RNA viruses among mammals amid prominent role of domestic species. https://doi.org/10.5061/DRYAD.P2NGF1VMG

160. Wolfe, N.D., Dunavan, C.P., Diamond, J., 2007. Origins of major human infectious diseases. Nature 447, 279–283. https://doi.org/10.1038/nature05775

161. Woolf, B.P., 2009. Machine Learning, in: Building Intelligent Interactive Tutors. Elsevier, pp. 221–297. https://doi.org/10.1016/B978-0-12-373594-2.00007-1

162. Woolhouse, M.E., Adair, K., Brierley, L., 2013. RNA Viruses: A Case Study of the Biology of Emerging Infectious Diseases. Microbiology Spectrum 1. https://doi.org/10.1128/microbiolspec.OH-0001-2012

163. Woolhouse, M.E.J., Gowtage-Sequeria, S., 2005. Host Range and Emerging and Reemerging Pathogens. Emerging Infectious Diseases 11, 1842–1847. https://doi.org/10.3201/eid1112.050997

164. Woolhouse, M.E.J., Howey, R., Gaunt, E., Reilly, L., Chase-Topping, M., Savill, N., 2008. Temporal trends in the discovery of human viruses. Proceedings of the Royal Society B: Biological Sciences 275, 2111–2115. https://doi.org/10.1098/rspb.2008.0294

165. Woolhouse, M., Gaunt, E., 2007. Ecological Origins of Novel Human Pathogens. Critical Reviews in Microbiology 33, 231–242. https://doi.org/10.1080/10408410701647560

166. Woolhouse, M., Scott, F., Hudson, Z., Howey, R., Chase-Topping, M., 2012. Human viruses: Discovery and emergence. Philosophical Transactions of the Royal Society B: Biological Sciences 367, 2864–2871. https://doi.org/10.1098/rstb.2011.0354

167. Wu, Y., Yang, F., Liu, Y., Zha, X., Yuan, S., 2018. A Comparison of 1-D and 2-D Deep Convolutional Neural Networks in ECG Classification. Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference. https://doi.org/10.1109/EMBC.2018.8512242

168. Wuchty, S., 2011. Computational Prediction of Host-Parasite Protein Interactions between P. Falciparum and H. sapiens. PLoS ONE 6, e26960. https://doi.org/10.1371/journal.pone.0026960

169. Yan, C., Duan, G., Wu, F., Wang, J., 2019. IILLS: Predicting virus-receptor interactions based on similarity and semi-supervised learning. BMC Bioinformatics 20. https://doi.org/10.1186/s12859-019-3278-3

170. Zaru, R., Magrane, M., Orchard, S., The UniProt Consortium., 2020. Challenges in the annotation of pseudoenzymes in databases: The UniProtKB approach. The FEBS Journal 287, 4114–4127. https://doi.org/10.1111/febs.15100

171. Zhang, A., He, L., Wang, Y., 2017. Prediction of GCRV virus-host protein interactome based on structural motif-domain interactions. BMC Bioinformatics 18, 145. https://doi.org/10.1186/s12859-017-1500-8

172. Zhao, P., 2016. R with Parallel Computing from User Perspectives. R-bloggers. https://www.r-bloggers.com/2016/09/r-with-parallel-computing-from-user-perspectives

173. Zheng, L., Li, C., Ping, J., Zhou, Y., Li, Y., Hao, P., 2014. The domain landscape of virus-host interactomes. BioMed Research International 2014, 867235. https://doi.org/10.1155/2014/867235

174. Zinsstag, J., Mackenzie, J.S., Jeggo, M., Heymann, D.L., Patz, J.A., Daszak, P., 2012. Mainstreaming One Health. EcoHealth 9, 107–110. https://doi.org/10.1007/s10393-012-0772-8

175. Zinsstag, J., Schelling, E., Waltner-Toews, D., Tanner, M., 2011. From "one medicine" to "one health" and systemic approaches to health and well-being. Preventive Veterinary Medicine 101, 148–156. https://doi.org/10.1016/j.prevetmed.2010.07.003

Appendices

Appendix I: Nextflow pipeline

Description:

The pipeline is complimentary to this thesis. It serves as a template to reproduce the method used in the thesis. It consists of binary and configuration directories named *bin* and *conf*, respectively. The *bin* directory consists of executable scripts used in the pipeline and the *conf* directory consists of configuration files for running the pipeline on different computational platforms. A global configuration file named *nextflow.config* is also included which, however, is not placed in the *conf* directory. The *nextflow.config* file also contains pipeline parameters. Pipeline parameters can also be provided using the *params.yml* file provided. Two directories, namely *modules* and *workflows*, contain the Nextflow scripts used to execute the various pipeline processes. Finally, a *main* Nextflow script is included which serves as the entry point for the pipeline thereby abstracting users from the multiple scripts. The pipeline is also available on github where its usage is described.

File:

- Zoon0PredV

GitHub Link:

- https://www.github.com/Rudolph-afk/Zoon0PredV

Appendix II: Containers definition and software dependencies

Singularity images were used for reproducing the environment used in this study. Two singularity containers were used and the software dependencies are listed in the definition files.

FCGR Singularity image definition

Description:

Singularity image definition for the container used for frequency chaos game representation (FCGR) (see Section 4.5)

Files:

- FCGR_container.def

Data processing and model processing Singularity image definition

Description:

Singularity image definition file for the container used for data cleaning, imputation, model training and evaluation (see Chapter 3:).

Files:

- tensorflow_container.def

Appendix III: Downloaded data

The data downloaded from various databases used in the study (see Section 3.2).

UniProtKB

Description:

Data downloaded from UniProtKB as FASTA and accompanying tabular metadata.

File:

- KW-1160.fasta
- KW-1160.tab.gz

NCBI Virus

Description:

Data downloaded from NCBI Virus as a comma-separated-value tabular file.

File:

- NCBIVirus.csv

EID2

Description:

Data obtained from EID2 as a comma-separated-value tabular file.

File:

- SpeciesInteractions_EID2.csv

Virus-Host DB

Description:

Data obtained from Virus-Host DB as a tab-separated-value tabular file.

File:

- virushostdb.tsv

Taxonomy database

Description:

Taxonomic database downloaded by the *ete3 toolkit* package.

File:

- taxdump.tar.gz

Appendix IV: Python and R scripts used in the pipeline

FCGR

Description:

The script containing the code used to implement FCGR. The dependencies are all listed in the FCGR Singularity image definition **(**see Appendix II:**)**.

File:

- chaos_game_representation_of_protein_sequences.R

Data cleaning and imputation

Description:

The script containing code used to clean the KW-1160 data obtained from UniProt. The script also includes code written to execute the imputation passes.

File:

- data_cleaning.py

Hyperparameter search

Description:

Contains the code executed for hyperparameter search algorithm used in the thesis. The script takes no parameters and returns a CSV file with results of the hyperparameter search.

File:

- hyperparameter_search.py

Evaluation metrics utility functions

Description:

A script containing utility functions used to define the metrics used when evaluating the model in this study.

File:

- metrics_helper.py

Model architecture

Description:

A script containing the model architecture as described in Section 4.6. Additionally, it contains a utility function to load a trained model.

File:

- model_definition.py

Model testing

Description:

Script used to evaluate the performance of the model created in the study.

File:

- test_zoonosis_model.py

Model training and validation

Description:

The script used to train the model defined in this study.

File:

- train_zoonosis_model.py

Data cleaning and imputation utility functions

Description:

A script containing utility functions used for data cleaning and imputation.

File:

- zoonosis_helper_functions.py

Appendix V: Model hyperparameter search results

Description:

Hyperparameter tuning was executed for 500 trials. This appendix shows results for all 500 trails executed.

Files:

- hyperparameter_results.csv

Appendix VI: Generated model

Description:

The model is complimentary to the thesis as it is the overall output. The model is saved as multiple files by the *keras* package and using it required installation of *TensorFlow*, version 2.6.0 and above, which is packaged with the latest version of the *keras* package.

Files:

- model

Appendix VII: FCGR Benchmark Results

Description:

Frequency chaos game representation benchmark results. Elapsed refers to the time taken to execute the function. Type refers to the name given to the type of execution used. Tests refers to the number benchmarking test which were executed.

Type	Tests	Elapsed
Asynchronous FCGR	100	1,326.52
Parallel + Asynchronous FCGR	100	148.19
Parallel FCGR: foreach	100	142.15
Parallel FCGR: parallel	100	207.59
Synchronous FCGR	100	1,228.46