Axel Christian Ørn Luiggi-Gørrissen
Trygve Johansen Woldseth

Beyond Traditional Boundaries:

Exploiting Open Dimensions for Enhancing Fake News Detection with Artificial Immune Systems

Master's thesis in Computer Science
Supervisor: Pauline Catriona Haddow
June 2023

Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

NTNU
Norwegian University of
Science and Technology

Axel Christian Ørn Luiggi-Gørrissen
Trygve Johansen Woldseth

Beyond Traditional Boundaries:

Exploiting Open Dimensions for Enhancing Fake News Detection with Artificial Immune Systems

Master's thesis in Computer Science
Supervisor: Pauline Catriona Haddow
June 2023

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

Beyond Traditional Boundaries:
Exploiting Open Dimensions for Enhancing Fake News Detection with Artificial Immune Systems

Axel Luiggi-Gørrissen, Trygve Johansen Woldseth

Master's thesis, Spring 2023

Artificial Intelligence Group
Department of Computer and Information Science
Faculty of Information Technology, Mathematics and Electrical Engineering

NTNU
Kunnskap for en bedre verden

Abstract

In the age of digital revolution, the rapid dissemination of information, particularly fake news, has posed significant challenges. Nearly half of U.S. adults turned to social media for news in 2021, creating an efficient system for the proliferation of fake news, which spreads more quickly and extensively than factual news. Especially in politically charged contexts, fake news has far-reaching effects, leading to potential manipulation and emotional distress among the general population. With such pervasive impacts, there is a need for computational models capable of swiftly identifying non-factual.

An Artificial Immune System (AIS) model is proposed as an intuitive solution to fake news detection. Drawing inspiration from the biological immune system's ability to distinguish self from non-self, the model treats a news piece as an antigen, thereby distinguishing fake news (non-self) from real news (self).

Novel adaptations within the AIS framework are investigated, addressing three research questions: adapting AIS to high-dimensional spaces, comparing embeddings and general textual features for fake news detection in an AIS model, and evaluating the proposed model against state-of-the-art methods. A novel dimension type for recognition regions, termed ”open dimensions”, is proposed, capable of managing high-dimensional spaces effectively and outperforming other dimension types on multiple benchmark classification sets. The model employing embeddings outperformed the model employing general textual features; however feature analysis was not performed for the model with general textual features, and effective feature selection may enhance the model's accuracy. Although the proposed model does not match neural network-based approaches' accuracy in the task of fake news detection, it exhibits competitive results with improved training time.

Although the research identified potential limitations such as a need for more in-depth exploration of open dimensions dynamics and further research into boosting methods and local search operators, a valuable foundation for future work in AIS and fake news detection is established. AIS models demonstrate potential in this field, with findings contributing towards the goal of combating the proliferation of fake news in our increasingly digital society.

Sammendrag

I den digitale revolusjonens tidsalder har den raske spredningen av informasjon, spesielt falske nyheter, skapt betydelige utfordringer. Nesten halvparten av voksne i USA brukte sosiale medier som kilde for nyheter i 2021, noe som har skapt et effektivt system for spredning av falske nyheter, som sprer seg raskere og mer omfattende enn ekte nyheter. Spesielt i politisk ladede sammenhenger har falske nyheter vidtrekkende effekter, noe som potensielt kan føre til manipulasjon av- og fysiske effekter blant befolkningen. Med slike langtrekkende effekter er det et behov for modeller som raskt kan identifisere falske nyheter.

En modell basert på kunstige immunsystemer (AIS) blir foreslått som en intuitiv løsning for påvisning av falske nyheter. Med inspirasjon fra det biologiske immunsystemets evne til å skille selv fra ikke-selv, behandler modellen en nyhetsartikkel som et antigen, og skiller dermed falske nyheter (ikke-selv) fra virkelige nyheter (selv).

Nyskapende elementer innenfor AIS blir undersøkt, og ser på tre forskningsspørsmål: tilpasning av AIS til høydimensjonale rom, sammenligning av *embeddings* og generelle tekstegenskaper for deteksjon av falske nyheter i en AIS-modell, og evaluering av den foreslåtte modellen mot eksisterende metoder. En ny dimensjonstype for *recognition regions*, kalt ”åpne dimensjoner”, blir foreslått, i stand til å håndtere høydimensjonale rom effektivt og yte bedre enn andre dimensjonstyper på flere velkjente klassifiseringssett. Modellen som bruker *embeddings* presterte bedre enn modellen som bruker generelle tekstegenskaper; men det ble ikke gjennomført en analyse av elementene for modellen som brukte generelle tekstegenskaper, og effektiv analyse av dette kan forbedre modellens nøyaktighet. Selv om den foreslåtte modellen ikke når nøyaktigheten til modeller basert på nevrale nettverk for deteksjon falske nyheter, viser den konkurransedyktige resultater med forbedret treningstid.

Selv om forskningen identifiserte potensielle begrensninger, som et behov for mer dyptgående utforskning av åpne dimensjoners dynamikk og ytterligere forskning på *boosting*-metode og lokale søkeoperatører, er det lagt et verdifullt grunnlag for fremtidig arbeid i AIS og deteksjon av falske nyheter. AIS-modeller demonstrerer potensial i dette feltet, med funn som bidrar til målet om å bekjempe spredningen av falske nyheter i vårt stadig mer digitale samfunn.

Preface

The following thesis was completed as the final part of our master's degree in Computer Science at the Norwegian University of Science and Technology.

First and foremost, we want to thank our supervisor Pauline Catriona Haddow for her invaluable guidance, support and encouragement throughout the project. Her unwavering optimism, deep understanding of Bio-AI and limitless curiosity were exceedingly inspiring and helped shape our understanding of the project. This thesis would have been far lesser without you.

We would also like to thank our fellow members of the CRAB lab for their insightful comments, enlightening discussions, and immense patience when listening to our grand ideas.

> If falsehood, like truth, had only one face, we would be in a better shape. For we would take as certain the opposite of what the liar said. But the reverse of truth has a hundred thousand shapes and a limitless field.
>
> Michel de Montaigne (1533–1592)

Axel Luiggi-Gørrissen and Trygve Johansen Woldseth
Trondheim, June 25, 2023

Contents

List of Figures

List of Tables

1 Introduction

1.1 Background and Motivation

The paradigm shift in news dissemination and consumption, driven by the digital revolution, has turned social media platforms into conduits of information, with nearly half of U.S. adults in 2021 reporting social media as their news source [3]. This transformation is intertwined with business models focused on monetising user engagement through the implementation of algorithms designed to grab the attention of their users. Unsurprisingly, this structure creates an efficient system for the rapid proliferation of fake news, which disseminates more quickly and extensively than factual news [67].

In particular, fake news spread markedly more rapidly when politics is the central theme [67]. This phenomenon might be attributable to the increasing use of social media for political manipulation. In 2020, disinformation and propaganda via social media were disseminated by state actors in 81 countries [5]. More alarmingly, the industry of political manipulation has grown beyond national boundaries, with signs of private firms engaged in organising such campaigns being identified in 48 countries [5].

The repercussions of fake news are not constrained to the political realm but extend into personal lives, causing physical harm and emotional distress. For instance, a 2021 study revealed that the circulation of disinformation during the COVID-19 pandemic correlated with heightened panic, fear, depression, and fatigue across various demographics [49]. Moreover, conspiracy theories fueled physical assaults against individuals of Asian origin due to the alleged Chinese origin of the pandemic.

Unravelling why people propagate fake news remains a challenging task. People's susceptibility to confirmation bias – seeking out and finding information which resonates with their beliefs – may offer some explanation. This predilection applies to both factual and non-factual content, with people demonstrating improved accuracy when assessing information consistent with their views [40]. Furthermore, people's capacity to discern deception barely surpasses random chance, achieving only around 54% accuracy [50]. Although accuracy improves to around 70% when assessing news headlines, the inclination to share non-factual content increases when not explicitly told to verify its accuracy, suggesting that social media's rapid pace might inadvertently foster inattentiveness and, consequently, misinformation [40].

As fake news can be harmful on multiple levels, there is a clear need to find methods to mitigate its spread and influence. To reduce the impact of people believing and spreading fake news, there needs to be an effort to inoculate the public by providing better information on how to discern factual content. Furthermore, social media platforms should be disincentivised from employing algorithms that only serve to increase engagement - they should focus on the *quality* of the content being disseminated to a greater extent. Crucially, the development of computational models capable of swiftly identifying non-factual content could prove instrumental in combating misinformation.

Artificial Immune Systems (AIS) draw inspiration from the biological immune system and present an intuitive approach to the problem of fake news detection. The biological immune system evolved over millions of years to protect a body from pathogens by efficiently distinguishing antigens which belong to the self from antigens which do not. Analogously, by treating a piece of news as an antigen, an AIS-based model could potentially discern fake news (non-self) from real news (self). Furthermore, just as the biological immune system is highly adaptive, Artificial Immune Systems are capable of adapting and learning from new information, making them apt for keeping up with the ever-evolving dissemination methods of fake news. This thesis puts forward an AIS-based model designed for the detection of fake news.

1.2 Goals and Research Questions

This section presents the goal and research questions for this thesis. The goal represents the overall intent of the presented work. The research questions are the testable sub-goals that will guide the experiments to answer the primary goal.

Goal *Identify methods to enhance fake news detection with Artificial Immune Systems through leveraging the semantic content in embeddings*

Lately, transformer-generated embeddings have gained popularity and proven their effectiveness for NLP tasks. Embedding vectors generally have a high dimensionality which can be a deteriorating factor for many AIS models. Investigating how to leverage the semantic content present in embeddings through an AIS model could also contribute to the field by identifying methods that can more appropriately handle high-dimensional spaces.

Research question 1 *How can an AIS model be adapted to operate on high-dimensional spaces?*

Given that the embeddings are high-dimensional, their direct application in traditional AIS algorithms might yield subpar results, owing to the curse of dimensionality (section 2.4.2). Exploring what modifications are needed to enhance coverage or manage the high-dimensional space more effectively is needed to properly draw from the semantic content in embeddings.

Research question 2 *How do embeddings compare to general textual features in terms of performance for fake news detection with an AIS?*

It is possible to perform content-based fake news detection by extracting textual features from a news article. Therefore, assessing the performance of embeddings against general textual features can provide valuable insights into the applicability of embeddings in the context of fake news detection.

Research question 3 *How does the accuracy of the proposed model compare to state-of-the-art fake news detection methods?*

While there are few AIS-based fake news detection models, state-of-the-art approaches generally perform training and testing using well-known datasets. Comparing the model to these serves as an evaluation of the model's applicability in the research field.

1.3 Literature Review

The literature review process applied in the project can be split into two parts: an initial, broader search, and a subsequent more directed one. The first part employed the pearl growing method, where an initial search term is used to find relevant articles, and keywords from these are then used to refine the search terms. This method displays a certain degree of stochasticity and is prone to get stuck in local term groups. For AIS research, this method failed to find many relevant recent articles since they often use unique terminology not found in older articles which established methods in the field.

The second search method was citation search, where the citations from one article are used. This allowed starting from one seminal AIS article and traversing the citation tree downwards to find relevant and recent articles. This method was used in concert with snowballing: looking into the bibliography of an article and traversing the citation tree the other way. By moving up and down the citation tree this way, a myriad of interesting articles were discovered. The main themes evaluated when deciding whether an article was useful for the project were: recency, application relevance, and technique relevance. For AIS articles, the recency criterion was relaxed as the field is not currently as productive as it has historically been. The criteria were used as fuzzy guidelines rather than strict policies with hard cutoffs; informative articles might appear in parts of the search space that would otherwise have been excluded. The relevant questions that were asked were:

- What is the state of the art for AIS-based classifiers?
- What are the advantages and disadvantages of the different components used in the identified AIS-based classifiers?
- What is the state of the art for fake news detection?
- What datasets are generally used for fake news detection?

1.4 Process Overview

The only restriction placed on the thesis topic was that it had to be related to Bio-AI. This freedom allowed the authors to explore many different ideas and approaches. A high-level overview of the process is shown in figure 1.1.

When considering the thesis topic, the authors found the application area of fake news interesting. This led to the early decision of keeping it as the application area. Deciding on the method required more debate. Neither of the authors was familiar with AIS and this led to other methods also being explored like neuro-evolution. However, due to their existing experience with neural networks and a shared curiosity to explore novel avenues within AI, neuro-evolution was eventually dismissed as a prospective thesis topic.

The authors found inspiration in the work of Sverdrup-Thygeson [60], which proposed a variety of promising pathways for future research. After careful deliberation, considering the authors' areas of interest, the decision was made to employ AIS as the principal method for the thesis.

An idea which was explored early on was to use temporal information. Mirroring the immune system's adaptability, news cycles are also dynamic, with fake news publishers frequently altering their topics. News that is fake at one point in time may not necessarily be so at another. Due to this, different approaches to harnessing this innate temporality were explored. However, due to issues with datasets not including temporal information and expanding project scope, the focus on temporality was eventually sidelined. Employing this temporal information still has merit, and could be an important topic in future research. This is discussed further in chapter 6.

Preliminary research into fake news detection unveiled that state-of-the-art models frequently incorporate some form of information fusion. Recent developments have explored the use of social context, visual media, and knowledge bases for fact-checking. It was decided that creating an AIS model employing only the content of a piece of news was preferable to designing an information fusion model for AIS. The primary motivation for this was that fake news detection using an AIS model has been relatively little researched. By maintaining a narrow focus, the authors could potentially contribute to the establishment of a robust base model, which could then be expanded upon, for instance, via information fusion, in future research.

Building upon the work by Sverdrup-Thygeson in leveraging the rich semantic content present in the embeddings from context-dependent language models was posited as a potential area of focus for the thesis. Investigating alternate methods to better exploit these embeddings, and comparing their performance with the general textual features directly derived from a news piece, might offer insights beneficial to the field, possibly opening up avenues for applying AIS to other domains dealing with high-dimensional data.

While the recent approach by Sverdrup-Thygeson employed both embeddings and semantic features, certain issues were identified with the use of embeddings. Initial investigations showed that improvements could be achieved to better leverage the semantic content in embeddings, potentially paving the way for a more successful approach.

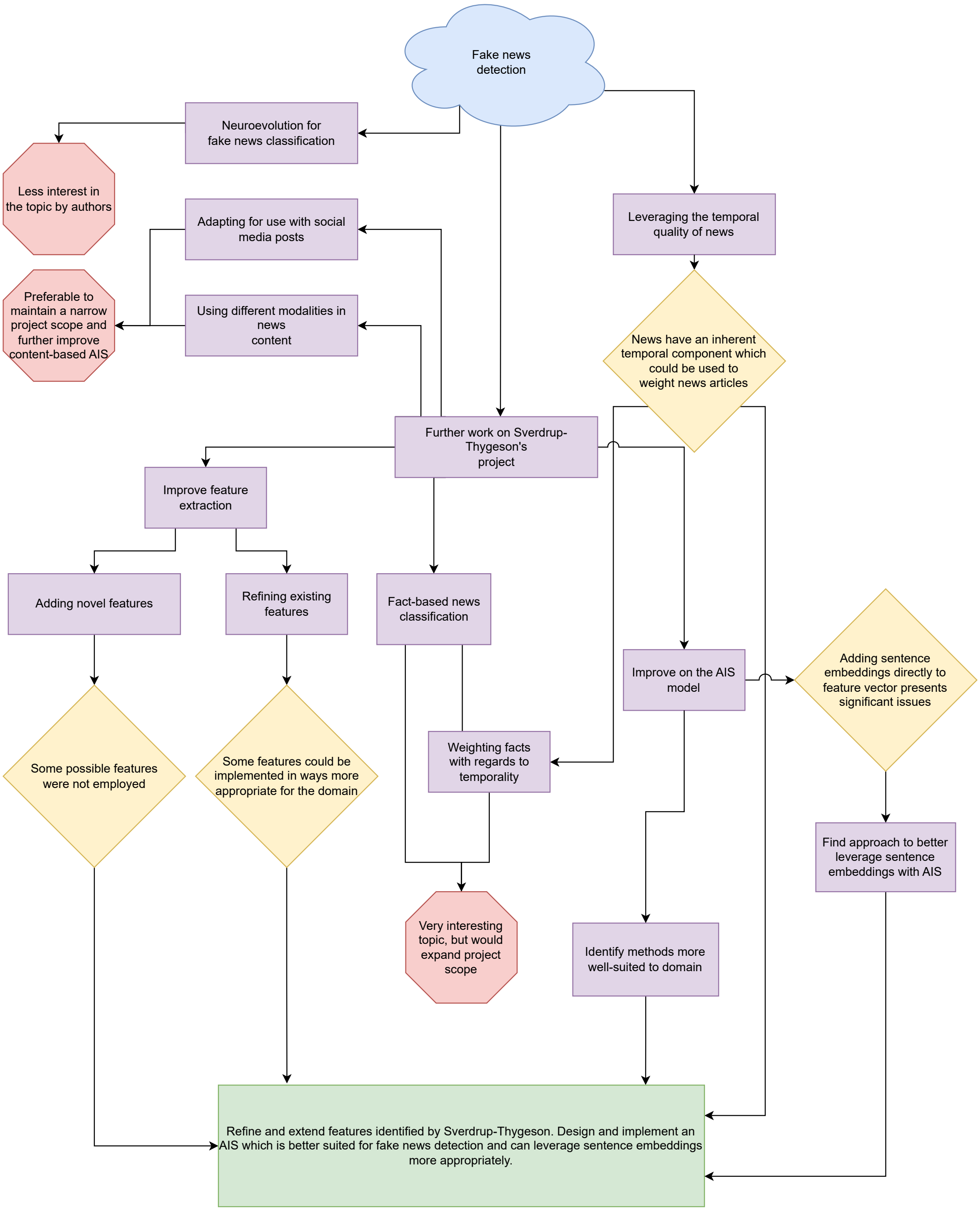

Fake news detection
Neuroevolution for fake news classification
Less interest in the topic by authors
Adapting for use with social media posts
Using different modalities in news content
Preferable to maintain a narrow project scope and further improve content-based AIS
Leveraging the temporal quality of news
News have an inherent temporal component which could be used to weight news articles
Further work on Sverdrup-Thygeson's project
Improve feature extraction
Adding novel features
Refining existing features
Fact-based news classification
Improve on the AIS model
Adding sentence embeddings directly to feature vector presents significant issues
Some possible features were not employed
Some features could be implemented in ways more appropriate for the domain
Weighting facts with regards to temporality
Find approach to better leverage sentence embeddings with AIS
Very interesting topic, but would expand project scope
Identify methods more well-suited to domain
Refine and extend features identified by Sverdrup-Thygeson. Design and implement an AIS which is better suited for fake news detection and can leverage sentence embeddings more appropriately.

Figure 1.1: High-level overview of thesis topic progression

1.5 Thesis Structure

The rest of the thesis is organised as follows. Chapter 2 gives an introduction to the theory necessary to understand both the field and the model. Following this, chapter 3 provides a concise survey of the state-of-the-art relevant to the thesis topic. The current state of the AIS field is treated here first in section 3.1 before moving on to the arena of fake news detection in section 3.2. The primary aim of chapter 3 is to delve into the advantages and disadvantages of various methods, with a special emphasis on their relevance to the research theme. Portions of the content from chapters 2 and 3 are adapted from a research project conducted during the Fall semester of 2022.

Chapter 4 presents the proposed model. Initially, in section 4.1, the core elements of the algorithm are elaborated. This is followed by a detailed explanation of the data pre-processing pipeline, as well as the features selected as salient for fake news detection within the proposed model.

Chapter 5 presents the experiments that were performed and discuss their results. Section 5.1 provides a synopsis of the experimental plan, which is split into three distinct phases. Sections 5.3 to 5.5 present these experimental phases. Each experiment section begins with a brief introduction delineating the experiment's purpose, hypothesis and setup. Following this, the results of the experiment are presented and discussed.

Chapter 6 concludes the thesis with an overview of the research outcomes and responding to the research questions posed in section 1.2. The subsequent section, section 6.2, details the primary contributions made to the field. This is followed by section 6.3, which outlines potential avenues for further research in this domain.

2 Theory

This chapter describes the background knowledge needed to understand the proposed model. In sections 2.1, 2.2 and 2.4, biological computation, the biological immune system and Artificial Immune Systems are covered. These sections relate to the core functioning of the proposed model. Knowledge of the biological mechanisms involved in the immune system is also necessary to better understand the concepts involved in AIS. Section 2.3 covers some central machine learning concepts relevant to the model, intended to aid in the understanding of certain properties of the model and experiments. In sections 2.5 and 2.6 different concepts used in text processing are explained, which is relevant to how features are extracted from a piece of news for classification in the model.

2.1 Bio-Inspired Computation

Bio-inspired computation refers to the branch of Artificial Intelligence that draws inspiration from biological behaviours and characteristics to design techniques that can be used to solve various computational problems. At the core of many bio-inspired models - as in nature - is the idea of natural selection. The central driver behind natural selection is that individuals best adapted to their environment are most likely to proliferate. A population of individuals progressively adapts to its environment over a number of generations through the processes of *genetic recombination* and *mutation*. Genetic recombination (i.e. reproduction) combines the genetic material of parents into offspring. Mutation refers to the process of random changes in genetic material, inserting diversity in the population that is not possible through genetic recombination alone [14].

2.1.1 Evolutionary Algorithms

Evolutionary Algorithms (EAs) are a class of algorithms that utilise the principles of evolution to solve optimisation problems. EAs work by simulating a population of encoded solutions over multiple generations. The solutions may change through crossover (genetic recombination) or random change (mutations). The algorithm's performance measure (termed fitness) gauges how well these solutions perform on the task at hand.

There exist myriad different evolutionary algorithms, but they all have certain properties in common. As an example, a general scheme common to many evolutionary algorithms is shown in algorithm 1.

Algorithm 1 General evolutionary algorithm

```
BEGIN
INITIALISE population
EVALUATE population
repeat
    SELECT parents
    RECOMBINE parents
    MUTATE resulting offspring
    EVALUATE offspring
    SELECT individuals for next generation
until TerminationCondition
END
```

Representation

The representation of individuals is key to the design of an evolutionary algorithm. The encoding of each individual is termed the *genotype*, while their expression of solutions is known as the *phenotype*. For example, if the solutions to a problem are integers, the phenotype of an individual might be 42, while the corresponding genotype could be the binary number 101010. The mapping from genotype to phenotype would in this case be a conversion from binary to decimal. In some cases, the genotype and the phenotype are equivalent, in which case no mapping is necessary.

Fitness

In order to rank individuals in a population it is necessary to define a fitness function that describes the soundness of a given solution. This is often used as the basis for selection later (both for reproduction and survival). This fitness function varies from application to application and is generally designed for the task at hand.

Variation Operators

As in nature, diversity is inserted into a population through the processes of mutation and genetic recombination (known as crossover in the context of EAs). The mutation operator is applied to a single genotype, producing a slightly different genotype as output. Mutation is stochastic - whether a genotype is subjected to it or not and what part of the genotype the mutation is applied to, is subject to chance. Crossover combines the genotype of two or more chosen parents into a number of offspring. Like mutation, crossover is a stochastic operator, the parts selected from the parents and how they are combined into offspring is random [14].

Exactly which variation operators to apply is dependent on the problem being tackled. For example, the travelling salesman problem generally encodes the genotype as a permutation of destinations, which the crossover and mutation operators need to take into account so that the genotype of each offspring is still a permutation.

Selection

Selection happens in two parts of the evolutionary algorithm loop: selection of parents for recombination, and survivor selection. One selection method is elitism, where the top-performing individuals are selected. However, this has the effect of reducing the diversity in the population over time since only individuals best suited to the current context survive, and the algorithm may therefore converge to a local optimum - a solution that is better than the surrounding ones but not the optimal one for the problem. An alternative selection mechanism is tournament selection. Here, the individuals compete amongst each other locally, in a subset of the population. This allows individuals who are not competitive on a population-wide scale for the problem at hand to pass on parts of their genetic material; elements of the genotype may still be viable even though the whole is not [14].

Crowding

Crowding is an approach used to preserve diversity in a population by ensuring that offspring replace similar members. The process is generally performed by choosing n members of the parent population to compare against and replacing the one most similar according to a given metric with the offspring. The effect of this is that subpopulations are preserved in *niches*, as offspring only compete against individuals that are similar to themselves [14].

2.2 The Biological Immune System

The immune system is a complex set of biological processes that enable an organism to resist unwanted change - whether from external pathogens, cancer cells, or foreign objects. The innate immune system is found in most organisms and has four implements to guard against pathogens: the anatomical, e.g. skin and membranes; the physiological, such as pH values and regulating body temperature; phago- and endocytic barriers, which are cells that break down and ingest foreign particles; and inflammatory, which is a complex immune response involving signalling from the vascular system.

The adaptive immune system is specific to vertebrates and has the ability to learn from previous pathogenic infections, which is necessary in cases where the innate immune system is unable to effectively stop an infection. It comes in addition to the innate immune system as an alternate line of defence. The main function of the adaptive immune system is to differentiate between ”self”-antigens, which belong to the body, and ”non-self”-antigens, which are found on the surface of pathogens. An antigen is a molecule which is able to be bound to an antibody or receptors on the surface of T-cells. Whereas the innate immune system is antigen-independent, the adaptive immune system is able to mount a response based on specific antigens, forming the key to the concept of immunological memory. The main components of the adaptive immune system are *lymphocytes*, more specifically T- and B-cells. The surface of a T-cell has many antigen-binding receptors (known as T-cell receptors, or TCR) of a unique type, which the cell is able to adapt in response to appropriate signals. The part of this receptor that binds to an antigen is known as the *paratope*. Similarly, the part of the antigen that binds to the paratope is known as the *epitope*. During the maturation of T-cells in the thymus, cells that have antigen receptors which react to self-antigens are eliminated through apoptosis (programmed cell death) by a process known as *negative selection*. This ensures that mature T-cells do not react to the cells of the body. T-cells depend on antigen-presenting cells (APCs) for activation which are cells that express antigen fragments on their surface, typically resulting from phagocytosis of a pathogen. The activation of a T-cell stimulates it into differentiation (transition into another type of cell), either into cytotoxic T-cells or T-helper cells. The cytotoxic T-cells are directly involved in eliminating pathogens by inducing apoptosis, while the T-helper cells take on a different role by directing other parts of the immune system through cellular signalling [34].

B-cells, like T-cells, are involved in the adaptive immune system and present a unique antigen receptor on their surface. However, they do not require APCs to discern antigens. When stimulated by an antigen that matches their receptor, they differentiate into either plasma cells or memory cells. Plasma cells produce antibodies, which are proteins that attach to antigens on the surface of pathogens, flagging them for elimination by other parts of the immune system. The function of memory cells is similar, but these are longer-lived, and provide the immune system with a ”memory” of the infection. If the infection reoccurs, these cells are able to quickly marshal a response by producing antibodies [34].

The binding strength between an antigen's epitope and an antibody's paratope is termed *affinity*. Affinity directly affects the proliferation of antibodies through the process of *affinity maturation*: successive exposure to an antigen will produce greater numbers of antibodies capable of binding to that specific antigen, with progressively higher affinity, illustrated in fig. 2.1. Whereas affinity is the strength of the monovalent binding between a single paratope and epitope, *avidity* is the measure of the multivalent binding strength of all binding sites of the antibody.

The adaptive immune system is theoretically able to respond to any pathogen invasion of the body. This is possible through the processes of *somatic hypermutation* and *clonal selection*. When a receptor of a B-cell is stimulated, the cell undergoes rapid proliferation, subject to a mutation rate many orders of magnitude higher than that seen in the other cells of the body. The process is known as somatic hypermutation, and concentrates the mutations in the parts of the genome coding for antigen recognition, leading to more specific receptors on the offspring antibodies. The resulting progeny need to compete for resources in the body. Through clonal selection, only the antibodies with the highest affinity towards the antigen survive, with the rest being eliminated

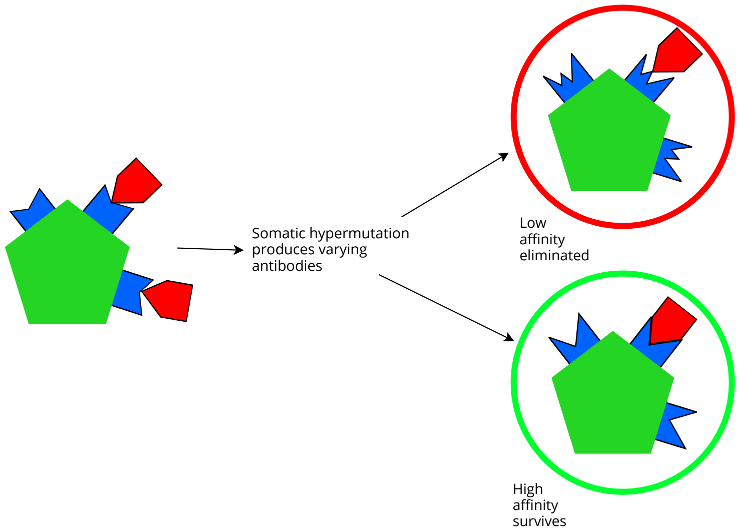


Figure 2.1: Affinity maturation in the biological immune system

2.3 Machine Learning Concepts

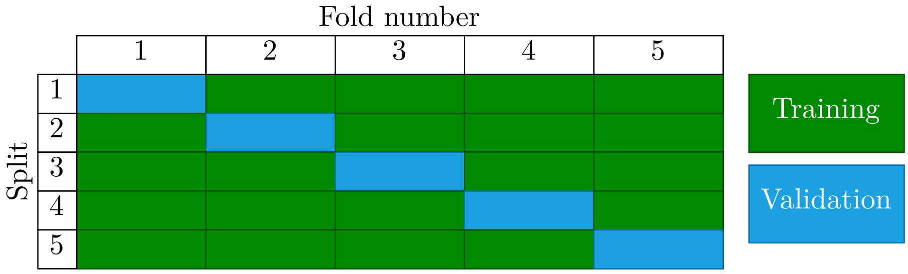
2.3.1 Classification

Classification refers to the task of determining which of a number of categories an observation belongs to. One way to do this is to encode the observation in a way that can be plotted in a coordinate system, e.g. by encoding each attribute in the observation as a real-valued *feature*. The resulting encoding of all the attributes is termed the *feature vector* and can be represented in an n-dimensional space, where n is the length of the feature vector [51].

In *supervised learning*, a machine learning model is given a set of inputs along with the appropriate outputs and attempts to learn a function to map a given input to an appropriate output. When used for classification, the model is given observations along with the correct class. The resulting mapping function can be thought of as partitioning the feature space in such a way as to properly classify new observations [51].

2.3.2 Cross-validation

Cross-validation is a technique used for better evaluating the performance of a machine learning model. The idea is that the training data is split into k *folds* with each containing $\frac{1}{k}$ of the training set. The model is then trained for k rounds, with each fold being used for validating the model once, and the $k-1$ remaining folds used to train the model. The downside of this is that it requires k times longer computation time, but the resulting average test score will provide a better estimate of the model's performance. It is important to note that the testing set should not be included when splitting into folds, as this would cause *knowledge leakage* - providing information to the model which can give it an unwanted advantage on the test set.



Split \ Fold number	1	2	3	4	5
1	🟦	🟩	🟩	🟩	🟩
2	🟩	🟦	🟩	🟩	🟩
3	🟩	🟩	🟦	🟩	🟩
4	🟩	🟩	🟩	🟦	🟩
5	🟩	🟩	🟩	🟩	🟦

🟩 Training

🟦 Validation

Figure 2.2: k-fold cross validation with 5 folds

2.3.3 Boosting

Boosting is a technique applied in machine learning that combines an ensemble of several weak learners into a single strong predictor. Most boosting algorithms work by weighting the data points according to whether or not they have been misclassified; data which is not appropriately classified gains a higher weight, while properly classified data have their weight lowered. The weak learners in the next round of boosting thus focus more on data with a higher weight, attempting to more appropriately classify them. The general formulation of a boosted classifier is given as follows:

$$F_T(x) = \Sigma_{t=1}^{T} f_t(x) \tag{2.1}$$

where f_t is a weak learner which takes some input x and returns a hypothesis as to which class x belongs to.

AdaBoost (Adaptive Boosting) is a boosting algorithm which has the property that as long as each weak learner performs better than random guessing on the training set, AdaBoost will eventually be able to predict the training set perfectly, given a large enough number of weak learners. In AdaBoost, the weak learners are adapted through a coefficient α_m, which is selected such that the training error err_m is minimised. First, the weights of all the observations are initialised as $w_i = \frac{1}{N}$ for $i = 1, 2, \ldots, N$, where N represents the total number of instances in the dataset. For training the weak classifiers, each iteration $m = 1$ to M, where M is the total number of weak classifiers, the following steps are undertaken:

1. *Fitting the classifier:* A weak classifier $G_m(x)$ is trained on the dataset using the current set of weights w_i.

2. *Calculating the weighted error rate:* The error rate for the classifier is calculated as follows:
$$err_m = \frac{\sum_{i=1}^{N} w_i I(y_i \neq G_m(x_i))}{\sum_{i=1}^{N} w_i}$$

Here, I is the indicator function which is 1 if $y_i \neq G_m(x_i)$ and 0 otherwise.

3. *Computing the classifier's coefficient:* Each classifier has a coefficient, α_m, which represents the amount of say the classifier has in the final vote. It is calculated as:

$$\alpha_m = \log\left(\frac{1 - err_m}{err_m}\right)$$

4. *Updating weights:* The weights of the observations are then updated according to the formula:

$$w_i = w_i \cdot \exp[\alpha_m \cdot I(y_i \neq G_m(x_i))]$$

This step increases the weights of the misclassified instances and decreases the weights of the correctly classified instances.

After iterating through all M classifiers, the final model, denoted by $G(x)$, is given by:

$$G(x) = \text{sign}\left(\sum_{m=1}^{M} \alpha_m G_m(x)\right) \tag{2.2}$$

In this equation, the sign function serves as the final classifier and takes a weighted vote over the M classifiers, each of which is weighted by α_m.

2.4 Artificial Immune Systems

Artificial Immune Systems (AIS) are models which draw inspiration for their architecture from concepts found in biological immune systems. It should be noted that these models generally do not attempt to be *bio-plausible*, in that they do not closely mimic the biological immune system, but rather extract ideas which can be usefully applied to the computational domain. As the biological immune system is incredibly complex, attempting to closely emulate it would require immense computing resources and may not have a tangible benefit for the application area.

AIS algorithms can be broadly classified into four categories: negative selection, clonal selection, immune network, and danger theory algorithms. However, there also exist hybrid algorithms which blur the line between these to some degree. Most AIS algorithms use the representational abstraction of the *shape-space*, an n-dimensional space in which the algorithms' antibodies and antigens operate [36].

Negative selection algorithms, or NSA, draw inspiration from the self-nonself dichotomy in the immune system, where it is able to classify elements as foreign (nonself) or part of the body (self). The core idea of these algorithms is to create random detectors in the shape-space, and eliminate any that overlap with the training data (the self). This mirrors the maturation of antibodies in the biological immune system.

The clonal selection principle was introduced by Burnet [6] in 1959, and presents the idea of affinity maturation in immune cells. The central tenet of this theory is that the immune system optimises itself against intrusion by mutating the cells with a rate inversely proportional to their affinity towards an antigen, and the proliferation of cells is proportional to their affinity. CLONALG [8], or Clonal Selection Algorithm, is based on this theory and has inspired a number of algorithms based on similar principles.

Immune network models (INMs) are based on the interactions and communications among immune cells, inspired by idiotypic interactions in the biological immune system first introduced by K [23]. These models have been applied to problem domains such as classification, optimization, and anomaly detection. INMs use concepts like antibody diversity, self/non-self discrimination, and immune response dynamics to develop computational algorithms.

Dendritic cell algorithms (DCA) are based on the "danger theory" explanation of the immune system. This theory posits that the immune system does not identify self/non-self entities, but rather reacts to *signals* emitted in the immune system. DCA algorithms eschew the self-nonself discrimination employed by other AIS algorithms and aim to reduce the number of false positives produced.

The biological immune system evolved to be able to combat any infection of the body, no matter the aetiology. This is in contrast to a central theorem of machine learning algorithms; the *no free lunch theorem* states that averaged over all possible problems, an algorithm will perform as well as any other. As a consequence, the best-performing algorithms for a specific application are generally tailored to that application. Artificial immune systems have been applied to a broad variety of problems, but they can generally be grouped into the categories of intrusion detection, classification, and optimisation. The remainder of this section will focus on concepts applied in AISs used for classification, as that is the most relevant to the task at hand.

2.4.1 Antigens and Antibodies

In an AIS, the antigens represent an observation or entry in a training set. This is generally encoded as a feature vector of n values. Accordingly, this creates an n-dimensional space where the values of each antigen's feature vector indicate its position in this space. The antibodies in an AIS are represented as entities located in a certain position in the shape-space, with an area known as the recognition region (RR) being the extent to which it is capable of interacting with an antigen in the shape-space. The simplest approach to antibody representation is an n-dimensional vector with the same number of feature values as the antigens; however, this limits the possibilities of the system by only allowing negative selection, and having the recognition region of every antibody be identical. By extending the antibody definition to contain information on the class it represents, an AIS can perform multi-class classification. Further extending the antibody definition with parameters that define the geometry of its RR can help create a more effective classifier, as each antibody is able to more appropriately cover its own local region of the shape-space. Variable-sized detectors can better fill "holes" in the shape-space not possible when employing constant-sized detectors [21].

An issue that can arise when performing classification using an AIS model with unbalanced datasets is that the population of antibodies tends to move towards the majority class, since a high accuracy can be achieved by classifying every antigen as the same class. Various approaches can be applied to mitigate this effect, for example by devaluing antibodies which cover many antigens already covered by others, and correspondingly increasing the value associated with covering undetected antigens.

2.4.2 Recognition Regions

The question of the shape of the antibodies' RR is important and is connected to both the representation and the affinity function used. The simplest recognition region, which is commonly employed, is an n-sphere (in which case it is usually called an **A**rtifical **R**ecognition **B**all (ARB)). The advantage of this is its simplicity, however, as shown by Hart [19], it is not a judicious choice for every problem. Some may require the different dynamics exhibited by other RR shapes to achieve effective classification. Hart also investigates two alternatives to the spherical RR shape: box-shaped, which is simply a hypercube; and cross-shaped, which is inspired by distinct binding sites on cells. Their results indicate that the shape and volume of the RR heavily influence the performance of the AIS model, though the article's scope is limited to only looking at these three distinct shapes. As the likelihood of the feature values of the antigens is not necessarily constant over the shape-space, some models (e.g. AISLFS [11]) employ *local feature selection*, which has the purpose of finding the appropriate subset of features associated with each antibody for a given area in the shape-space.

The degree to which an antigen matches an antibody is termed affinity, like its biological counterpart. There are various ways to calculate this, but the most common is inverse Euclidean distance, i.e. the affinity is higher for antigens closer to the centre of an antibody. While an antigen will bind to all antibodies with an enclosing RR, the antibodies may be of different classes. In this case, a method is needed to assign a class to the antigen. Some models use a voting heuristic to classify the antigen in this case, weighting the vote from each antibody by the affinity towards them. Other models employ nonoverlapping RRs, in which case this question becomes moot.

The *curse of dimensionality* refers to the problem of having vectors with a large number of features, resulting in a high-dimensional shape-space. This causes bounded recognition regions to fill a progressively smaller amount of the shape-space as the dimensionality increases. For example, consider an antibody with a RR that covers half of the space for each feature. For a 1-D space, it would cover half, 2-D a quarter, 3-D an eighth, and so on. For an n-dimensional space it would only cover $\frac{1}{2^n}$ of the total volume of the space, despite covering half the length along each feature axis [36]. Figure 2.3 shows how increasing dimensionality affects the coverage of the shape-space. For example, whereas a 2-D RR needs to cover only half of each axis to fill 25% of the volume of the shape-space, a 10-D RR needs to cover 87% along each axis to fill the same volume.

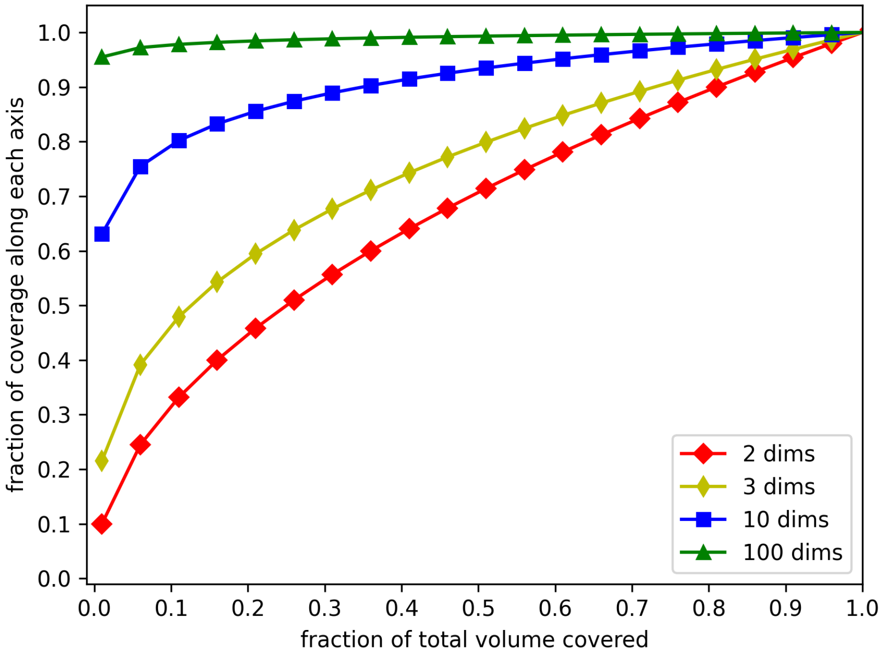


Figure 2.3: Effect of increasing dimensionality on shape-space coverage (adapted from [36])

2.5 Natural Language Processing (NLP)

NLP refers to using computers to communicate and learn from the natural language used by humans. There are a wide array of methods to help achieve this, from simple statistical approaches like term counting, to state-of-the-art language models which are able to capture the semantics of the human language.

2.5.1 Tokenisation

Tokenisation is the process of splitting text into separate units, which are termed tokens [51]. This is an important first step in many NLP pipelines. The tokenisation process is more than just splitting into spaces and periods: e.g. in the sentence ”I won’t go to the C.I.A.” the second word, ”won’t”, is generally split into the tokens ”will” and ”not”. Further, the periods in the abbreviation ”C.I.A” will be ignored and produce one token (except the last period since it marks a sentence full stop). There exist various ways to accomplish tokenisation, and choices made here can have ramifications concerning how a model performs.

2.5.2 Term Frequency

Calculating the frequency of the occurrence of a word in a text is known as term frequency (TF), defined as is the number of occurrences of a word divided by the number of words in the text. TF gives an indication of the relative importance of a word in a text. A related method is Term Frequency - Inverse Document Frequency (TF-IDF). This method lessens the importance of words that occur often, yet carry little semantic information, e.g., ”a”, ”the”, ”to”, etc. TF-IDF decreases the importance of words that appear often in multiple texts in a corpus while increasing the importance of those words that occur relatively more rarely. It is defined as the following:

$$\textbf{TF-IDF} = \frac{f_{t,d}}{\Sigma_{t' \in d} f_{t',d}} \cdot \log\left(\frac{N}{1 + d_t}\right) \tag{2.3}$$

where $f_{t,d}$ is the number of times that a term t appears in a document d, N is the total number of documents in the corpus, and d_t is the number of the documents where term t occurs.

2.5.3 Lemmatisation

A text often includes several inflected forms of the same word, as well as synonymous words carrying the same general meaning. Lemmatisation is the process of grouping these similar words, i.e., ”better, great, good” are all mapped to ”good”. This contrasts the related process of *stemming*, which naively attempts to group words by truncating them so that only the word stem remains. A downside to stemming is that it may also lead to unrelated words being stemmed to the same root, e.g. the words ”business” and ”busy” might be stemmed to the root ”bus-”. Lemmatisation extends stemming with morphological analysis to only truncate inflected variants of the word, and to return the dictionary form of it (the *lemma*).

2.5.4 Embeddings

Embeddings are distributional representations of content in a vector space which capture semantic information by grouping similar words. Embedding meaning may be done both to the individual words or to sentences as a whole. In the embedding space, the different hyper-dimensional directions encode meaning. As an example: if the embedding vector for the word ”royalty” is subtracted from the embedding vector for the word ”King” the resulting point in the embedding space lies closest to the embedding for the word ”man”.

Word2vec [37] and GloVe are examples of context-independent embedding models. In effect, this means that the embedded representation of a word produced by the model remains the same no matter the context. The key difference between the output of these models is that word2vec produces an output dependent on the local context. In contrast, GloVe uses global word co-occurrence (meaning it considers the entire training corpus). BERT, another embedding model, is based on the architecture of *transformers*, which use an attention mechanism to capture contextual information through the position of each word in a sentence.

Whitening and dimensionality reduction of embeddings

The contextual embedding space of BERT suffers from anisotropy, where vectors are not uniformly distributed, but are instead concentrated in a relatively narrow ”cone”. As a result, the average cosine similarity of any two word-embeddings is 0.99 [15], complicating comparison. A similar challenge afflicts the BERT sentence embedding space [28].

To address the issue of anisotropy in the BERT sentence embedding space, Su *et al.* [59] have proposed to apply whitening, which is a technique that changes the covariance matrix of the input vectors to the identity matrix, resulting in the inputs being uncorrelated. Whitened BERT sentence embeddings display improved performance on semantic similarity tasks. Moreover, the method provides for dimensionality reduction. The goal of the whitening process is to find a linear transformation for the set of input embeddings $\{x_i\}_{i=1}^N$ such that the mean value of the transformed embeddings is 0 and the covariance matrix is the identity matrix:

$$\tilde{x}_i = (x_i - \mu)W$$

First, the mean vector μ and covariance matrix Σ of a given set of input embeddings are calculated as follows:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\Sigma = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^T(x_i - \mu)$$

To transform the covariance matrix Σ into the identity matrix, there needs to be a matrix W that has the following property:

$$I = W^T \Sigma W$$

Equivalently:

$$\Sigma = (W^{-1})^T W^-1$$

As the covariance matrix can be expressed as follows by singular value decomposition (where Λ is a diagonal matrix):

$$\Sigma = U \Lambda U^T$$

and W^-1 therefore is equivalent to $\sqrt{\Lambda}U^T$, W becomes:

$$W = U\sqrt{\Lambda^-1}$$

To apply dimensionality reduction, only the first k columns of the transformation matrix W are applied in the transformation shown in equation 2.5.4. Furthermore, the process can be updated with new embeddings:

$$\mu_{updated} = \frac{n_{old}}{n_{old} + n_{new} + 1}\mu_{old} + \frac{n_{new}}{n_{old} + n_{new} + 1}\mu_{new}$$

$$\Sigma_{updated} = \frac{n_{old}}{n_{old} + n_{new} + 1}\Sigma_{old} + \frac{n_{new}}{n_{old} + n_{new} + 1}\Sigma_{new}$$

where n_{old} and n_{new} are the number of embeddings in the existing set and the number of new embeddings respectively.

2.5.5 Transformers

Transformer models have rapidly become the predominant technique for NLP tasks, largely due to how they leverage *attention*. The principles of cognitive attention inspire attention mechanisms in NLP. They work by enhancing some parts of a given input while diminishing others, essentially focusing on the important parts of the input. Self-attention is the specific type of attention which lies at the core of the transformer architecture. This mechanism relates the input to itself, allowing the transformer to learn which parts of a sequence are important, given the rest. Self-attention is accomplished by projecting the input vector into three different vectors (known as the key-, query-, and value vector) by applying learned projection matrices. The encoding result of the i-th word in the input to the model is calculated as follows:

$$\begin{aligned} r_{ij} &= (\mathbf{query}_i \cdot \mathbf{key}_j)/\sqrt{d} \\ a_{ij} &= e^{r_{ij}}/(\sum_k e^{r_{ik}}) \\ \mathbf{embedding}_i &= \sum_j a_{ij} \cdot \mathbf{value}_j \end{aligned} \tag{2.4}$$

where d is the dimensionality of the key vector. Through the process of *multiheaded attention*, the transformer can achieve better granularity in the information contained in the embeddings. This is done by applying the self-attention mechanism separately to subsets of the input and concatenating the result. Employing multi-headed attention ensures that information is not lost by being averaged out over the whole of the input.

The self-attention mechanism opens the door to context-dependent embeddings. For example, the word ”strike” when used in the sentences ”the workers were on strike” and ”she bowled a strike” have different meanings from the context of the words accompanying it, and will correspondingly produce different embeddings.

2.6 Fake News Detection

Fake news detection strategies can be divided into four overarching approaches [55]. Some models also employ a hybrid system with multiple strategies in concert.

1. **Knowledge-based** strategies attempt to assess the veracity of the claims made in the news content.
2. **Content-**, or semantic-based approaches, try to classify a piece of news based on its writing style and semantic features, such as misspellings, modal adverbs, subjectivity, etc.
3. **Propagation-based** models look at the dissemination of news in social media. As fake news spread in a different pattern and at a different speed than real news, this approach seeks to leverage the difference in propagation for detecting fake news.
4. **Credibility-** or source-based approaches attempt to classify content with regards to their source, e.g. author, site, and/or references.

The architecture presented in section 4 uses a content-based approach. Accordingly, the other methods won't be explored in-depth.

Content-based detection of fake news hinges on extracting and representing useful features from the content of a piece of fake news. Linguistic features capture information relating to the writing style of the piece of news, e.g. negations, vulgar words, sentence complexity or typos. Since fake news is created with the intent to deceive an audience, the writing style of fake news differs from real news, containing a style designed to stimulate the reader's attention to a greater degree. Sentiment-based features attempt to model the emotion that a piece of news attempts to convey. To capture this, term-frequencies of words carrying positive and negative connotations can be used. The frequency of exclamation marks can also indicate sentiment. Latent features are not directly observable, but can be extracted with language models to produce embeddings carrying semantic information which is not readily apparent. Some models may also employ features extracted from other modalities than textual, e.g. images, video, or sound.

3 State of the Art

This chapter presents current research relevant to the presented work. First, the current state of the art in AIS models will be presented with a short introduction to application areas. Following this, some key design decisions in recent models are highlighted, as well as their advantages and disadvantages. Finally, different approaches for fake news classification are discussed along with their advantages and disadvantages.

3.1 Artificial Immune Systems

Different algorithms vary in their closeness to the biological inspiration and what parts of it they derive their methods from. This section will give a short introduction to the applications of state-of-the-art AIS algorithms. In contrast, the following sections will discuss central AIS characteristics and how current models handle them.

Some recent applications of AIS algorithms include job shop scheduling [30], fault detection [29][72], and stock-market manipulation detection [48]. The application of intrusion detection is naturally suited for AIS models [1] and is the area of highest research output today [2][10][53][13].

The biological immune system is able to accurately distinguish different antigens, in effect acting as a natural classifier [18]. Accordingly, the development of AIS for classification is a logical step, and there have been several models proposed for applications ranging from hyperspectral image classification [73] to spam detection [52]. AIS models have also been applied for document classification, demonstrating suitability for classification tasks involving natural language [63].

For the specific task of fake news detection with an AIS, there has been little research; however, Sverdrup-Thygeson [60] has proposed a model which shows promise for further AIS development targeting this domain.

3.1.1 Recognition Region

As demonstrated by Hart [19], the dynamics of the RR can significantly affect the model's performance. To address this, some models have explored the use of non-spherical shapes, such as Ellipsoidal-AIS [39], which investigates the use of evolved elliptical RRs. On the other hand, AISLFS [11] and CLONALG-LFS [69] use non-spherical RRs in another way by constructing uniquely shaped RRs for each antibody by selecting only relevant features in a local part of the shape-space, known as local feature selection.

The likelihood of a feature's value is not uniform across the shape-space, and values that are highly probable in one region may not be present in another. In order to address this issue, AISLFS and CLONALG-LFS employ a binary vector that represents the active features for each antibody, i.e. the ones used for affinity calculation with the antigens. This approach has the advantage of reducing the dimensionality of the RRs by eliminating redundant features in a local part of the shape-space. Although both AISLFS and CLONALG-LFS are similar algorithms, their distinctions lie in their initialisation and affinity calculation, which are further discussed in sections 3.1.4 and 3.1.3.

AISLFS mutates the set of active features for each antibody, which results in an uneven RR in the shape-space of the entire feature vector. As an example, deactivating a single feature from a feature vector consisting of three elements, while the other two remain activated, will transform the RR in the shape space from a sphere to a cylinder, parallel to the axis of the deactivated feature. The benefit of this approach is that the algorithm can better take advantage of the varying probabilities of features in the shape-space. For instance, if a feature is unlikely to impact the classification in a certain region, deactivating the feature for an antibody in that region can enhance classification by covering the shape-space to a greater extent, encompassing more antigens.

On the other hand, Ellipsoidal-AIS [39] approaches non-spherical RRs differently by allowing for mutation of the length and orientation of the axes of an n-spherical RR, creating a hyper-ellipsoidal RR. The motivation behind employing ellipsoidal RR is to allow for better coverage of antigens in regions where some features lie closer than others, similar to the challenge addressed by AISLFS and CLONALG-LFS. By stretching an axis towards an arbitrary length and allowing for free rotation along all axes, antigens can be effectively grouped by a single antibody which would not be possible with spherical RR. The centre of each antibody is also subject to mutation, in contrast to AISLFS, where they are static and only the active feature vector is mutated.

Ellipsoidal-AIS outperforms several other AIS methods on multiple benchmark classification data sets, thus supporting the findings of Hart that the model's performance is dependent on the dynamics of the RR. While AISLFS and CLONALG-LFS also achieve high scores compared to other state-of-the-art AIS classifiers, they score slightly lower - around 3-5 percentage points - than Ellipsoidal-AIS for the Diabetes and Heart Statlog datasets (the sets that all three algorithms report results for). The downside of Ellipsoidal-AIS is that it introduces additional complexity both in space and time - the model takes around 2 hours per fold for 5-fold cross-validation on the Diabetes benchmark set. In comparison, AISLFS and CLONALG-LFS have much lower computational complexities.

3.1.2 Evolutionary Process

The evolutionary process presented in section 2.1 (algorithm 1) is also relevant to AIS models. This section discusses how selection, variation and diversity are approached in recent models.

Parent & Survivor Selection

Clonal selection (sections 2.2 and 2.4), first employed in CLONALG [8] is widely used in AIS models. In CLONALG, every antibody in the population is exposed to each antigen. The n antibodies with the highest affinity for each antigen are selected and cloned proportionally to their affinity. Then, the affinity of each clone to the antigen is calculated, with the best-performing clone becoming a candidate for inclusion in the set of memory cells - the set of antibodies with the highest affinity towards each antigen. Additionally, the d highest-affinity clones replace the d lowest-affinity individuals in the population.

A slightly different approach to parent selection is applied in AISLFS. Here, every antibody is selected as a parent and proliferates n clones each generation. The motivation behind this approach is that every antibody should already be placed in an ideal location through the initialisation procedure. Another method is applied in the model by Sverdrup-Thygeson. Here, tournament selection is used when choosing parent antibodies, where potential parents compete amongst a population subset - in contrast to the elitism approach of CLONALG. Allowing for the selection of sub-optimal parents as in these models introduces greater diversity in the antibody population which could aid in escaping local optima.

Survivor selection is also accomplished differently in AISLFS. The n best-performing of the resulting clones replace the n worst-performing antibodies in the population, regardless of whether they perform better or worse than the antibodies they replace. On the other hand, MAIM applies survivor selection to the combined populations of offspring and parents. The model employs fitness-proportionate selection, where the probability of an antibody becoming part of the next generation equals its affinity divided by the sum of affinities of all antibodies in the population. Both the survivor mechanisms of MAIM and AISLFS function to increase diversity. MAIM's survivor selection is stochastic, though if an antibody has much higher fitness than the rest of the population, it will nearly always be selected. In AISLFS, the same antibody will likely oscillate in and out of the population due to children strictly replacing parents.

CLONALG-LFS [69] draws inspiration from CLONALG for its parent selection mechanism. Meanwhile, it uses a different survivor selection method: parents and offspring are sorted in descending order of affinity and only the best are chosen for the next generation.

Variation Operators

The mutation procedures in CLONALG-LFS [69] and AISLFS [11] do not directly mutate the antibodies' feature values, but only the set of included features. Each clone of an antibody has a unique mutation in one element of the feature vector. A consequence of only using this mutation is that the feature vectors are static and the antibodies are locked to their parents' location in the shape-space.

Meanwhile, Ellipsoidal-AIS employs three mutation operators: orientation, where the centre of the ellipse is adjusted; length, where the lengths of the axes of the ellipse are adjusted; and orientation, where the ellipse is rotated in two randomly chosen directions about the centre. On the other hand, VALIS [26] uses mutation to adjust the centre of each antibody in the shape-space and the radius of its recognition region. The mutation probability is $1/(1+n_f)$ for both parameters (where n_f is the number of features in the feature vector) until at least one mutation occurs. Both approaches present advantages and disadvantages. VALIS is computationally much simpler but lacks the improved classification accuracy that a non-spherical RR may provide.

While adding more mutation operators may allow for the creation of more accurate antibodies it also introduces additional complexity as the algorithm has a larger space of mutations to search.

Crossover is not commonly used in AIS models. However, certain models apply it. VALIS [26] is one such model. Here, crossover is used on the radius and centre parameters of the antibodies. The effect of this is not investigated. MAIM [4], an island model AIS with elements inspired by VALIS, uses crossover to distribute genotypes across the islands better, with uncertain effects. Another recent model that draws inspiration from VALIS eschews the use of crossover altogether while maintaining comparable results [60], indicating that crossover may not provide a significant benefit for classification. On the other hand, Dai *et al.* [9] propose a crossover operator for CLONALG inspired by quantum mechanics, achieving improved convergence speed and accuracy for travelling salesman problems. They do not investigate the applicability of the operator for classification tasks.

Diversity Preservation

To preserve diversity in the antibody population, VALIS takes inspiration from the concept of *niching* and includes a term in its fitness function to model fitness sharing so that the evolutionary process is encouraged to spread the antibodies over the search space. The effect of the sharing factor is investigated, showing improved classification accuracy. This sharing factor is also employed in MAIM Baug *et al.*, a model inspired by VALIS. A different approach to diversity preservation is proposed in Li *et al.*, where a crowding heuristic is applied to reduce population density in crowded areas of the shape-space. For the application of job shop scheduling, the model employing the crowding heuristic achieves better results than the model without the heuristic applied.

3.1.3 Affinity Calculation and Class Assignment

The most common approach to affinity calculation is to use either the Euclidean distance or the Manhattan distance in the shape space between an antigen and an antibody's centre, which has the advantage of simplicity. The affinities are then used when performing class assignments for the antigens.

In AISLFS the affinity between an antibody and an antigen is inversely proportional to the distance between them - antigens which lie closer to the centre of an antibody's RR have greater affinity than those close to the edge. The model explores using both Manhattan distance and Euclidean distance as distance measures, with similar results on most sets [26]. The affinity is only calculated for the selected feature subset of each antigen. In contrast, the affinity function adopted by CLONALG-LFS does not employ parameters dependent on a RR, instead relying on a local clustering approach to minimise the intra-class distance and maximise the inter-class distance between antigens and antibodies. The advantage of the approach is that it is independent of a classification method and does not explicitly define a RR in its training loop, allowing for more freedom when choosing how to perform class assignment.

VALIS [26] uses binary affinity, where the affinity between an antigen and an antibody is 1 if the antigen lies inside the antibody's RR, otherwise 0. The model's performance is improved by calculating a weighted sum of the affinities from every antibody covering the antigen. Ellipsoidal-AIS also adopts binary affinity but foregoes considering affinity between a single antigen and antibody. Instead, it looks at antibody fitness, defined as the ratio of correctly classified to incorrectly classified antigens in an antibody's RR. The approach used in Ellipsoidal-AIS has the advantage of being computationally efficient compared to VALIS, as each antigen is covered by exactly one antibody, thus not requiring the calculation of weighted voting from every overlapping RR as done in VALIS. Still, it may lose some granularity in selection and RR placement, as the method does not allow for considering the placement of the antigen in the RR.

The question of how to perform class assignment can be approached in many different ways. The simplest method is to count the antigens in an antibody's RR and assign the antibody's class to the query antigen. A different strategy is to sum up the combined affinities for each class for every antigen in an antibody's RR and assign the class with the highest sum. These approaches are employed in AISLFS [11], with the affinity-weighting mechanism scoring higher. VALIS [26] uses a similar affinity-weighting scheme, but also includes an additional weighting term to account for the ”reliability” of the antibodies' prediction. It is defined as the sum of affinities to antigens of the antibody's class in its RR divided by the sum of affinities towards antigens of a different class in its RR. The advantage of the weighting term is that antibodies with large overlapping RRs containing many antigens of different classes have less effect on classification than more accurate antibodies. VALIS also includes a term that weighs the influence of antibodies that cover few antigens less heavily to avoid overfitting caused by antibodies only covering a small percentage of the antigens.

Imbalanced datasets are challenging since the antibodies will likely become biased towards the class with the largest number of samples. The weighting term used in VALIS to prevent overfitting may exacerbate the bias, as antibodies that cover few antigens relative to others are likely to be in a minority class. The idea of memory cells, introduced by CLONALG [8], offers a potential solution to this challenge. Ellipsoidal-AIS is one algorithm adopting this method by conserving the fittest antibody relative to each antigen for the next generation. This approach, combined with the fitness score in Ellipsoidal-AIS, helps mitigate the potential risk of overfitting caused by small antibodies.

3.1.4 Initialisation

Initialisation of the antibodies is important, as it can drastically reduce running time and increase convergence rate. When distance is used as an affinity measure, the features of the data are usually normalised so that the distance measure is not affected by unequal feature scaling, which would weigh some features as more important than others.

Random initialisation of the antibodies is the simplest and is commonly used [69] [70] [72]. With this approach, the initial antibodies are randomly placed in the shape-space. The advantage of this is decreased computational complexity and higher initial diversity [30], however, it may lead to longer training times since initial fitness will likely be low.

Another approach to initialisation is to use the training set to seed the antibodies directly [11] [29]. This strategy aims to ensure that every antigen is initially covered by at least one antibody, which can be especially valuable for sparse datasets. VALIS adopts an approach leveraging antibody seeding, wherein centres for each antibody are chosen from the antigens without replacement during initialisation. The radius of the RR is set as the distance to a random antigen of the same class. However, a potential drawback of this method is that it may result in antibodies that cover antigens of different classes if antigens of the same class are not clustered together in the shape-space. MAIM [4] compared the use of random initialisation and the method proposed in VALIS and found that VALIS' strategy significantly improved performance.

In AISLFS, the radius of each RR is determined individually during initialisation based on the *cross-reactivity threshold*, defined as the greatest radius possible without the RR containing an antigen of a different class. This approach is similar to the one adopted in VALIS but avoids the issue of covering antigens of a different class. While CLONALG-LFS is a similar algorithm, it approaches initialisation differently, inspired by CLONALG. A pool of antibodies is initialised for every antigen, with the size governed by the parameter N_{pop}. The algorithm is also more robust in its initialisation by having the r antibodies with the lowest affinity replaced by randomly initialised antibodies, reducing the dependence on representative training samples seen in AISLFS.

3.2 Fake News Detection

Before attempting classification, it is necessary to delineate what constitutes ”fake news”, as the definition of the term has been somewhat diluted. In fake news research, there is no generally agreed definition of the term, however, there is emphasis placed on three key concepts: the *authenticity*, the *intention*, and the *origin*. From the field of forensic psychology, several theories imply that expressed falsehood differs from factual content by writing style [64], textual features like word count [35], the specificity towards events [22], and the sentiments that are expressed [76]. While these theories relate to deceptive statements and not fake news directly, they are closely interlinked concepts.

To define fake news, the most intuitive, though broad definition is simply the following:

Definition 3.2.1. Fake news is false news.

As this is too general to be useful in most cases (since the term ”false” is highly subjective), several authors employ a more narrow definition of fake news [75]:

Definition 3.2.2. Fake news is *intentionally published*, contains *non-factual information*, and is published by a *news outlet*.

Research on fake news detection models has increased dramatically in recent years [38] and many different approaches have been proposed, ranging from relatively simple models like support vector machines to complex deep learning models. Khan *et al.* [27] performed a benchmarking study examining the relative performance of traditional machine learning approaches, deep learning approaches (with GloVe embeddings) and advanced pre-trained language models (BERT, ELMo, etc.). Their results found that pre-trained language models achieve superior results compared to other deep learning methods, while the performance of traditional machine learning is inferior to both. One notable exception is the naive Bayes classifier, which attains comparable results to deep learning approaches. A recent trend in fake news detection models is the adoption of information fusion, such as social contexts, visual information, or fact-checking, making them examples of hybrid systems (section 2.6). One recent example is KAHAN, which combines social context and external knowledge with news content [62], achieving state-of-the-art results on the FakeNewsNet dataset. Another recent trend in fake news detection is graph neural networks, which shows promise as a deep-learning approach. The motivation behind using graph-based architectures is to better exploit information propagation patterns - essential for fake news detection methods that leverage the social context of fake news [43]. The model used by Lotfi *et al.* [31] for detecting rumours in Twitter conversations is an example of a model which incorporates graph-convolutional networks in its architecture. Rumour detection is related to fake news detection, but considers only the spread of a statement on social media without public news content. The model demonstrates superior performance compared to other rumour-detection approaches, as well as having the ability for early detection of rumours. Another recent model proposes to apply an AIS-based model for the detection of fake news [60]. Intuitively, the niche of fake news detection fits the domain of AIS. Just as the biological immune system is able to recognise and attack foreign pathogens in the body, AISs can be trained to identify patterns which are not typical of real news. Moreover, since AISs are adaptive and can learn from new data over time, they are able to continuously improve their accuracy and effectiveness - well-suited to the dynamic environment of news.

3.2.1 General Textual Features

The general textual features in this context refer to the non-latent features that are extracted from a piece of news, e.g. term counts and grammatical quality. The most salient features often differ from dataset to dataset and striking the right balance for real-world fake news classification is a difficult task.

In a recent benchmarking study by Gravanis *et al.* [17] three articles were compared based on feature sets across multiple datasets. Their results indicate that a combination of 57 general textual features from all three sets, coupled with word embedding features, yielded the best results. Moreover, the study also revealed that the importance of various features varied across the datasets, which was influenced by the properties of each set. For instance, the Kaggle set, which only contains true articles from a single source (Reuters) was able to achieve accurate classification using a single feature (typos). In contrast, the sets that incorporated a more diverse range of sources necessitated a greater number of features to discern language, style and intention [17].

In their research, Rashkin *et al.* [46] investigated the language used in fake news articles and the features that were most common compared to real news. Their findings suggest that the strongest predictor for fake news is swear words, which appear seven times more often than in real news. Second-person pronouns are also a strong predictor, being 6.73 times more prevalent in fake news articles. Additionally, specificity was identified as an indicator of real news, consistent with the theory by Johnson *et al.* [22]. Horne *et al.* [20] looked at the language characteristics of fake news articles, finding that headline content is a strong differentiating factor. Fake news headlines were generally longer, with simpler language and more capitalised words and proper nouns. As proper nouns correlate to a certain extent with specificity, this may seem at odds with the findings of Rashkin *et al.* However, it is important to note that this finding extends only to the title, and not to the body of the article.

Sverdrup-Thygeson [60] conducted a study on feature relevance for fake news classification with an AIS model, selecting 19 general textual features. Of these, 17 were chosen based on existing research, with two novel features added: quotation mark frequency and divisive topics. The results of the study show that the novel features performed well, ranking among the top 5 in terms of feature importance. However, the accuracy testing on the proposed AIS model only used 9 of these features.

3.2.2 Embeddings

Numerous fake news detection methods currently in use incorporate embeddings in some form, whether context-independent models such as word2vec or GloVe, or context-dependent ones such as ELMO or BERT. However, when compared to word2vec and GloVe, context-aware embeddings have the drawback of having relatively higher dimensionality, as well as being more computationally expensive to produce [42]. To address the high dimensionality issue, one possible solution is to apply dimensionality reduction techniques. For instance, Raunak *et al.* [47] suggested using principal component analysis along with a post-processing algorithm for dimensionality reduction of word2vec embeddings. Their results indicate that the reduced embeddings produce comparable or superior outcomes across various natural language processing tasks. Additionally, Su *et al.* [59] have proposed a whitening method that enhances the performance of BERT sentence embeddings on semantic textual similarity tasks, while also providing dimensionality reduction advantages (section 2.5.4). The benefit of the method is that it is computationally simpler than comparable methods to address the anisotropy in the

BERT sentence embedding space.

The model by Gravanis *et al.* [17] uses word2vec to produce an embedding vector, finding that accuracy is better when using the vector alone compared to only textual features, achieving accuracies of 86.1% and 93.7% respectively. Their model is further enhanced by combining textual features and embeddings, with an accuracy of 94.9%. These accuracies were averages of several datasets, indicating that embeddings are beneficial for fake news classification.

The FakeBERT model introduced by Kaliyar *et al.* [25] investigates the use of both BERT and GloVe embeddings, in combination with LSTM (long short-term memory) and CNN (convolutional neural network) models for classification. The findings of their study indicate that the CNN-based architecture outperforms the LSTM architecture. Furthermore, both BERT and GloVe embeddings yielded similar levels of accuracy. Verma *et al.* [66] also proposed a model, MCred, that incorporates embeddings and a CNN. In this approach, both GloVe and BERT-based embeddings are employed simultaneously. The results of their study show that MCred achieves comparable performance to FakeBERT on the Kaggle dataset. These findings suggest that combining different types of embeddings may not lead to improved results. It should be noted that both MCred and FakeBERT are tested on the Kaggle set, which has some characteristics which can cause the main advantage of BERT embeddings (contextual information) to be unnecessary for accurate classification, discussed in section 3.2.3.

Another recent model [45] uses BERT embeddings along with an LSTM architecture, with performance testing done on the FakeNewsNet dataset (section 3.2.3) using headlines only. The model achieves accuracies of 88.75% and 84.10% on the Politifact and GossipCop portions respectively. The results indicate that headline information alone may be enough to detect fake news accurately.

Sverdrup-Thygeson's [60] study explored adopting BERT embeddings for an AIS model, showing improved results on one dataset (Kaggle) compared to only using general textual features. However, the model's performance diminished on two other datasets (LIAR and FakeNewsNet), possibly due to issues with the versions of the FakeNewsNet and LIAR sets not being fully representative of fake news. The variants of the sets employed for testing only include a fact-checking statement and not the article text the statement refers to. Furthermore, the model treats embedding features the same way as general textual features, thus implicitly placing higher importance on them due to their larger dimensionality. The output of the base BERT model produces a 768-length vector so the resulting shape-space would be 768-dimensional. Therefore, since the model mutates values of the feature vector with the same probability for each value, it is far more likely that it will mutate a value belonging to an embedding feature.

Interestingly, the model performs better when using *head* + *tail* embeddings (the first and last sentences of the article) compared to the headlines, in contrast to the findings of Horne *et al.* [20] which indicate that headlines are generally a good indicator of falsehood [20]. This may also be due to the aforementioned dataset issues, as the first line of the fact-checking statement often includes the verdict. The study does not look at separating head from tail embeddings.

3.2.3 Relevant Datasets

Creating datasets for fake news is a difficult task. What constitutes fake news is dynamic; a fake statement at one point in time may be true at another. Therefore, there is a shortage of high-quality datasets relevant to the proposed model. Many authors use sets based on expert knowledge from fact-checking sites like Politifact or Snopes [62] [12] [56]. Others construct their own sets, making the comparison more difficult as these often are not published due to copyright issues stemming from the articles or the social context [17].

Kaggle

The Kaggle dataset is a large-scale dataset posted initially on the data science competition site kaggle.com. The dataset is a binary classification of reliable/unreliable news sources from text news content. An issue with this set is that it only includes articles classed as true from one source, Reuters. The challenge is that a classifier trained on the set will likely be equipped to identify Reuters' editorial style more than any other. The benchmark study by Gravanis *et al.* discusses this dataset compared to a similar dataset, containing the same set of fake articles but with a greater diversity of sources for real articles (the McIntire set). The classifier trained on the Kaggle set produced noticeably poorer results than the one trained on the McIntire set, indicating that diversity in sources is necessary for an effective classifier [17]. Another issue with the set is that every real article begins with the word "REUTERS-", further increasing the likelihood of any classifier trained on the set effectively converging to a source-based classifier for Reuters.

FakeNewsNet

The FakeNewsNet set [54] nominally includes 1,056 news articles with labels from Politifact and 22,865 articles with labels sourced from the gossip fact-checking site GossipCop. It also contains 602,659 Twitter statements related to the articles for use in classifiers leveraging the social context. This dataset has a greater diversity of sources for both true and false news compared to the Kaggle set, which enables a classifier to generalise better and avoids creating a simple source-based classifier.

An issue with the set is that it is not served directly; the user must fetch the articles from a set of links. Consequently, each researcher using this set might have different data due to, e.g. sites being taken down or being taken over by others since the set was initially made available. This effectively reduces the size of the corpus, as well as rendering comparison to other models less accurate.

BuzzFeedNews

The BuzzFeedNews dataset [44] comprises articles sourced from nine distinct outlets, three each from hyperpartisan left- and right-wing publications, as well as three mainstream media outlets. The set includes 2282 articles, each of them manually checked for factual content by five BuzzFeed journalists and subsequently assigned to one of four classes. This dataset is highly imbalanced, with one class representing approximately 80% of the samples.

LIAR

The LIAR dataset [68] consists of 12,836 short political statements from American politicians . The data set has six classes according to the degree of truthfulness, ranging from ”true” to ”pants on fire”. The classes for each statement are based on fact-checks performed by Politifact. Due to the brevity of the samples and the fact that this set requires six-class classification, it is generally considered one of the more complex sets to classify accurately.

4 Model

This chapter presents the different components of the proposed model. First, the parts of the core AIS model will be presented. Subsequently in section 4.2, the data pre-processing pipeline will be detailed, as well as the features chosen for defining a news article as an antigen. The code repository containing the model can be found at:
https://github.com/Tryxel-Industries/ais.

4.1 Flexible-Dimension-AIS

This section describes the different aspects of the proposed Flexible-Dimension-AIS (FD-AIS) algorithm. First, the high-level structure of the model will be presented, and then the different components of the model will be discussed in-depth.

4.1.1 Model Overview

The general functioning of the model is shown in fig. 4.1. First, pre-processing of the data is performed through a pipeline described in section 4.2. Thereafter, the initial training conditions are set up, defined by the parameters in table 4.1. The model has two general modes of operation, depending on whether or not boosting is enabled. If boosting is not enabled, the algorithm runs through the core training loop once. The model first initialises the population, the particulars of which are described in section 4.1.4. The initial population is then evaluated and scored. Evaluation differs from scoring in that evaluation finds which antigens are detected by each antibody while scoring uses the evaluation information to give the antibodies a fitness score, described in section 4.1.5. After the population has been initialised the model enters the main training loop, which is run for g_n generations.

The first step in the loop determines the number of antibodies to replace, which is governed by the parameter AB_{rr}. This parameter undergoes a linear decrease throughout the training, encouraging exploration during the initial generations and exploitation in the latter. The replacement procedure is described in section 4.1.8. Thereafter, parents are selected and subsequently cloned. The offspring are then subjected to mutation, described in section 4.1.7. The resulting offspring are then evaluated and scored. The algorithm can also be run with the option to inject additional diversity through ”leaking”, which inserts randomly instantiated antibodies into the population. Leaking is further explained in section 4.1.9. The population may be replaced with the next generation, and the loop continues. After the training has run for g_n generations, the final antibodies are stored and training is complete.

If boosting is enabled, the algorithm incorporates a few additional steps. The initial weights for each antigen are computed, along with the size of each boosting round. Each round of boosting executes the basic training scheme described above and recalculates the weights based on the results. The boosting method employed is described in more detail in section 4.1.11.

4.1.2 Model Parameters

Table 4.1 shows the hyperparameters of the algorithm. The AIS is highly flexible and has many parameters which can be tuned for the task at hand.

Table 4.1: Parameters for the proposed model

Model Parameters	
Parameter	**Description**
E_m	Model evaluation method
B_n	Rounds of boosting
AB_p	Antibody population size as a fraction of antigen set
AB_{rr}	Antibody replacement ratio
L_F	Leak fraction
L_R	Fraction of the leaked population which is randomly instantiated
g_n	Number of generations to run the training
α	Weight of the Correctness part of the fitness score
β	Weight of the Coverage part of the fitness score
γ	Weight of the Uniqueness part of the fitness score
ϵ	Weight of the Valid Avidity part of the fitness score
ζ	Weight of the Invalid Avidity part of the fitness score
M_{ow}	Mutation offset weight
M_{mw}	Mutation multiplier weight
M_{mlsw}	Mutation multiplier local search weight
M_{rw}	Mutation radius weight
M_{vtw}	Mutation value type weight
M_{lw}	Mutation label weight
N_c	Maximum number of clones produced by a parent

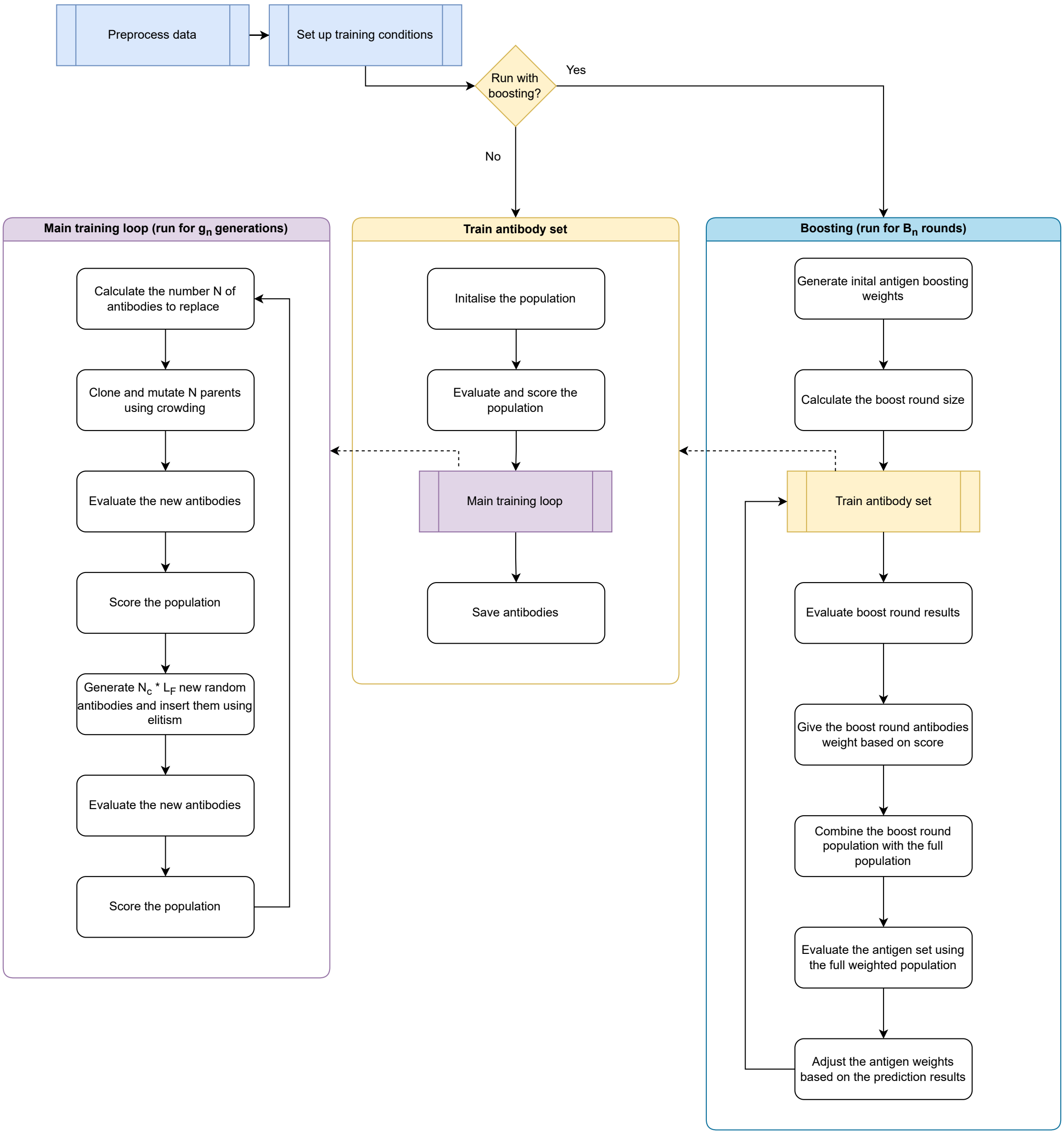

Preprocess data
Set up training conditions
Run with boosting?
Yes
No
Main training loop (run for g_n generations)
Calculate the number N of antibodies to replace
Clone and mutate N parents using crowding
Evaluate the new antibodies
Score the population
Generate N_c * L_F new random antibodies and insert them using elitism
Evaluate the new antibodies
Score the population
Train antibody set
Initalise the population
Evaluate and score the population
Main training loop
Save antibodies
Boosting (run for B_n rounds)
Generate inital antigen boosting weights
Calculate the boost round size
Train antibody set
Evaluate boost round results
Give the boost round antibodies weight based on score
Combine the boost round population with the full population
Evaluate the antigen set using the full weighted population
Adjust the antigen weights based on the prediction results

Figure 4.1: Model diagram

4.1.3 Antibody and Antigen Structure

The antigens each contain a vector that defines the feature values for the antigen. For the antibodies, each contains a vector representing its centre in the shape-space, as well as one containing the radii of the antibodies' RR. In addition, the antibodies have a vector of multipliers which is used for scaling the dimensions along each feature axis. This scaling capability allows antibodies to mutate from n-spheres to ellipsoids.

The proposed model introduces a novel element: the dimension-type vector. This addition enables antibodies to individually mutate their recognition regions (RR) for each feature axis into one of three categories: circular, open, or disabled.

To test whether an antigen lies within the RR of an antibody, eq. (4.1) is employed. Here, d is the dimensionality of the shape-space, ag_v are the feature values of the antigen which is being tested against, ab_o is the offset vector of the antibody, ab_m is the vector of multipliers for each dimension of the antibody, ab_e is the vector of exponents for each antibody dimension, which determines the dimension type, and ab_t is the antibody threshold value. If $c(ab, ag) < 0$, the antigen lies within the antibody's RR.

$$c(ab, ag) = (\sum_{n=1}^{d}((ag_v^{(n)} + ab_o^{(n)}) * ab_m^{(n)})^{ab_e^{(n)}}) - ab_t \tag{4.1}$$

Circular:

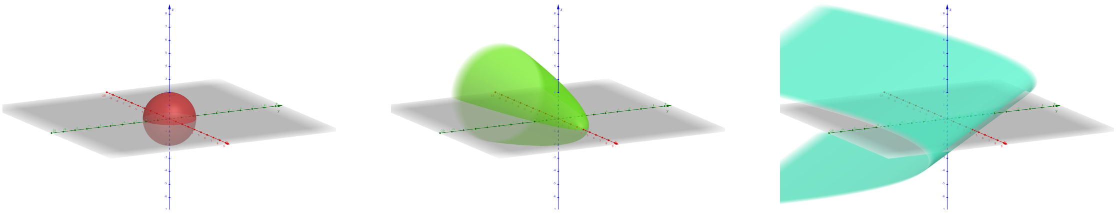
Antibodies operating with this type of dimension use a closed RR. The name circular is a slight misnomer, as these dimensions have the capability of mutating into ellipsoids through the multiplier, inspired by Ellipsoidal-AIS [39]. However, in contrast to Ellipsoidal-AIS, the proposed model does not allow for rotation of the RRs, as preliminary testing showed the exponential increase in space- and time-complexity to not justify the potential increase in accuracy.

Open:

By setting an element in the ab_e vector to 1 the corresponding dimension of the antibody "opens", allowing it to expand unbounded in one direction. How the recognition region changes when dimensions are opened is shown in fig. 4.2. If all dimensions are open a hyperplane will form. The core motivation behind the open dimension type is that it will provide a greater degree of coverage of high-dimensional shape-spaces (such as the ones produced by embedding models).

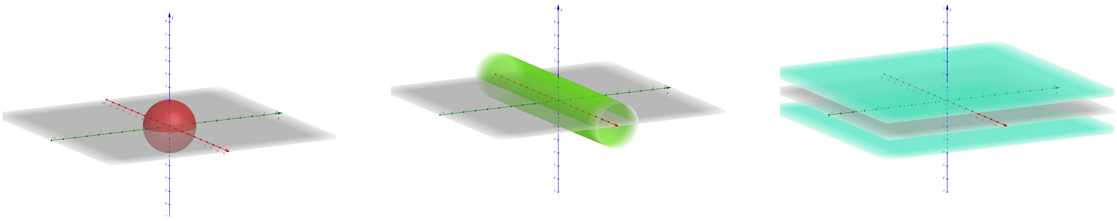
Disabled:

The last dimension type is inspired by the feature subset selection employed in AISLFS [11] and CLONALG-LFS [69] which have demonstrated the benefit of disabled dimensions. By disabling dimensions, the model gains the ability to perform local feature selection by reducing the local shape-space for the antibody to one containing a subset of the features. How this affects the recognition region is shown in fig. 4.3.



(a) Zero open dimensions (b) One open dimension (c) Two open dimensions

Figure 4.2: Antibody RR with zero, one and two dimensions open in a 3d space



(a) Zero disabled dimensions (b) One disabled dimension (c) Two disabled dimensions

Figure 4.3: Antibody RR with zero, one and two dimensions disabled in a 3d space

4.1.4 Initialisation

The proposed model utilises a combination of random initialisation and antigen-based initialisation. To initialise an antibody based on an antigen, a weighted selection scheme is employed. The probability of selecting an antigen is inversely proportional to the number of antibodies that already cover the antigen. The antibody boosting relevance value is also considered when boosting is enabled. For random and antigen-based initialisation, the algorithm can expand the antibody radius to the maximum size where no antigens of a different class are covered. This approach is inspired by the cross-reactivity threshold initialisation method employed in AISLFS [11]. The size of the population is governed by the parameter AB_p, which is either set manually or as a fraction of the size of the antigen set. The initial dimension types for the antibodies are chosen at random with a uniform probability distribution, however, this can be adjusted through parameters.

4.1.5 Affinity and Fitness Calculation

Antibodies with open dimensions do not have a geometric centre. Therefore, the algorithm employs affinity which does not require a centre. The algorithm utilises two affinity functions which are used as components in antibody fitness. The first is binary affinity, (eq. (4.2)), which is 1 if the antigen is covered by the antibody, otherwise 0. This is similar to the approach used by VALIS [26] and Ellipsoidal-AIS [39]. The second affinity function takes the result of eq. (4.1) and feeds it through a limiting function, described in eq. (4.8). This is a novel approach for use in an affinity function and has the purpose of clamping the value to a specified interval. This affinity function is designed to better optimise open dimensions.

$$Aff(Ab_i, Ag_j) = \begin{cases} 1, & \text{if } c(ab_i, ag_j) \leq 0 \\ 0, & \text{otherwise} \end{cases} \tag{4.2}$$

Additionally, the model employs a five-fold weighted fitness calculations scheme, where each part of the fitness formula attempts to emphasise some wanted property of the final antigens. The fitness score is calculated using Equation 4.3, where α, β, γ, ϵ, and ζ are the parameter weights for the different fitness components. How the Correctness, Coverage, Uniqueness, Avidity-valid, and Avidity-invalid values are calculated is explained below

$$Fitness = \alpha * Correctness + \beta * Coverage + \gamma * Uniqueness + \epsilon * Av_g + \zeta * Av_b \tag{4.3}$$

Correctness

The correctness of the fitness score is shown in Equation 4.4, where TP is true positives, FP is false positives, and η is the scaling factor for the error. Equation 4.4 is mostly the same as the one used by Ozsen *et al.* [39] except for the error scaling factor η, which is added to more strongly penalise errors. The correctness score is a number between $-\eta$ and 1 and favours antibodies with a high level of precision. The default value of η used is 2 since it proved effective in preliminary testing.

$$Correctness = \frac{TP - (FP * \eta)}{FP + TP} \tag{4.4}$$

Coverage

The coverage of the antigens is calculated according to eq. (4.5), where TP is true positives, and $Positives$ is the total number of the same class in the population. This is also known as recall. The coverage score is a value between 0 and 1. This part of the fitness value favours large antibodies over smaller ones to increase generalisation ability.

$$Coverage = \frac{TP}{Positives} \tag{4.5}$$

Uniqueness

The uniqueness of the antibodies is calculated according to Equation 4.7. The shared affinity value is calculated according to eq. (4.6). For each antigen, the number of antibodies that cover it is known as ”shared count” (SC). For each antibody, the shared affinity (SA) is the sum of the inverse of the sharing count of each antigen covered by the antibody. The uniqueness score of an antibody is then equal to the ratio of shared affinity to true positives (TP), the number of antigens correctly classified by the antibody. The uniqueness score is a value between 0 and 1 and favours antibodies covering antigens which are not detected by many others to improve coverage of the shape-space. This approach is analogous to the concept of fitness sharing in evolutionary computing [14].

$$SA = \sum_{matched} \frac{1}{SC} \tag{4.6}$$

$$Uniqueness = \frac{SA}{TP} \tag{4.7}$$

Valid and Invalid Avidity

To more appropriately value antibodies with open dimensions, a heuristic affinity function was employed. For an antigen, the result of eq. (4.1) is passed through the modified sigmoid function in eq. (4.8). The sigmoid function is employed to avoid the issue of the value being unbounded for open intervals by clamping the values to a limited interval.

$$\sigma(x) = 1 - \frac{1}{1 + e^{x}} \tag{4.8}$$

After determining the affinity heuristic, both Avidity-Invalid and Avidity-Valid values are calculated. Avidity in biological immune systems is a measure of the binding strength over all binding sites on an antibody (section 2.2). Correspondingly, the avidity functions employed are the mean of the heuristic affinity function eq. (4.8). Avidity-Valid is the mean of each correctly classified antigen by the antibody, while Avidity-Invalid is the mean of each incorrectly classified antigen by the antibody. The motivation behind these parts of the fitness function is to appropriately adjust open intervals.

4.1.6 Parent Selection

To select antibodies for cloning and mutation, a random selection method was adopted. The reasoning behind not using tournament selection as in [60] or elitism as in [26] was that initial testing showed tournament selection led to a decrease in diversity that negatively impacted performance. Similarly, elitism would lead to even lower diversity. Instead, random selection was deemed suitable given the effective selection pressure in other areas of the model. Each antibody selected for reproduction produces m clones proportional to the parent fitness. m calculated for the i-th antibody according to eq. (4.9). The clones are subsequently subjected to mutation and matched with antigens.

$$m^{(i)} = round(max(F^{(i)}_{max}, 0.2) * N_c) \tag{4.9}$$

where F_{max} is the ratio of the antibody's fitness to the maximum antibody fitness in the population. To ensure that a selected parent generates clones, the maximum of 0.2 and F_{max} is chosen as a scaling coefficient for the N_c parameter.

4.1.7 Mutation Operators

Each offspring of an antibody is subject to a single mutation, with the mutation type chosen through a weighted random pick using varying weights defined in the model parameters. The algorithm employs five types of mutation operations, with all but the local search, and dimension type mutations being scaled according to the antibody fitness. This scaling principle is based on the idea that antibodies demonstrating higher fitness necessitate finer adjustments compared to their less fit counterparts, inspired by the somatic hypermutation mechanism in CLONALG [8]. Notably, crossover was not employed in the model, as the literature review did not uncover any instances where it demonstrated a benefit for classification with an AIS.

Shift mutation

The shift mutation shifts the centre of the antibody along one random axis by adding or subtracting an offset scaled by the fitness of the antibody. This effectively translates the entire antibody in the shape-space.

Multiplier mutation

Scales the multiplier for one random dimension. For ellipses, this will shorten or lengthen an axis and for open dimensions, this will modify the shape by narrowing or widening the opening.

Dimension type mutation

This mutation changes the dimension type of one random feature of the antibody to a different one listed in 4.1.3. When the dimension-type changes for a dimension the new dimension is unlikely to have a good fit for the recognition region in that dimension. To counteract this and shift the RR front in that dimension towards a more optimal position, it is possible to perform a local search of the multiplier of the resulting dimension if enabled in the model.

Threshold mutation

Scales the detection threshold in eq. (4.1), used to decide whether or not an antibody detects an antigen. For an antibody with only circular dimensions this amounts to increasing the radius. For an antibody with only open dimensions this amounts to moving the hyperplane in the positive direction along all axes. For antibodies with a combination of dimension types, the RR will move in the direction of the sum of the axes with open dimensions which will in turn increase the radius of all circular dimensions.

Local-search scale mutation

This mutation type iteratively searches for the multiplier which maximises R_{ce}, the ratio of correctly classified antigens to the total number of classified antigens for a RR (eq. (4.10)). The motivation behind this is to improve the generalisation ability of the model by avoiding the creation of small RRs only suitable for the training set. Initially, a method inspired by the cross-reactivity threshold applied in the initialisation scheme of AISLFS [11] was employed. However, preliminary testing showed that the alternative approach produced better results across several reference datasets. This is likely because simply minimising the number of errors will quickly move the antigens into a local optimum. Further exploration of different local search techniques is beyond the scope of this thesis but could be an interesting area for future research.

$$R_{ce} = \frac{\ln(TP+2)*1.5}{\ln(TP+2)*1.5+\ln(FP+2)} \tag{4.10}$$

4.1.8 Replacement

The model incorporates a crowding operator (section 2.1.1) to manage the replacement of parent antibodies by the cloned offspring. All offspring and the parent compete for a single survivor spot in the population for the next generation. The percentage of the population to be replaced is determined based on a linear decrease determined from the AB_{rr} parameter.

Both VALIS and the model by Sverdrup-Thygeson [60] display a certain degree of instability, possibly through the high replacement ratios employed. To mitigate this to some extent in the proposed model, elitism and crowding are employed. Since replaced antibodies must belong to the same class and yield a better score in order to survive the proposed model should display improved stability.

4.1.9 Leaking

In order to promote diversity within the population, the model employs a technique whereby a random subpopulation of antibodies is generated and integrated into the main population via elitism. This approach enables the algorithm to explore antibody centres outside densely populated areas of the shape-space, thus increasing the likelihood of antigens that span multiple distinct clusters of antigens simultaneously. The leaked antibodies also have a chance to be initialised with centres at the antigen values. The antigens to initialise from are selected using the process described in section 4.1.4. Furthermore, the process of leaking, which is implemented alongside local search and open intervals, facilitates greater coverage of the shape-space through the instantiation of random antibodies. The L_F parameter is used to determine the number of antibodies that are leaked each generation, and this parameter also controls the size of the additional leaked population that is merged with the replaced population. The L_R parameter is utilised to decide the ratio of antibodies that are instantiated randomly versus those based on antigens. It is important to note that the leaked population is generated and selected independently of the antibodies replaced through normal clonal selection each generation. The approach is widely used in genetic algorithms to explicitly increase diversity [14], and is also employed in CLONALG-LFS [69].

4.1.10 Ratio Locking

To avoid excessive antibody drift between classes, the algorithm uses ratio locking, which involves preserving the antibody population's original ratios of the classes. The rationale behind employing ratio locking is to avoid the tendency for the antibody population to gravitate towards the predominant class in imbalanced datasets. The technique was found to be somewhat effective in model refinement testing (exp. MR2).

4.1.11 Boosting

Since an AIS can be considered a set of weak predictors (the antibodies), it was theorised that implementing a boosting scheme could be beneficial. Boosting is a novel idea for an AIS and the proposed model employs a slightly modified version of Adaboost section 2.3.3. The modifications are two-fold, the first difference is not counting undetected antigens as errors to better handle challenges associated with large antigen populations, thereby making the method more scalable. The other change is employing a retry system for rerunning boosting rounds that have an accuracy below 50% and thereby ensuring that weights are correctly updated.

4.1.12 Class Assignment

Two different class assignment schemes were tested: count assignment and fractional-sum assignment. The fractional-sum assignment scheme was inspired by the vote-allocating procedure employed in VALIS [11].

Count assignment

When operating with count assignment, the model only considers binary antibody affinity, counts the number of registered antibodies of each class, and picks the class with the most registered antibodies.

Fractional-sum assignment

With fractional-sum assignment, the final training accuracy of the antibody is considered when performing class assignments. When the algorithm classifies an antigen, it sums up the accuracy of each antibody which detects the antigen and picks the class with the highest accuracy sum. As an example, if an antigen is covered by two antibodies with 90% accuracy in class 1 and four antibodies with 40% accuracy in class 2. The algorithm will classify the antigen as class 1

4.2 Data Pre-Processing and Feature Generation

To make the data digestible for the AIS model, multiple pre-processing steps were needed. The pre-processing pipeline for the embeddings and the semantic features differed in multiple aspects - this section outlines the differences and the commonalities. The general functioning of the pipeline is shown in fig. 4.4.

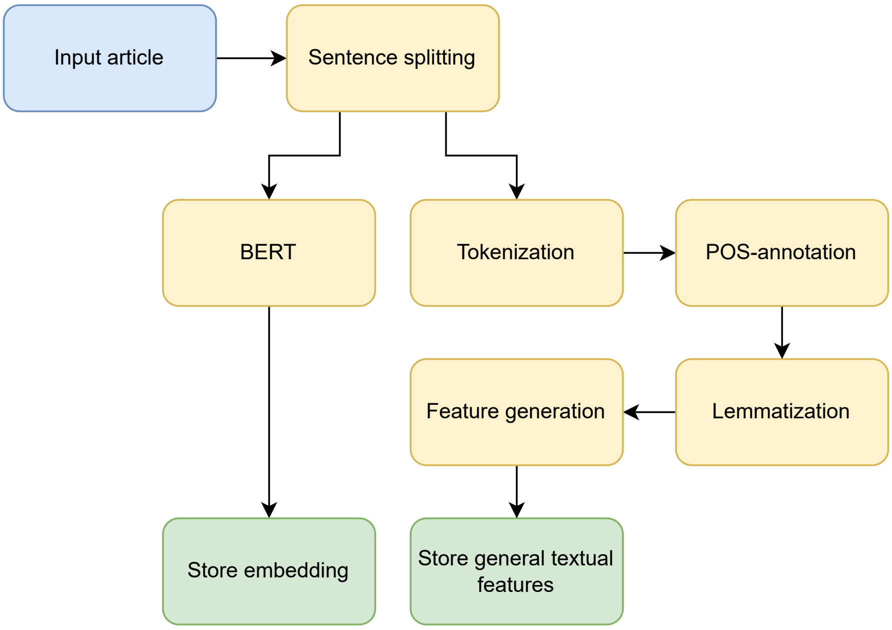


Figure 4.4: The pre-processing pipeline

4.2.1 Pre-processing overview

For a given news article, it is passed through several steps before it is suitable for use in the algorithm. First, the sentences are split. This is necessary for both the embeddings and general textual features. To generate embeddings, the split sentences are fed through a pre-trained BERT model and the resulting outputs are then stored. Generation of the general textual features necessitates some additional steps. The split sentences are tokenised, POS-annotated and lemmatised. This was all performed through the Stanford CoreNLP toolkit [33]. Subsequently, the chosen features were generated and stored.

4.2.2 Normalisation

To avoid the predicament of unequal feature scaling in feature values, input values were normalised between 0 and 1, which is a common approach for AIS models [26] [8]. The normalisation was performed according to eq. (4.11):

$$\tilde{x} = \frac{x - min(x)}{max(x) - min(x)} \tag{4.11}$$

where $min(x)$ and $max(x)$ are the lowest and highest values for the feature respectively. Values falling outside the range found in training were clipped after normalisation to the range [0, 1]. Embedding values were not normalised, as this would remove some of the semantic information carried in the embedding vector.

4.2.3 General textual features

For the general textual features several from the study by Sverdrup-Thygeson [60] were adopted, as they have demonstrated effectiveness for fake news classification with an AIS. This section will outline what modifications were done to those features, if deemed necessary, as well as any new features implemented.

Many of the general textual features used in the model are based on frequency counting (section 2.5.2). The term-frequency features were primarily derived from the ones identified by Sverdrup-Thygeson [60], with adaptations made to the lexicon used for certain features following an evaluation that identified some potential issues. First- and second-person pronouns, strongly subjective words and action adverbs displayed satisfactory performance and employed appropriate lexicons. Accordingly, these features were adopted for use in the proposed model. Similarly, numbers, exclamation and question marks, and word count were also included as features. Additionally, the model adopted Sverdrup-Thygeson's novel features of quotation mark frequency and divisive topics, which also performed well. However, while swear words are a good indicator of fake news [46], this feature did not perform well in Sverdrup-Thygeson's study. The low performance may be due to the lexicon being too general, including for example words such as ”women's”, ”UK”, ”sick” and ”fire” as swear words. Therefore, the proposed model adopted swear words as a feature while employing a more narrowed lexicon.

However, Sverdrup-Thygeson's study also revealed that certain features did not enhance the model's performance. Manner adverbs, modal adverbs, and negative- and positive opinion words were either ineffective or had a negative impact on the model's performance and were accordingly not used.

Some novel features were also included. One such feature is the number of capitalised words in the text, which was inspired by Horne *et al.*'s [20] study on headline content in fake news articles. The hypothesis was that the presence of capitalised words could extend beyond the headlines and manifest in the body of the articles as well. Moreover, intensifier adverbs, which strengthen the meaning of other expressions and indicate emphasis, have been included as a novel feature (e.g. ”absolutely”, ”absurdly”, ”extremely”, ”completely”). This is based on the assumption that such phrases may offer insight into the sentiment conveyed by an article. The number of words and phrases that convey emphasis has also been integrated as a feature, as it may suggest that the authors are attempting to elicit an emotional response from the reader, a hallmark of fake news [20]. Lastly, the number of words and phrases that display generalisation has been adapted to serve as a feature. The motivation behind this is that fake news has a lower specificity than real news [22], and should therefore display an increased tendency toward generalisation. The lexicons for the emphasis, intensifier adverb and generalisation features were derived from the MPQA arguing lexicon [57].

The Flesch-Kincaid grade level feature performed well in Sverdrup-Thygeson's study and is consequently also used in the proposed model.

Table 4.2: General textual features adopted for use in the model

Feature #	**Description**
1	Swear words
2	First-person pronouns
3	Second-person pronouns
4	Action adverbs
5	Superlative forms
6	Comparative forms
7	Strongly subjective words
8	Negations
9	Negative opinion words
10	Positive opinion words
11	Numbers
12	Exclamation and question marks
13	Quotation marks
14	Word count
15	Divisive topics
16	Effect word sum
17	Capitalised words
18	Intensifier adverbs
19	Emphasis
20	Rhetorical questions
21	Generalisation
22	Inconsistency
23	Conditionals
24	Necessity

4.2.4 Embeddings

Sentence embeddings are employed as a feature for the embedding part of the proposed model, as the results produced by FakeBERT [25] and the model in [17] have demonstrated their utility for the task of fake news classification. Additionally, embeddings were used in the model by Sverdrup-Thygeson [60], with interesting results. While there were issues with some of the datasets, as noted in section 3.2, the best-performing feature set included a sentence embedding for the headline, indicating that embeddings are viable for use in an AIS model.

While models based on both context-independent [17] and context-dependent [25] embeddings have shown promising results, BERT was selected as the embedding model over word2vec or GloVe. BERT has specific attributes that are more relevant to the proposed model, such as its ability to create embeddings over entire sentences, not just words, which reduces the number of antigens necessary to store. Furthermore, it was theorised that contextual information present in BERT embeddings would be beneficial to the proposed model. To counteract the high dimensionality of the BERT output embeddings, the base version of BERT was selected which has an output dimensionality of 768, compared to 1024 in Bert-large. Moreover, the uncased version was chosen due to its widespread adoption for fake news classification [60] [25]. For the embedding process, the articles with split sentences were fed into a BERT transformer network to create the sentence embeddings.

The use of embeddings in the proposed algorithm was done by treating sentences from a news article as separate antigens, with the class label of the original news article used as the class label of each antigen.

Whitening of embeddings and dimensionality reduction

To mitigate the challenges that come with high dimensionality, the model implements whitening of sentence embeddings and dimensionality reduction through the approach outlined in section 2.5.4.

As whitening of sentence embeddings has shown promising results on semantic similarity tasks [59], it was hypothesised that this may transfer to the semantic classification in the proposed model, providing a benefit in addition to dimensionality reduction.

5 Experiments and Results

The following chapter presents and discusses the experiments performed using the algorithm presented in chapter 4. First, the experimental plan is outlined in section 5.1, which briefly touches on the experiments performed, divided into three distinct phases. Following this, section 5.2 presents the datasets used for testing as well as the specifications of the experimental machine. Sections 5.3 to 5.5 describe the experiments performed to determine the workings and performance of the model. Each experimental phase is bookended by a brief introduction and a discussion summarising the key findings of the phase.

5.1 Experimental Plan

The goal of the experiment plan was to gain a deeper understanding of the workings of the model and to answer the research questions presented in section 1.2. The experiment plan was split into three phases, where phases 1 and 2 deal with the model while phase 3 considers the application.

Experiment Phase 1: Model refinement

Exp. MR1: Investigate the effect of locking the ratios of antibodies of different classes to mitigate the impact of imbalanced datasets

Exp. MR2: Investigate the effect of crowding to preserve diversity in the population

Exp. MR3: Investigate the effect of boosting to improve performance

Exp. MR4: Investigate the effect of inserting random individuals through leaking to promote diversity

The purpose of this phase was to investigate the applicability of various methods to mitigate challenges that appeared through preliminary model implementation. Exp. MR1 examined the effect of the proposed ratio-locking operation, intended to aid the model in maintaining diversity in class label distribution and avoid unwanted drift towards majority-class classification in imbalanced datasets. To further maintain diversity through the evolutionary procedure, a crowding operator was implemented, the effect of which was investigated in exp. MR2. Following this, exp. MR3 was performed to determine whether a boosting method could be appropriate for the proposed model to improve performance. Finally, the effect of leaking (section 4.1.9) was examined in exp. MR4.

Experiment Phase 2: Explore impact of recognition regions

Exp. RR1: Investigate the impact of allowing ellipsoidal RRs by varying the multiplier

Exp. RR2: Investigate the effect of local search for multipliers

Exp. RR3: Investigate the impact of ”open” dimensions for RRs

Exp. RR4: Examine effect of adjusting AB_p with open dimensions

The second experiment phase explored different definitions of the RR of the antibodies. First, the impact of the multiplier was examined, with exp. RR1 looking at performance with and without an evolved multiplier. Then, local search of the multiplier was examined in exp. RR2. The intention here was to determine whether an operator to find multipliers not easily reachable through mutation could improve the model. Exp. RR3 looked at the effect of the different dimension types detailed in section 4.1.3, by employing three models, each with one dimension type removed. Finally, the effect of adjusting the AB_p parameter for a model without the disabled dimension type was investigated in exp. MR4.

Experiment Phase 3: Embeddings compared to semantic features in model

Exp. ES1: Investigate the effect of whitening and dimensionality reduction in the model employing embeddings

Exp. ES2: Evaluate the performance of the model on fake news detection when run with general textual features

In the final experiment phase, the model resulting from the findings of the first two phases was applied for fake news classification. Exp. ES1 considered the effects of whitening and dimensionality reduction, while exp. ES2 compares the performance of embeddings and general textual features.

5.2 Experimental Setup

To evaluate the model on classification tasks, several benchmark datasets were utilised in initial testing, outlined in table 5.1. The selection of these datasets was based on their attributes and current use by state-of-the-art AIS classifiers. This allowed for a thorough assessment of the model's ability to adapt to different datasets, as well as enabling comparison with alternative algorithms.

For the experiments conducted in Phase 3 (section 5.5), fake news datasets were utilised, as listed in table 5.2. Similar to the benchmark sets used for classification, these were selected to include varying levels of complexity, balance ratio, and class labels. These sets are also commonly employed for benchmarking by other models, which aids to position the proposed model within the context of current state-of-the-art fake news detection techniques.

The specifications of the machine on which the experiments in this chapter were run are listed in table 5.3.

Table 5.1: Benchmark datasets employed for testing in Phase 1 and 2

Dataset	**Samples**	**Classes**	**Balance ratio**	**Features**
Wine [58]	178	3	0.4/0.33/0.27	13
Diabetes [24]	768	2	0.35/0.65	8
Ionosphere [65]	351	2	0.36/0.64	34
Sonar [61]	208	2	0.47/0.53	60
Iris [16]	150	3	0.33/0.33/0.33	4
Glass [16]	214	6	0.33/0.35/0.08/0.06/0.04/0.14	9

Table 5.2: Fake news datasets employed for Phase 3

Dataset	**Samples**	**Classes**	**Balance ratio**
FakeNewsNet (Politifact portion) [54]	838	2	0.38/0.62
Kaggle set	20717	2	0.5/0.5
BuzzFeedNews [44]	1627	4	0.78/0.13/0.05/0.04

Table 5.3: Relevant specifications of machine for experiments

Component	**Description**
OS	Ubuntu 22.04.2 LTS (Jammy Jellyfish)
CPU	AMD Ryzen 9 7950X @ 4.5 GHz
GPU	NVIDIA GeForce RTX 4090 (24 GB VRAM)
RAM	64GB DDR5 @ 5200 MHz

5.3 Experiment Phase 1: Model Refinement (MR)

The model refinement phase sought to evaluate the applicability of some of the proposed approaches in the model to iteratively improve it for classification tasks. Unless otherwise noted, the experiments in this phase were run with the parameters displayed in table 5.4. All experiments in this phase were performed using 10-fold cross-validation for 10 runs, with the average of all runs as the result. The experiments in this phase were also run with open dimensions, local search, and an adaptive multiplier, the effects of which are further examined in the following phase.

Table 5.4: General parameters for the models in Phase 1

Phase 1 Parameters		
Parameter	**Value**	**Description**
E_m	Fraction sum	Model evaluation method
B_n	0	Rounds of boosting
AB_p	1.0	Antibody population size as a fraction of antigen set
AB_{rr}	0.6-0.01	Antibody replacement ratio
L_F	0.0	Leak fraction
L_R	0.0	Fraction of the leaked population randomly instantiated
g_n	300	Number of training generations
α	2.5	Weight of the Correctness part of the fitness score
β	1.0	Weight of the Coverage part of the fitness score
γ	1.2	Weight of the Uniqueness part of the fitness score
ϵ	0.0	Weight of the Valid Avidity part of the fitness score
ζ	1.4	Weight of the Invalid Avidity part of the fitness score
M_{ow}	1	Mutation offset weight
M_{mw}	1	Multiplier mutation weight
M_{mlsw}	1	Multiplier local search mutation weight
M_{rw}	1	Radius mutation weight
M_{vtw}	1	Dimension type mutation weight
M_{lw}	0	Class label mutation weight
N_c	10	Maximum number of clones produced by a parent

Experiment MR1 - Ratio Locking

The purpose of this experiment was to test the effect of ratio locking on the model, particularly on unbalanced sets. During the initial phases of model refinement, a significant tendency was observed where the antibodies drifted toward the majority class. To counterbalance this, it was proposed to maintain the initial label distribution when instantiating the antibodies and to keep this distribution constant throughout training. This experiment also tests the effect of disabling ratio locking. When disabled, the antibodies can mutate their class membership, inspired by immunoglobulin class-switching in the biological immune system (section 2.2).

Hypothesis

The motivation behind ratio locking was that by forcing a constant distribution of antibody class labels, the observed effect of antibody drifting could be stopped. However, since locking the ratios only accounts for the class distribution of the antigens, not their spatial distribution, the effect could be somewhat limited. For example, if a majority class is clustered closely together in the shape-space, it will likely require fewer antibodies to accurately classify than a class which is more sparsely distributed. By disabling ratio-locking and allowing for mutation of class membership, the model might be able to more appropriately account for the spatial distribution of the antigens but could be liable to evolve towards the majority class. Furthermore, adding an extra mutation operator increases the search-space of the algorithm, thereby taking longer to converge.

Model setup

For this experiment, two models were used, the first without ratio locking and the second with ratio lcocking. For the model without ratio locking, the class label mutation weight (M_{lw}) was set to 1, so that the model would be able to adapt the membership distribution.

Results and observations

The impact of disabling ratio locking was less than expected, possibly due to heavy selection pressure elsewhere in the model. In addition to the ratio-locking, elitism for survivor selection was employed, which could have ameliorated the unwanted antibody drift to some extent by eliminating antibodies with low fitness.

Figure 5.1 shows the mean distribution of the antibodies for the first 20 generations of training on the Glass reference set, after which the distribution stabilises. There is some slight variation in the final class distribution, with classes 1 and 2 increasing their proportion of antibodies, and classes 3, 5 and 6 decreasing. Class 7 remains relatively stable. This minimal variation was much lower than expected and may indicate that the spatial distribution of the antigens has less of an impact on the antibody distribution than hypothesised. The results demonstrate that ratio-locking could be a viable alternative to allowing for the mutation of class labels.

Tables 5.5 and 5.6 show the results from running the model without and with ratio locking respectively. As can be seen, the performance of the models is very similar, with a marginally higher runtime when employing ratio locking. Noticeably, the variation in training accuracy is slightly improved, which indicates that ratio locking can produce more consistent results. Especially on the complex Glass set, variation is much lower on both the train and test sets. As the impact on runtime was negligible, and since enabling ratio locking removes the need for an additional mutation operator, it was decided to continue to use ratio locking in the model.

Table 5.5: Results of the model without ratio locking on reference sets

Dataset	**Test Accuracy**	**Train Accuracy**	**Avg runtime/fold**
Wine	0.779 (0.116)	0.787 (0.082)	0.51 (0.03)
Diabetes	0.593 (0.116)	0.592 (0.111)	1.73 (0.20)
Ionosphere	0.532 (0.144)	0.553 (0.128)	0.83 (0.03)
Sonar	0.528 (0.136)	0.557 (0.099)	0.52 (0.02)
Iris	0.868 (0.095)	0.866 (0.060)	0.37 (0.02)
Glass	0.412 (0.147)	0.431 (0.106)	0.44 (0.07)

Table 5.6: Results of the model with ratio locking on reference sets

Dataset	**Test Accuracy**	**Train Accuracy**	**Avg runtime/fold**
Wine	0.771 (0.119)	0.796 (0.086)	0.55 (0.03)
Diabetes	0.590 (0.113)	0.601 (0.106)	1.76 (0.21)
Ionosphere	0.550 (0.154)	0.560 (0.121)	0.91 (0.03)
Sonar	0.539 (0.130)	0.570 (0.092)	0.57 (0.02)
Iris	0.872 (0.102)	0.876 (0.069)	0.39 (0.03)
Glass	0.422 (0.110)	0.448 (0.077)	0.48 (0.07)

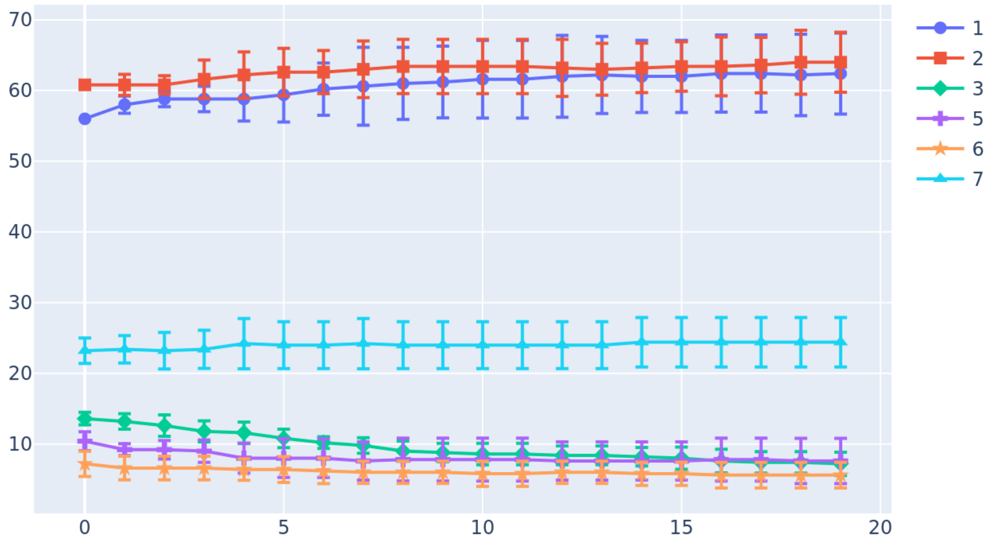


Figure 5.1: Evolution of class distribution without ratio locking for the first 20 generations on the Glass set

Experiment MR2 - Crowding

The second experiment tested how implementing a crowding scheme would affect the model's accuracy. Low diversity was an issue in preliminary testing and led to early convergence of the model, as shown from the results of exp. MR1. Therefore, finding a way to maintain the diversity throughout the evolutionary process was important to avoid local optima and produce an effective classifier.

Hypothesis

As shown by Li *et al.* [30], crowding is a possible approach to the challenge of low diversity in an AIS. Therefore, it was theorised that implementing such a scheme in the proposed algorithm could be beneficial, given the observed tendency towards low diversity in the population. Furthermore, it was theorised that the crowding implementation could counterbalance the high replacement ratio to a certain extent, allowing for progressively increasing the model's accuracy while avoiding the trap of local optima.

Model setup

The model was run for 300 generations with 10-fold cross-validation for each of the reference datasets presented in section 5.2. Following the results of exp. MR1, ratio-locking was used for this experiment, accordingly M_{lw} was set to 0 and the antibodies could no longer switch their class. All other parameters were set as in table 5.4.

Results and observations

The results indicate that crowding is crucial for the proper functioning of the model. When run without crowding, the model displays tendencies to get stuck in local optima, likely due to the selection pressure being too high, as seen from the results in exp. MR1. The implementation of the crowding operator significantly increased the performance of the model, achieving drastically higher accuracy across all reference sets. With crowding, the model is able to better fit the training set as shown from the difference between the results presented in tables 5.6 and 5.7. The variance in the model with crowding is much lower than without, demonstrating that the model can better avoid local optima.

Figure 5.2 shows the average score components of the population when run with crowding. An interesting observation is that even though η was set to 0, the model was able to somewhat optimise Valid Avidity. This suggests that information relating to Valid Avidity is latent in other fitness components being optimised. All fitness components displayed high stability, with little variation throughout the evolutionary loop, which demonstrates that crowding aids the model in becoming more stable.

Figure 5.3 shows the average fitness score of the antibody population without and with crowding on the Ionosphere set. Without crowding, the model very rapidly reaches a local optimum it is unable to escape, as shown from fig. 5.3a. With crowding enabled, the model is able to conserve diversity through the evolutionary process, as shown by the slower increase in fitness in fig. 5.3b. The model is also more stable and achieves greater overall fitness with crowding enabled.

Table 5.7: Results of the model with ratio locking and crowding

Dataset	**Test Accuracy**	**Train Accuracy**	**Avg runtime/fold**
Wine	0.985 (0.027)	1.0 (0.0)	0.53 (0.02)
Diabetes	0.762 (0.034)	0.898 (0.019)	1.98 (0.03)
Ionosphere	0.936 (0.031)	0.947 (0.006)	0.88 (0.13)
Sonar	0.834 (0.105)	0.992 (0.008)	0.50 (0.01)
Iris	0.967 (0.065)	0.993 (0.007)	0.47 (0.01)
Glass	0.711 (0.132)	0.946 (0.016)	0.63 (0.05)

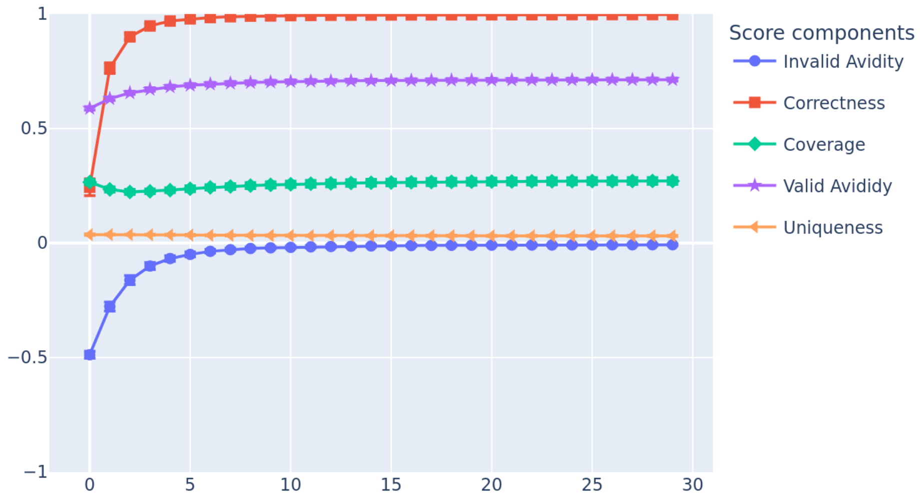
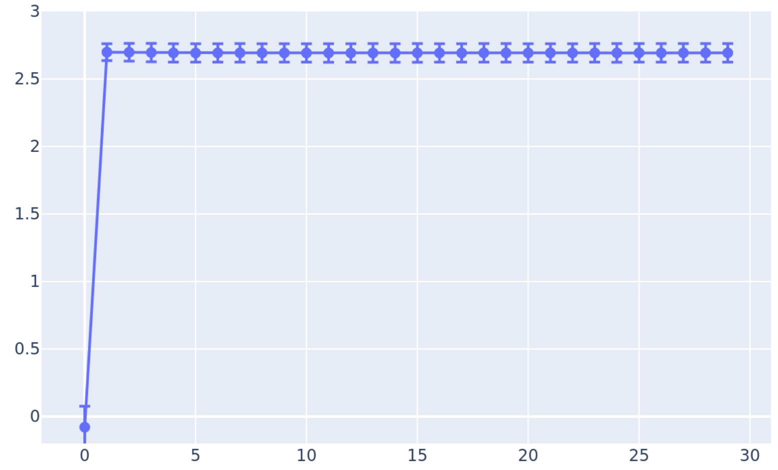
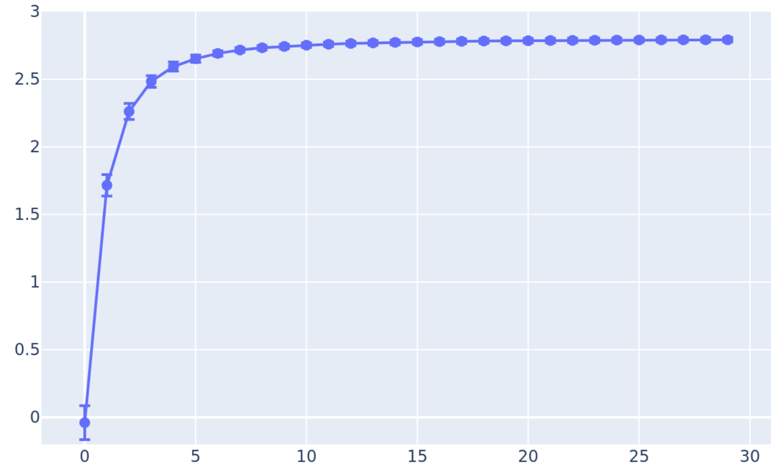


Figure 5.2: Score components on the Ionosphere set with crowding (sampled every 10 generations)



(a) Without crowding



(b) With crowding

Figure 5.3: Average fitness score of the antibody population on the Ionosphere set without and with crowding (sampled every 10 generations)

Experiment MR3 - Boosting

This experiment tested whether boosting could aid the model in better fitting the classification sets. The use of a boosting method is novel for the field of AIS, but it can be said to naturally fit the domain. A boosting classifier runs a weighted set of weak predictors and merges them into a stronger one. As these kinds of classifiers are highly similar to how an AIS uses a lot of weak predictors (antibodies) and merges the results to make a prediction, it was theorised that boosting may provide a benefit to the model. Furthermore, boosting has proven effective in solving many problems (section 2.3.3), therefore, investigating its applicability for AIS was an interesting area of study.

As described in section 4.1.11 the presented model uses a slight variation of ADA-Boost. The purpose of this experiment was to examine whether or not the novel idea of implementing boosting in an AIS model could benefit the presented model.

Hypothesis

The theory underpinning this experiment was that the integration of boosting techniques could bolster an AIS model's performance on the training set. However, boosting techniques may potentially overfit, which could, in turn, degrade the testing accuracy. Additionally, a substantial increase in runtime was anticipated, given that each boosting round executes the full training loop for g_n generations.

Model setup

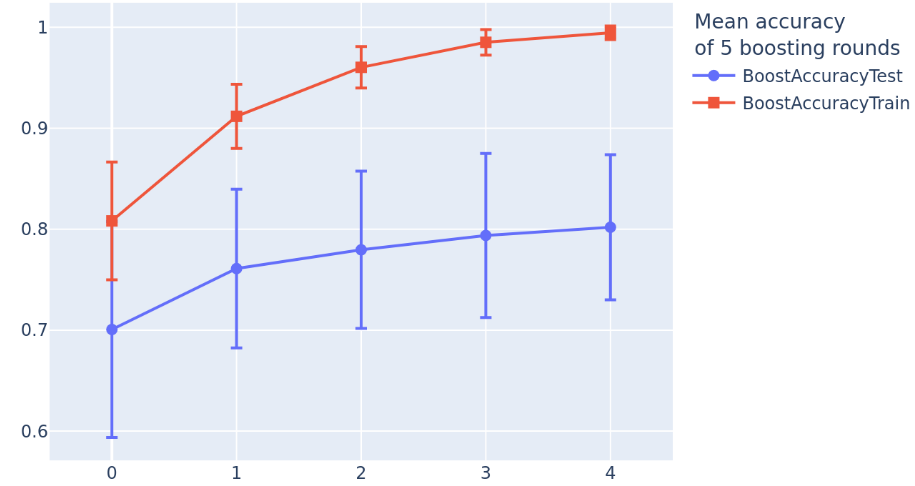
For this experiment, the model was run with $B_n = 5$. Other parameters were set as in table 5.4. The model was again run ten times with 10-fold cross-validation.

Results and observations

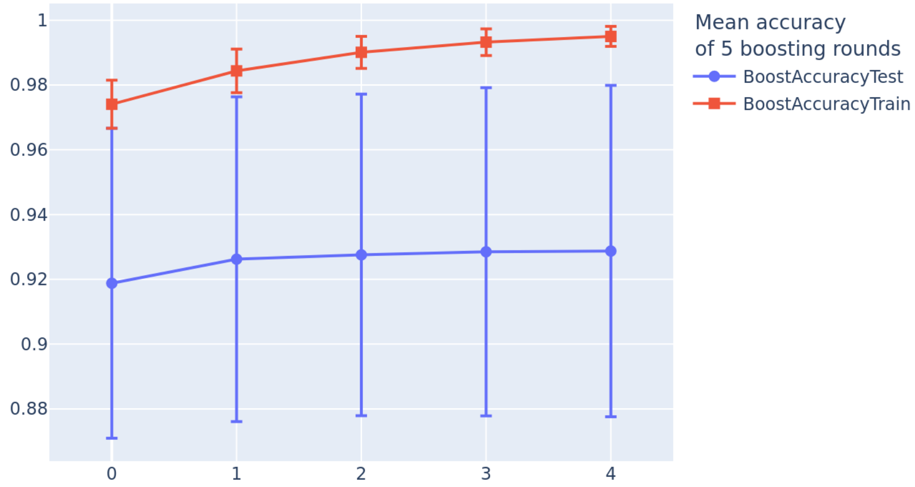
Boosting did not perform quite as well as expected. Figure 5.4 shows the mean training- and testing accuracies of five rounds of boosting over ten runs with 10-fold cross-validation on the Sonar and Ionosphere sets. In both cases, there is a clear trend of the model increasing the training accuracy, with very little variation in the final boosting round. For Sonar, this translates to the test set as well, with testing accuracy gradually increasing. For the Ionosphere set, however, no clear conclusion can be drawn. While mean testing accuracy marginally increases, the variation in results is very high.

Table 5.8 presents the results of boosting on the benchmark datasets. As hypothesised, there was a significant impact on runtime resulting from running the algorithm with boosting enabled. The achieved results were not as good as without boosting (table 5.7); however, the model run with boosting showed increased training accuracy, particularly on the Diabetes and Ionosphere sets. This could indicate that boosting causes the model to overfit, which was also expected. Likewise, there was a considerable increase in the runtime of the model with boosting.

One notable outlier is the high standard deviation for runtime on the Iris set. This is likely an outlier occurring due to a scheduled task running during one of the folds.



(a) Sonar



(b) Ionosphere

Figure 5.4: Aggregate results of 5 rounds of boosting over ten runs with 10-fold cross-validation

Table 5.8: Results of the model with boosting, $B_n = 5$

Dataset	**Test Accuracy**	**Train Accuracy**	**Avg runtime/fold**
Wine	0.984 (0.027)	1.0 (0.0)	1.83 (0.07)
Diabetes	0.760 (0.044)	0.926 (0.017)	8.65 (2.82)
Ionosphere	0.929 (0.051)	0.995 (0.003)	3.04 (0.07)
Sonar	0.799 (0.079)	0.995 (0.007)	1.68 (0.03)
Iris	0.959 (0.055)	0.998 (0.004)	2.71 (7.01)
Glass	0.684 (0.111)	0.922 (0.018)	3.33 (0.21)

Experiment MR4 - Leaking

This experiment was performed to study the impact of leaking (section 4.1.9) on the algorithm. Leaking was implemented as an additional component to aid in increasing the diversity of the model since a lack of diversity was an issue during model refinement, as noted previously.

Hypothesis

It was theorised that increasing L_F (the leak fraction) would inject additional diversity and therefore reduce the stability of the model somewhat. However, it was expected that leaking could enhance the algorithm's performance by aiding it in escaping local optima and exploring other parts of the search-space.

Model setup

To determine the effects of the leaking operation, this experiment used two models: one with L_F set to 0.2, and one with L_F set to 0.8. In both cases, L_R (fraction of leaked population randomly initialised) was set to 0.5. These two models were chosen to give an indication of the effects of leaking by the magnitude of L_F. All other parameters were set as in table 5.4.

Results and observations

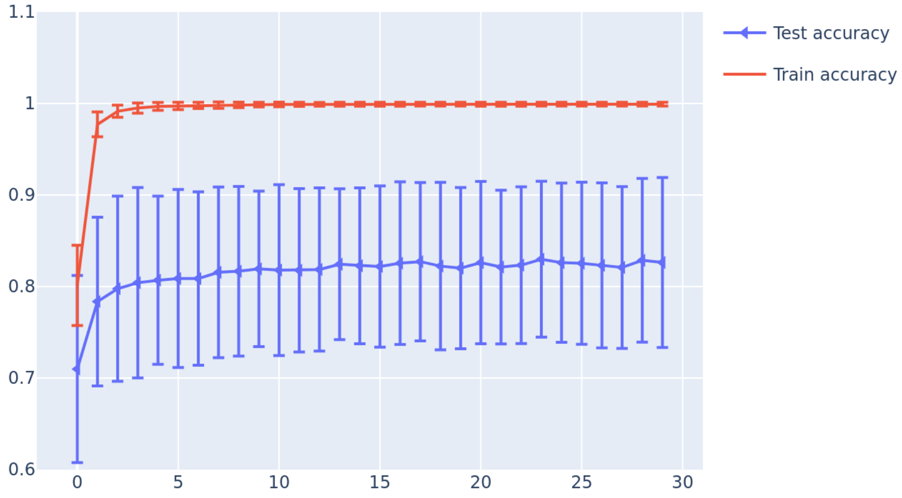
Contrary to expectations, leaking did not have a beneficial effect on the model. It was even detrimental to performance on the more complex Diabetes and Glass sets. As shown from tables 5.9 and 5.10, leaking led to a decrease in training accuracy on the Diabetes and Glass sets, but increased it on the other sets compared to running the algorithm without leaking. Interestingly, for the Wine set, a leak fraction of 0.8 improved testing accuracy, showing that leaking can be beneficial in some circumstances, and may produce better results from tuning the model. A counter-intuitive finding from this experiment was that increasing L_F often increased the model's stability, the opposite of the hypothesised effect. Figure 5.5 shows the mean Train- and Test accuracies over ten runs of 10-fold cross-validation on the Sonar set. The variation in testing accuracy is markedly lower with a higher L_F. The cause of this phenomenon is not readily evident but may be due to stochasticity from L_R being set to 0.5.

Table 5.9: Results of the model with $L_f = 0.2$

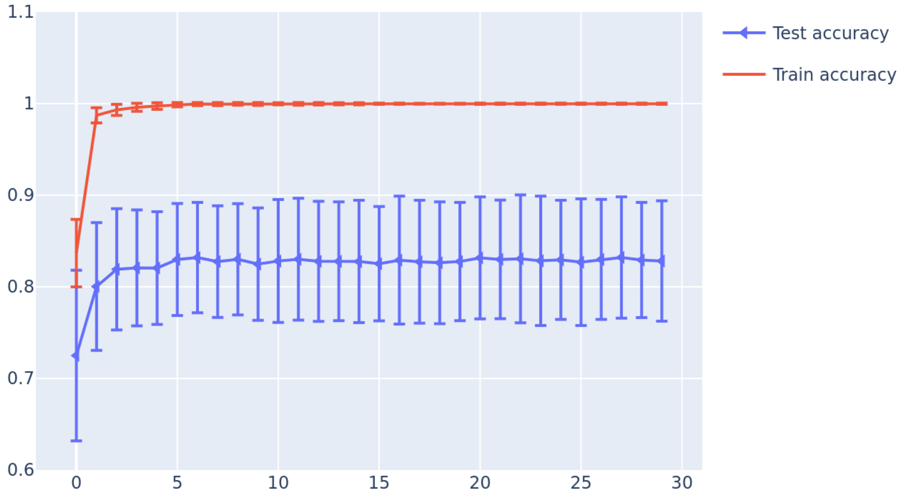
Dataset	**Test Accuracy**	**Train Accuracy**	**Avg runtime/fold**
Wine	0.985 (0.026)	1.0 (0.0)	0.53 (0.02)
Diabetes	0.724 (0.045)	0.823 (0.012)	1.98 (0.03)
Ionosphere	0.934 (0.040)	0.994 (0.003)	0.88 (0.13)
Sonar	0.824 (0.091)	0.999 (0.002)	0.50 (0.01)
Iris	0.961 (0.039)	0.995 (0.005)	0.47 (0.01)
Glass	0.653 (0.110)	0.821 (0.042)	0.63 (0.05)

Table 5.10: Results of the model with $L_f = 0.8$

Dataset	**Test Accuracy**	**Train Accuracy**	**Avg runtime/fold**
Wine	0.991 (0.020)	1.0 (0.0)	0.53 (0.02)
Diabetes	0.707 (0.036)	0.789 (0.013)	1.98 (0.03)
Ionosphere	0.926 (0.027)	0.993 (0.004)	0.88 (0.13)
Sonar	0.829 (0.066)	0.999 (0.001)	0.50 (0.01)
Iris	0.957 (0.061)	0.995 (0.004)	0.47 (0.01)
Glass	0.578 (0.114)	0.721 (0.056)	0.63 (0.05)



(a) With $L_F = 0.2$



(b) With $L_F = 0.8$

Figure 5.5: Difference in Test- and Train accuracy on the Sonar set with varying L_F (sampled every 10 generations)

5.3.1 Phase Discussion

The results of experiments MR1 and MR2 demonstrated that ratio-locking and crowding were beneficial to the model's performance. Boosting and leaking did not conclusively show to benefit the algorithm, but did display some interesting properties which could warrant further research. The models in this section employed evolution and local search of the multiplier, as well as all dimension types. The impacts of these are examined in the following phase.

Comparison to state-of-the-art AIS classifiers

To determine the position of the proposed model in the landscape of the current state-of-the-art of AIS classifiers, the results of several of these algorithms on the benchmark sets detailed in table 5.1 were collated and are shown in table 5.11. While the proposed model did not perform the best on any one set, it attained highly competitive results, outperforming Sverdrup-Thygeson's model [60], MAIM [4] and VALIS [26] on all sets, and outperforming AISLFS [11] on several. It should also be noted that the proposed model was not parameter tuned for any set in particular, but rather used the same parameters for all of them. Furthermore, while most of the compared models do not state their runtime, Ellipsoidal-AIS does, allowing for a rudimentary analysis of their relative efficiency. For the Diabetes reference set, Ellipsoidal-AIS states a mean training time of 119.8 minutes per fold. For comparison, the proposed model achieves a mean training time of 1.98 seconds per fold.

Table 5.11: Model results compared to other AIS classifiers on benchmark sets (best results in **bold**)

Dataset	**FD-AIS**	**AIS for fake news**[60]	**MAIM**[4]	**VALIS**[26]	**AISLFS**[11]	**Ellipsoidal-AIS**[39]
Wine	0.985 (0.027)	0.944 (0.005)	0.967 (0.003)	0.972 (0.005)	0.978 (0.058)	**0.998 (0.026)**
Diabetes	0.762 (0.034)	0.702 (0.006)	0.757 (0.006)	N/A	0.742 (0.090)	**0.804 (0.024)**
Ionosphere	0.936 (0.031)	N/A	0.671 (0.021)	0.928 (0.007)	0.944 (0.049)	**0.971 (0.016)**
Sonar	0.834 (0.105)	N/A	0.675 (0.085)	0.818 (0.020)	**0.881 (0.118)**	N/A
Iris	0.967 (0.065)	0.959 (0.005)	0.965 (0.006)	0.956 (0.005)	0.957 (0.038)	**0.998 (0.012)**
Glass	0.711 (0.132)	N/A	0.640 (0.021)	0.689 (0.024)	**0.754 (0.161)**	N/A

5.4 Experiment Phase 2: Recognition Regions (RR)

Phase 2 was designed to examine the effects of different modifications to the RRs employed in the model. Whereas the preceding phase sought to iteratively increase the model's performance by determining desirable features, this phase examines how various RR schemes impact the model. Refer to table 5.7 for the results of the model where all features examined in this phase are enabled. Following the results of Phase 1, the experiments in this section were run with both crowding and ratio locking but without boosting or leaking, as they did not demonstrate a beneficial effect on the classifier's performance in experiments MR3 and MR4.

Experiment RR1 - Impact of Ellipsoids

The first phase 2 experiment was performed to test the effect of the multiplier in the algorithm. The multiplier allows for constructing ellipsoidal RRs by elongating the axes of a circular RR and should allow for finer selection by adjusting the extent of the RRs in the shape-space.

Ellipsoidal-AIS [39] demonstrated that employing ellipsoidal RRs can produce a highly effective classifier, achieving state-of-the-art results for an AIS. However, the high memory footprint ($O(n^2)$) of the Ellipsoidal-AIS model - associated with the use of rotation matrices to allow for rotation in arbitrary spaces - is problematic for use with high dimensional spaces such as those produced from embeddings. A method to leverage some of the effects of ellipsoidal RRs while forgoing the use of rotation was therefore sought. The model attempts to accomplish this through a multiplier associated with each axis for the antibodies. The purpose of this experiment was to evaluate the effect of non-rotational ellipsoids on the model's performance.

Hypothesis

It was theorised that allowing for a variable multiplier could improve the performance of the model by giving it the ability to more appropriately adapt to local areas of the shape-space. The multiplier also affects the open dimensions, by widening or narrowing their opening, and is theorised to provide benefits for that dimension type in addition to the circular one. No impact on run-time was expected from disabling the multiplier mutation

Model setup

Following the results of phase 1, the model was run with ratio locking and crowding, but without boosting or leaking. The model also employed all dimension types (the effect of disabling these are investigated in exp. RR3). The results are the average of ten runs of 10-fold cross-validation. To test the multiplier's effect, two models were used. The first model disabled the mutation which affects the multiplier, and all multipliers were instantiated with a static value of 1. The second model allowed for mutating the multiplier. Additionally, the model does not perform local search to determine the optimal multiplier, which is examined in the subsequent experiment (exp. RR2).

Results and observations

Table 5.12 presents the results of the model with a static multiplier of 1, while table 5.13 shows the model with an evolved multiplier. For most sets, the impact was not as drastic as expected, however, some sets display a marked difference. Performance on the Sonar and Glass sets is noticeably improved through ellipsoidal RRs, presumably due to the way the antigens are distributed in the shape-space for these sets benefiting from ellipsoids. Furthermore, while the standard deviations on the train and test sets are roughly equal, evolving the multiplier results in more stable test accuracy. This could imply that ellipsoidal RRs are better able to generalise.

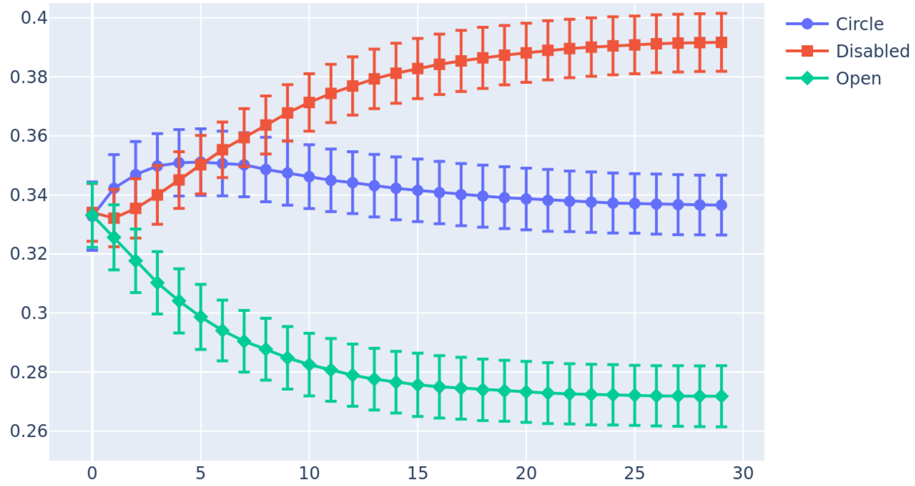
Interestingly, the antibodies show a very slight predisposition toward disabling dimensions, possibly due to this dimension type not being affected by the multiplier. Figure 5.6 shows the distribution of dimension types over 300 generations (subsampled every 10) on the Wine set, with and without a static multiplier. The model with a static multiplier increases the number of disabled dimensions more rapidly and ends up with a marginally higher number of disabled dimensions, and correspondingly a lesser amount of open and circular dimensions. This result may indicate that the model compensates somewhat for the lack of a variable multiplier by its freedom in the mutation of dimension types.

Table 5.12: Results of the model with static multiplier

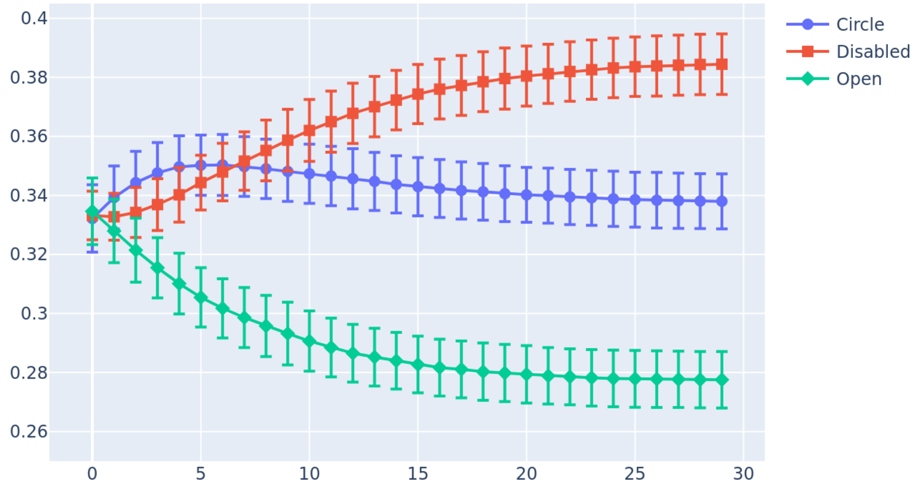
Dataset	**Test Accuracy**	**Train Accuracy**	**Avg run-time/fold (s)**
Wine	0.979 (0.031)	1.0 (0.0)	0.17 (0.01)
Diabetes	0.756 (0.044)	0.815 (0.012)	0.84 (0.02)
Ionosphere	0.902 (0.046)	0.954 (0.007)	0.29 (0.01)
Sonar	0.770 (0.092)	0.961 (0.029)	0.14 (0.01)
Iris	0.951 (0.078)	0.992 (0.007)	0.15 (0.01)
Glass	0.555 (0.072)	0.627 (0.054)	0.21 (0.02)

Table 5.13: Results of the model with evolvable multiplier

Dataset	**Test Accuracy**	**Train Accuracy**	**Avg run-time/fold (s)**
Wine	0.978 (0.032)	1.0 (0.0)	0.17 (0.01)
Diabetes	0.745 (0.033)	0.805 (0.010)	0.87 (0.01)
Ionosphere	0.891 (0.070)	0.948 (0.009)	0.29 (0.01)
Sonar	0.781 (0.077)	0.953 (0.033)	0.15 (0.01)
Iris	0.949 (0.043)	0.994 (0.006)	0.15 (0.01)
Glass	0.568 (0.101)	0.637 (0.051)	0.22 (0.02)



(a) With static multiplier



(b) With evolved multiplier

Figure 5.6: Difference in dimension type distribution relating to the multiplier on the Wine set (sampled every 10 generations)

Experiment RR2 - Local Search

The second experiment of Phase 2 was designed to evaluate the effect of employing local search for the multiplier. Local search has demonstrated benefits on a range of machine learning problems, and it was theorised that it could be advantageous to the proposed model as well. The specifics of the local searching method are described in section 4.1.7.

To aid the goal of having the antibodies cover a greater number of antigens, local search enables the antibodies to find multipliers not easily reachable through mutations. A potential issue with this approach is the likelihood of ending in local optimums due to decreased diversity resulting from offspring using the same multiplier.

Hypothesis

Applying local search was expected to negatively impact the run-time of the model, due to the local search operator taking longer to compute. On the other hand, it was predicted that employing local search would lead to improved results since the antibodies would be able to cover the antigens more appropriately.

Model setup

For the experiment, the model was run with local search for the multiplier enabled, but without the multiplier mutation. For reference, table 5.7 shows the result of enabling both local search and the multiplier mutation. The other model parameters are the same as in Phase 1, shown in table 5.4.

Results and observations

The application of local search enhances the algorithm's performance across nearly all datasets. However, the results are not quite up to par with the model run with both local search and multiplier mutation (table 5.7). This disparity may likely stem from the local search model reaching a saturation point, consequently leading to reduced population diversity.

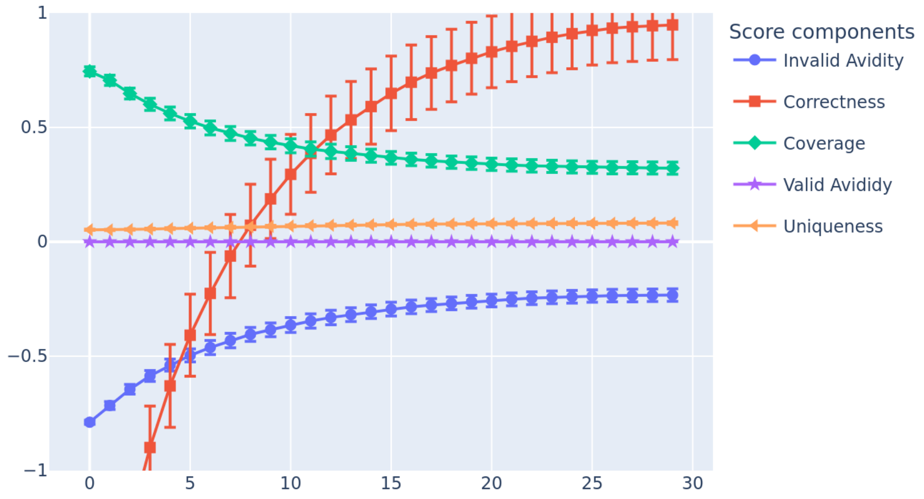
As predicted from the hypothesis, a noticeable increase in the mean run-time of the model was observed. This was expected due to the inherently greater computational demands of the local search operator compared to multiplier mutation.

Figure 5.7 shows the evolution of the average score of each fitness component for the models with an evolved multiplier and a local searched multiplier over 300 generations on the Glass set. The first data point in the graphs occurs after the first generation, meaning that the population has already been subjected to a round of mutation and replacement. An interesting observation is that local search results in much less variation in several score components, particularly Correctness. Recall from section 4.1.5 that Correctness measures the ratio of correctly classified to incorrectly classified antigens by an antibody. Hence, it can be inferred that local search promotes the rapid proliferation of antibodies that exhibit superior detection accuracy, relative to those generated by the model leveraging an evolved multiplier. Local search also shows increased stability for the Correctness fitness component, indicating that a greater proportion of antibodies have a high Correctness value. Also noteworthy is the local search model's ability to better optimise Valid- and Invalid Avidity. These fitness components are most relevant for open dimensions and are highly sensitive to an appropriate multiplier.

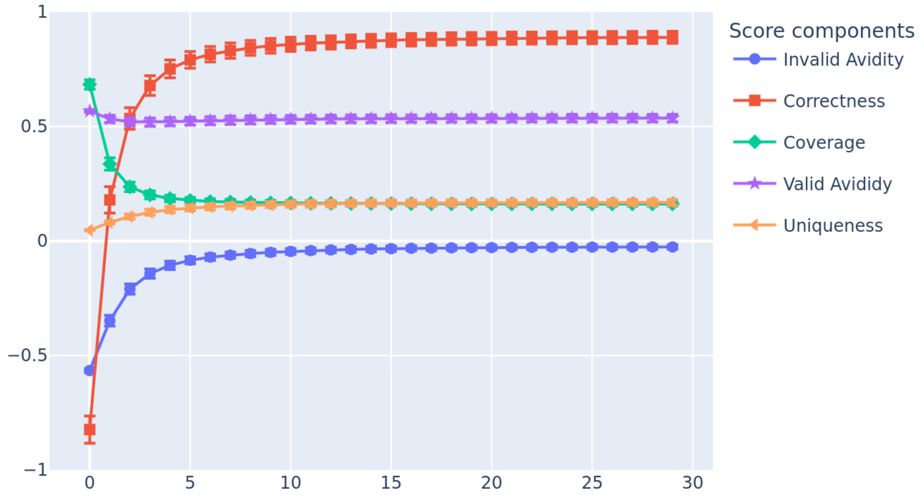
The model employing the local search operator displays a marked increase in training accuracy on all benchmark sets compared to the one with an evolved multiplier. This difference is particularly pronounced in the case of the Glass set, presumably due to the set's inclusion of six classes. Discovering an optimal multiplier that does not interfere with other antibodies is important to attain good performance on the set.

Table 5.14: Results of the model with local search for multiplier and without mutation of the multiplier on reference sets

Dataset	**Test Accuracy**	**Train Accuracy**	**Avg run-time/fold (s)**
Wine	0.985 (0.031)	1.0 (0.0)	0.63 (0.03)
Diabetes	0.752 (0.04)	0.917 (0.016)	2.32 (0.04)
Ionosphere	0.934 (0.036)	0.995 (0.003)	1.04 (0.02)
Sonar	0.827 (0.088)	1.0 (0.001)	0.59 (0.02)
Iris	0.957 (0.054)	0.999 (0.003)	0.56 (0.01)
Glass	0.698 (0.110)	0.926 (0.023)	0.90 (0.08)



(a) With evolved multiplier



(b) With local searched multiplier

Figure 5.7: Difference in evolution of score components with evolved multiplier compared to local searched multiplier on Glass set (sampled every 10 generations)

Experiment RR3 - Effect of Dimension Types

The third Phase 2 experiment examined how the various dimension types, in particular the novel ”open” dimension type, impact the algorithm.

As bounded RR are afflicted with various issues when applied to high-dimensional shape-spaces (section 2.4.2), it was theorised that adopting unbounded RRs in the model might mitigate this to some extent. Therefore, a new open RR type was proposed, which allows for the dimensions of a RR to be opened, creating an unbounded RR intended to improve coverage of the shape-space (section 4.1.3). While the efficacy of disabled dimensions in concert with circular dimensions has already been demonstrated by AISLFS [11], for the purposes of the proposed model it was interesting to see how the combination would interact. This experiment was therefore also run with a model with disabled and circular dimensions. Finally, eschewing circular dimensions altogether was also tested. Recall from section 4.1.7 that when circular dimensions are not used, the RRs become hyperplanes, effectively transforming the model into a collection of linear classifiers.

Hypothesis

It was theorised that giving the antibodies the ability to create RRs which could cover larger areas of the shape-space would yield better performance on high-dimensional data-sets, without increasing the number of antibodies needed. Therefore, the model with open and circular dimensions was expected to perform better than the one with disabled and circular dimensions. The model with only open and disabled dimensions was expected to achieve good results for simple sets, but would likely fare worse on more complex ones, especially those requiring multi-class classification. The run-time was expected to be slightly lower for the models employing the disabled dimension type since it results in simplified calculations.

Model setup

The experiment involved the use of both crowding and ratio locking. Additionally, both an evolved multiplier and local search were enabled for this experiment. To properly examine the impact of the open dimension type, three models were used. The first was run without open dimensions, including only circular and disabled types. Subsequently, a model was run without disabled dimensions, considering only open and circular types. Finally, a model was run without circular dimensions, focusing solely on disabled and open types. The general parameters were again as in table 5.4.

Results and observations

Table 5.15 shows the results of the model when open dimensions are turned off, while table 5.16 shows the results of running without disabled dimensions. Table 5.17 shows the results from the model with only open and disabled dimensions.

The findings from the Glass dataset across the three models are noteworthy. The model without disabled dimensions is able to achieve a better fit to the training data than the other models. With only open and disabled dimensions, the training accuracy is particularly lower, suggesting that this model is unable to fit the Glass set as well as the other model. As noted, this set is quite sensitive to appropriate antibodies, and the hyperplane RRs of the model without circular dimensions may interfere with one another and cause this low training accuracy. This model's accuracy is also worse on the Diabetes training set compared to the others. However, here the model is able to achieve results on the test set on par or better than the other models. This may indicate a degree of overfitting on this set on the part of the others. The model with only open and circular dimensions produced the overall best results across the sets, which demonstrates the utility of the open dimension type.

There is an impact on run-time associated with the model without disabled dimensions. This is likely due to the fact that the disabled dimensions result in simpler calculations for antibody coverage, while the more complicated calculations associated with open dimensions may result in longer computation times.

Table 5.15: Results of the model without open dimensions on reference sets

Dataset	**Test Accuracy**	**Train Accuracy**	**Avg run-time/fold (s)**
Wine	0.975 (0.031)	0.999 (0.001)	0.45 (0.02)
Diabetes	0.738 (0.033)	0.921 (0.026)	1.73 (0.02)
Ionosphere	0.700 (0.043)	0.982 (0.007)	0.72 (0.02)
Sonar	0.352 (0.097)	0.943 (0.016)	0.37 (0.01)
Iris	0.949 (0.060)	0.989 (0.007)	0.43 (0.01)
Glass	0.668 (0.118)	0.909 (0.026)	0.55 (0.05)

Table 5.16: Results of the model without disabled dimensions on reference sets

Dataset	**Test Accuracy**	**Train Accuracy**	**Avg run-time/fold (s)**
Wine	0.986 (0.026)	1.0 (0.0)	0.66 (0.03)
Diabetes	0.754 (0.054)	0.918 (0.011)	2.36 (0.03)
Ionosphere	0.926 (0.046)	0.996 (0.003)	1.09 (0.02)
Sonar	0.814 (0.069)	0.999 (0.001)	0.61 (0.02)
Iris	0.959 (0.033)	0.999 (0.003)	0.57 (0.02)
Glass	0.696 (0.107)	0.943 (0.017)	0.77 (0.06)

Table 5.17: Results of the model without circular dimensions on reference sets

Dataset	**Test Accuracy**	**Train Accuracy**	**Avg run-time/fold (s)**
Wine	0.983 (0.028)	1.0 (0.0)	0.57 (0.03)
Diabetes	0.754 (0.047)	0.800 (0.009)	1.83 (0.04)
Ionosphere	0.874 (0.050)	0.923 (0.007)	0.74 (0.02)
Sonar	0.796 (0.088)	0.998 (0.003)	0.51 (0.02)
Iris	0.941 (0.047)	0.970 (0.012)	0.43 (0.01)
Glass	0.653 (0.129)	0.701 (0.025)	0.55 (0.04)

Experiment RR4 - Effect of Population Size with Open Dimensions

As open dimensions were thought to provide better coverage of the shape-space with fewer antigens, adjusting AB_p (the antibody population fraction) could result in further insight into open dimensions in the context of the proposed algorithm.

Hypothesis

It was theorised that effective classification would be possible even with a small number of antibodies employing open dimensions, as they should be able to envelop a greater amount of antigens. The model with a lower AB_p was not thought to increase the performance - the results were expected to be somewhat lower than those displayed in table 5.16 from exp. RR3. However, lowering AB_f should ameliorate the run-time, as the algorithm no longer has to perform calculations for as many antibodies. Meanwhile, increasing the antibody population fraction from 1.0 may provide benefits by reinforcing the number of antibodies participating in a given antigen's class assignment, and may even boost the model's performance. An increase of AB_p was expected to result in a higher run-time.

Model setup

For this experiment, two models were employed, one decreasing AB_p from 1.0 to 0.5, and one increasing it to 2.0. The models were run with ratio-locking, crowding and local search enabled, and without boosting. The allowed dimension types were circular and open. All other parameters were as shown in table 5.4

Results and observations

Decreasing the antibody ratio does not have as drastic an effect as was expected, performing comparably to the results of exp. MR2 on most sets. Two outliers are the Diabetes and Glass sets, where the model performs under par. On the Glass set, the model shows a tendency towards underfitting during training, achieving around 60% accuracy. This may be attributed to the smaller number of antibodies not being sufficient for the six-class classification of the Glass set. When run with a smaller antibody fraction, the model achieved a certain degree of improved efficiency, though this was not as significant as hypothesised.

For the model with the increased antibody ratio, the results are interesting. While the performance on most sets is comparable to the model employed for exp. MR2, on the Glass set, it achieves superior results, possibly indicating that the open dimension type may be advantageous for classification on sets with many labels. Furthermore, the model achieves improved training accuracy in most cases, which could further indicate that open dimensions can result in improved coverage of the shape-space, supporting the findings of exp. RR3. However, there is a significant impact on the run-time of the model resulting from increasing the antibody ratio due to a greater amount of antibodies needing to be calculated. While the antibody population is four times as large in the model with $Ab_p = 2.0$, the run-time is roughly doubled, which was less than expected. This likely comes from optimisation and implementation specifics.

Table 5.18: Results of the model with open and circular dimensions, $Ab_p = 2.0$

Dataset	**Test Accuracy**	**Train Accuracy**	**Avg run-time/fold (s)**
Wine	0.989 (0.022)	1.0 (0.0)	0.93 (0.04)
Diabetes	0.757 (0.036)	0.961 (0.009)	3.51 (0.03)
Ionosphere	0.932 (0.045)	0.998 (0.002)	1.44 (0.01)
Sonar	0.826 (0.079)	1.0 (0.0)	0.82 (0.02)
Iris	0.960 (0.033)	1.0 (0.0)	0.93 (0.02)
Glass	0.747 (0.106)	0.981 (0.009)	1.19 (0.09)

Table 5.19: Results of the model with open and circular dimensions, $Ab_p = 0.5$

Dataset	**Test Accuracy**	**Train Accuracy**	**Avg run-time/fold (s)**
Wine	0.985 (0.032)	1.0 (0.0)	0.46 (0.02)
Diabetes	0.719 (0.044)	0.835 (0.023)	1.66 (0.02)
Ionosphere	0.922 (0.052)	0.991 (0.004)	0.77 (0.02)
Sonar	0.821 (0.061)	0.989 (0.008)	0.41 (0.01)
Iris	0.957 (0.045)	0.994 (0.005)	0.40 (0.01)
Glass	0.627 (0.109)	0.795 (0.037)	0.55 (0.05)

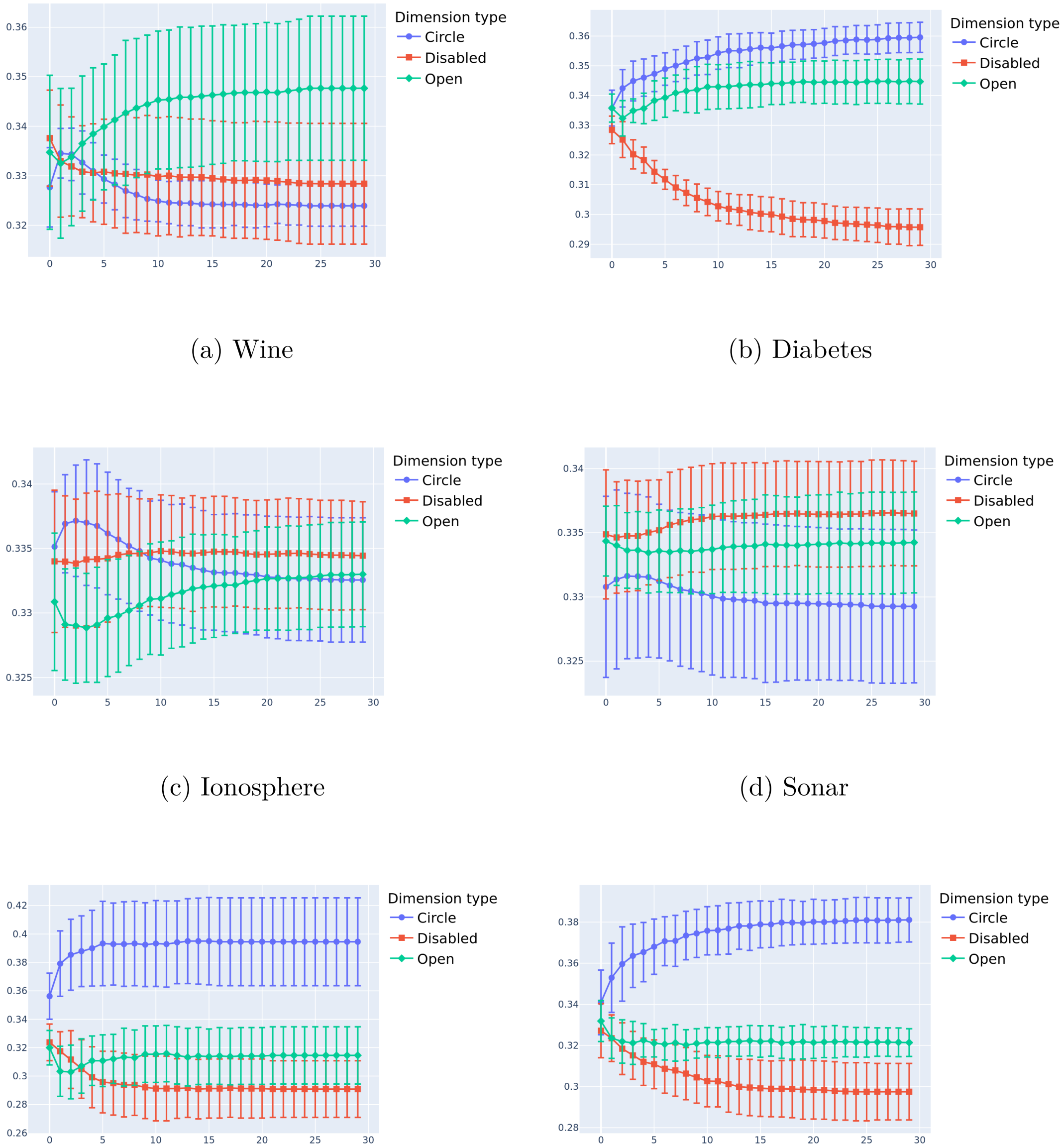
5.4.1 Phase Discussion

The results from the experiments in this phase show that alternate RR geometries can produce markedly different results, in line with the results expected from the research by Hart [19]. The algorithm tends to perform better with circular and open dimensions than combinations of the other dimension types, as shown from exp. RR3. It should be noted that this does not conclusively demonstrate that open dimensions are an appropriate choice for all models, but they do show an advantage in the context of the algorithm proposed herein. Furthermore, increasing the population ratio of the model with open and circular dimensions produced results on par or better than the model with all dimension types (results of which are shown in table 5.7). Open and circular dimensions also showed a lower variation in results, seen from the standard deviations listed in table 5.18. Particularly on the Glass set, open dimensions displayed a marked advantage, which may indicate that this dimension type could be useful for multi-class classification of higher arity.

While mutation of the multiplier did not have as marked an impact as hypothesised compared to a static multiplier, leveraging local search to determine an appropriate multiplier led to the faster convergence of the model and better results with less variation in fitness across the runs, shown in fig. 5.7. Local search did have an impact on the runtime of the model though not as large as expected, likely due to the optimisation of the operator. Research into the effects of local search operators in other AIS models could be an interesting area for further work.

Difference in evolved dimension type distribution

During the experiments in this Phase, it was observed that the final dimension type distribution of the model varied across the sets, which could indicate that the dimension types that are most suited differ from set to set. Allowing the model to mutate the dimension type should ensure that it is eventually able to find a distribution which is appropriate for the task at hand. Figure 5.8 shows the evolved distribution of dimension types on the reference sets. The models were run with all dimension types, ratio-locking, crowding, local search and an evolved multiplier. Note that e.g. Wine evolves a relatively greater amount of open dimensions, while Diabetes progressively moves away from disabled dimensions. On Ionosphere and Sonar, the difference is less pronounced, with Ionosphere converging to a roughly equal amount of each dimension type, and Sonar displaying heavy variation. On Iris and Glass, the circular dimension type seems most appropriate. Of note is the difference between the dimension type distribution presented in fig. 5.6 and fig. 5.8 on the Wine set. In the latter, local search for the multiplier is enabled, resulting in an increased proclivity towards open dimensions not present without local search. This result further substantiates the notion that an appropriate multiplier is important for open dimensions.



(a) Wine

(b) Diabetes

(c) Ionosphere

(d) Sonar

(e) Iris

(f) Glass

Figure 5.8: Difference in evolved distribution of dimension type on reference sets

5.5 Experiment phase 3: Embedding and semantic features for fake news detection (ES)

Phase 3 was concerned with evaluating the proposed approach for fake news detection. Exp. ES1 looks at the embedding approach proposed in chapter 4 and the effect of whitening and dimensionality reduction on the model's performance. Exp. ES2 employs the general textual features selected in section 4.2.3 and evaluates the performance of the model on the fake news datasets in section 5.2.

Experiment ES1 - Effect of whitening and dimensionality reduction of sentence embeddings

This experiment explored the impact of whitening and dimensionality reduction of sentence embeddings used in the model. While whitening has demonstrated increased performance for semantic similarity tasks, it is interesting to examine whether this transfers to classification with an AIS model.

Hypothesis

It was hypothesised that leveraging the rich semantic content present in embeddings would benefit the model's performance as embeddings have demonstrated their applicability for many NLP domains. The novel ”open” dimension type was expected to provide good coverage of the high-dimensional embedding space and be able to detect the majority of the antigens.

Reducing the anisotropy in the sentence embedding space should improve the classification accuracy for an AIS, as it will reduce the ”cone” effect described in section 2.5.4 and therefore allow antibodies to better group antigens in the shape-space. Additionally, dimensionality reduction should mitigate the impact of the curse of dimensionality and allow for better coverage of the shape-space per antibody.

Model setup

All the initial dimensions were set to open. The motivation behind this was to cover a large portion of the space initially; however, the model was allowed to mutate into other dimension types to better tailor each RR to local areas in the shape-space. All other parameters were again set as in table 5.4.

For this experiment, boosting was employed. The theory was that the high coverage resulting from initialising all dimensions as open would result in weak learners with a good chance of producing better than random hypotheses, thus being well-suited for boosting. The model was configured with $B_n = 10$.

ζ, the weight corresponding to Invalid Avidity, was set to 0 for this experiment. The motivation behind disabling the avidity parts of the fitness score was that the large RRs would respond poorly to optimising these fitness components. Furthermore, preliminary testing supported this notion.

The antibody population was set to a static size of 300 per boosting round, as the large size of the datasets precluded the possibility of running the model with a one-to-one antigen-antibody ratio. Furthermore, the open dimensions were theorised to provide adequate coverage even with a low population size, supported by the findings from exp. RR4.

For this experiment, it was decided not to use cross-validation, but rather employ a greater train-test split, as the datasets contained a large number of observations and would give a good indication of the model's performance without needing to run the model as many times. For the Kaggle set, the model employed a 25-75 train-test split, amounting to around five thousand articles for the training set. For the BuzzFeed and PolitiFact sets, an 80-20 train-test split was used.

Results and observations

Table 5.20 shows the result of the model employing embeddings on the fake news datasets with varying degrees of whitening.

The impact of whitening and dimensionality reduction was greater than expected. Due to the curse of dimensionality, a higher number of dimensions makes coverage of the shape-space more difficult (section 2.4.2). The results on the model without dimensionality reduction may imply that there is a limit to how well the open dimension type can cover the shape-space.

Some inconsistency in accuracy was observed between runs with the same model. One phenomenon that may explain the inconsistency between the results is that the model predicts the majority class after a boosting round. This leads all the misclassified antigens to progressively get higher and higher weights until the predictor "flips" and the algorithm oscillates forward and backwards between classes, as can be seen in fig. 5.9. Notice that the model has a stable accuracy of the majority class, but after several rounds, the boosting weights for the other values get so high that the model starts to flip and change accuracy. The situation is sometimes recoverable, as in fig. 5.11 where the model first predicts the majority class with 60%, then flips to the other class with 40%, before finally stabilising and gradually getting better. This phenomenon occurs sporadically; however, when it does not the results produced are significantly better. For the model run on the Kaggle set with whitening and 256 dimensions, the result achieved was around 50% due to the described issue. In runs where this oscillation does not occur the accuracy is markedly higher (0.8560).

The findings from this experiment suggest that the model is able to leverage the semantic information present in embeddings, even when some of this information is condensed through dimensionality reduction. Further work is needed on the dynamics and interactions of the open dimension type to determine the causes of the observed oscillation phenomenon.

Table 5.20: Results for the model employing embeddings on the fake news sets with varying dimensionality reduction

Dataset	**90 dims**	**256 dims**	**No whitening (768 dims)**
Kaggle	0.8270	0.4998	0.4996
PolitiFact	0.8494	0.6869	0.4062
Buzzfeed	0.7793	0.3547	0.6751

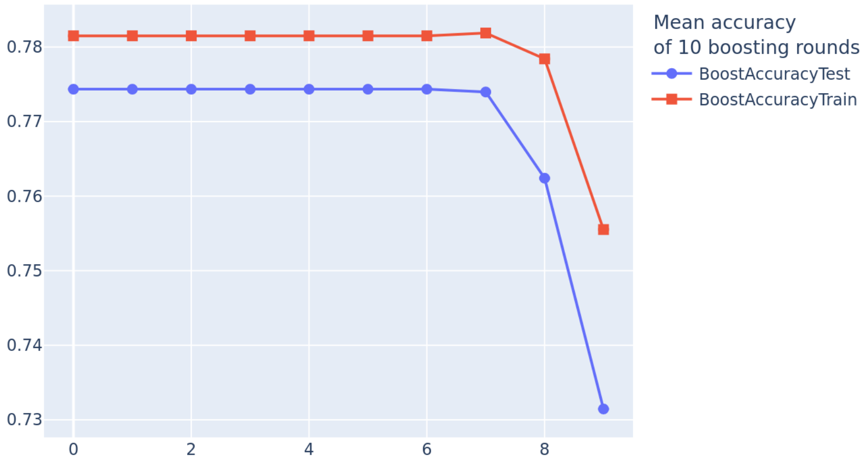


Figure 5.9: Buzzfeed with whitening and reduced to 90 dimensions

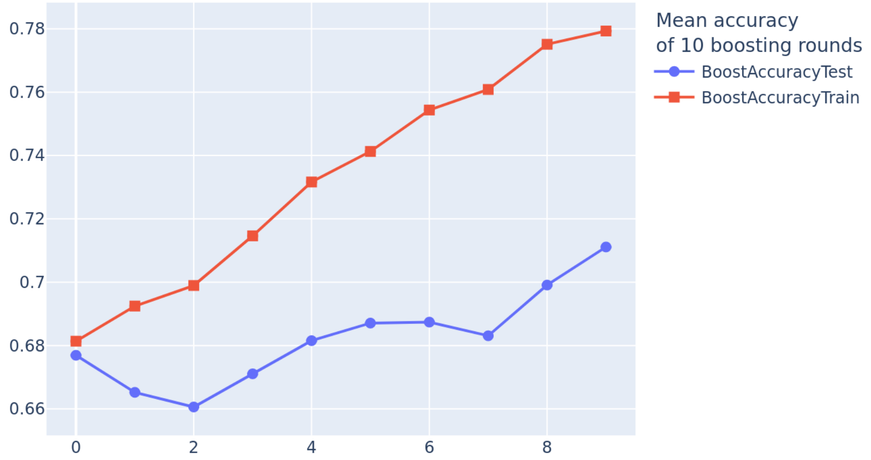


Figure 5.10: Kaggle with whitening and reduced to 90 dimensions

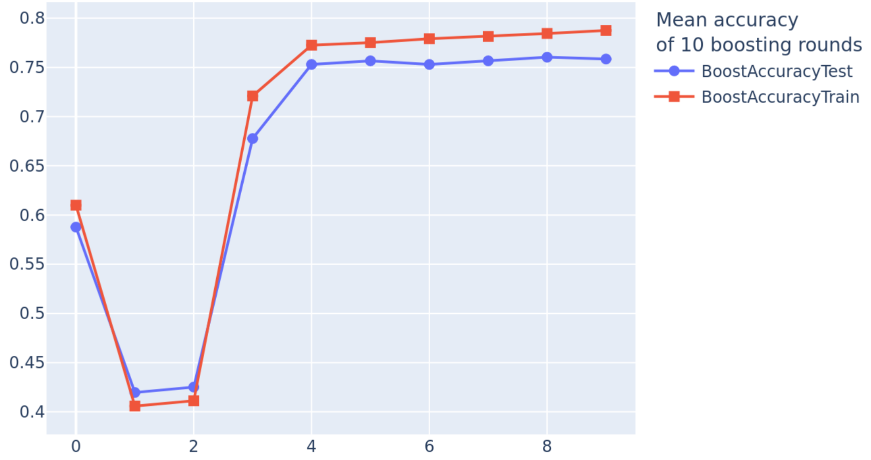


Figure 5.11: Politifact with whitening and reduced to 90 dimensions

Experiment ES2 - Fake news detection with general textual features

This experiment evaluated the model's performance for fake news detection when employing the general textual features from section 4.2.3. Both run-time and accuracy were considered.

Hypothesis

While content-based classification with an AIS has already demonstrated viability, the proposed model displayed promising performance on the benchmark classification sets. Therefore, it was theorised that it would translate to the task of fake news detection and achieve high performance there as well. However, there is a limitation in that feature analysis of the chosen features was not performed, and those selected were instead done so on the basis of the results from the literature review. Therefore, there was some uncertainty as to how the model might fare when employing these features.

Model setup

The parameters for this experiment were again set as in table 5.4. Notably, neither boosting nor leaking were employed, as the results from experiments MR3 and MR4 showed that they were not beneficial for non-embedding-based classification. All dimension types were used.

Results and observations

Table 5.21 shows the result of the model when run with the general textual features selected in section 4.2.3.

The model's performance on the fake news sets was not as good as expected from the results on the benchmark sets. A possible cause of this is that the chosen features were not well selected for the purposes of the model. Performing feature analysis will likely improve the performance of the classifier by reducing the number of irrelevant features. On the Kaggle and Politifact sets, the model is able to avoid the trap of majority classification. On the other hand, the results on the BuzzFeed set suggests that more work is needed to counterbalance the majority class tendency for heavily imbalanced datasets. Here, the model ends up performing majority classification.

Table 5.21: Results of the model on the fake news data sets with general textual features

Dataset	**Test Accuracy**	**Train Accuracy**	**Avg runtime/fold (s)**
Kaggle	0.735	0.634	39.406 (0.481)
Politifact	0.774	0.948	2.1006 (0.037)
Buzzfeed	0.777	0.777	6.3377 (0.111)

5.5.1 Phase discussion

The algorithm displays fairly good performance when run with embeddings. The open dimension type demonstrated substantial coverage of the shape-space. Even in the absence of dimensionality reduction, the algorithm successfully covers all antigens, demonstrating the potential of open dimensions in AISs intended for high-dimensional problems. Refining the open dimensions and enhancing generalisation ability by identifying suitable fitness functions might further improve the model.

When run with semantic features, the results were inferior to what had been expected from the performance on the benchmark classification sets. The drop in performance may be explained by issues with feature selection. The adopted features were not selected based on feature analysis, but rather on the basis of the literature review.

On the other hand, when run with general textual features, the results did not meet the expectations set from the results of the benchmark classification sets. This decline in performance could be attributed to potential challenges in feature selection. The chosen features were not the outcome of a feature analysis but were rather selected based on the literature review. Performing a feature analysis to identify the most salient features will likely elevate the classification accuracy for the model with general textual features.

Comparison to state-of-the-art fake news detection methods

The method does not achieve state-of-the-art performance on most sets, with the performance on the Kaggle set being especially inferior compared to the rest of the field. On the Politifact set, however, the model is able to achieve results performing on par with several models. It should be noted that not all models compared against are content-based. For instance, the models by Tseng *et al.* and Rai *et al.* are information fusion models, and incorporate a knowledge base and leverage the social context surrounding the news articles.

Table 5.22: Comparison of model performance on the Kaggle set

Kaggle set	
Model	**Accuracy**
FD-AIS(Embeddings)	0.856
FD-AIS(General Textual)	0.735
Sverdrup-Thygeson [60]	0.917
Kaliyar *et al.* [25]	0.989
Verma *et al.* [66]	**0.991**
Mandical *et al.* [32]	0.870

Table 5.23: Comparison of model performance on the Politifact set

Politifact set	
Model	**Accuracy**
FD-AIS(Embeddings)	0.850
FD-AIS(General Textual)	0.774
Dun *et al.* [12]	0.858
Tseng *et al.* [62]	**0.925**
Rai *et al.* [45]	0.888
Yang *et al.* [71]	0.837
Shu *et al.* [56]	0.878
Pérez-Rosas *et al.* [41]	0.811
Castillo *et al.* [7]	0.794
Zhou *et al.* [74]	0.892

Table 5.24: Comparison of model performance on the BuzzFeedNews set

BuzzFeedNews set	
Model	**Accuracy**
FD-AIS(Embeddings)	0.779
FD-AIS(General Textual)	0.777
Shu *et al.* [56]	0.864
Zhou *et al.* [74]	**0.879**
Pérez-Rosas *et al.* [41]	0.755
Castillo *et al.* [7]	0.789

6 Evaluation and Conclusion

This chapter provides an overall assessment of the work presented in this thesis and the contributions to the field and some avenues. Finally, some proposals for future research relating to the proposed model are presented.

6.1 Goal Evaluation

Research question 1 *How can an AIS model be adapted to operate on high-dimensional spaces?*

Through the introduction of the novel open dimension type for the RRs of antibodies within the AIS model, it was able to more effectively operate in high-dimensional space, an area which had proven challenging with bounded RRs. Open dimensions demonstrated improved performance on benchmark classification sets. Furthermore, exp. ES1 showed that open dimensions provide good coverage for the shape-space resulting from employing embeddings. The results from exp. RR3 suggest that open dimensions might outperform disabled dimensions in terms of overall performance. The experiment described in exp. RR4 provided evidence that even with a smaller population size of antibodies, open dimensions can deliver effective classification. This observation could make AIS models more computationally efficient by reducing calculations associated with antibodies. However, the results from the experiment in exp. RR2 highlighted the sensitivity of open dimensions to an appropriate multiplier for proper coverage of the shape-space. Therefore, optimization of this aspect should be a consideration when designing AIS models incorporating open dimensions.

Overall, open dimensions showcase the potential of unbounded RRs, and work to further optimise them might advance the field by producing AIS models which are well-suited for high-dimensional shape-spaces.

Research question 2 *How do embeddings compare to general textual features in terms of performance for fake news detection with an AIS?*

The performance of models employing embeddings was compared with those of models using general textual features within the context of fake news detection using an AIS model. The findings indicate that embeddings outperformed general textual features in terms of accuracy for fake news detection; however, this may be due to limitations stemming from the selected general textual features.

The performance of the model employing general textual features on the fake news sets was mixed. On the other hand, the model operating with embeddings displayed satisfactory performance on the Kaggle and Politifact sets. However, for the Buzzfeed set, the model resorted to majority classification. Therefore, additional research is required to better manage imbalanced datasets when using embeddings.

Whitening and reducing the dimensionality of the embeddings improved the algorithm's performance (exp. ES1), showing that it is able to leverage the semantic content in embeddings even when the content has been condensed. The findings may also suggest that the observed effect of increased performance on semantic similarity tasks by Su *et al.* [59] translates to semantic classification.

However, there are certain limitations associated with using embeddings. The method is less scalable, suggesting that scalability could be a challenge when applying this approach to large-scale fake news detection systems. Potentially, other approaches to employing embeddings as features could be investigated. In such cases, open dimensions could prove beneficial.

There is a limitation associated with the results of the model employing general textual features in that feature analysis was not performed. The selection of appropriate features for the general textual features might further enhance the performance of the AIS model using these.

Research question 3 *How does the accuracy of the proposed model compare to state-of-the-art fake news detection methods?*

The accuracy of the proposed model was evaluated in comparison to state-of-the-art fake news detection methods. The findings indicate that while the proposed model does not perform as well as state-of-the-art models employing neural networks or information fusion, it still displays highly competitive results on many datasets, with the added benefit of improved training time.

Although the proposed model may not achieve the same level of accuracy as neural network-based models, it still demonstrates competitive performance in terms of fake news detection. This suggests that the proposed model offers a viable alternative with efficient training times.

6.2 Contributions

In conclusion, this thesis focused on investigating the application of AIS for fake news detection, exploring various techniques to improve the accuracy and effectiveness of the model, particularly with regard to high-dimensional spaces. A novel dimension type for recognition regions was proposed which was initially intended to handle the high dimensionality of the embedding space. However, open dimensions demonstrated superior performance to disabled dimensions on other classification tasks as well, which could benefit other AIS models. When employed with embeddings, however, some unexpected phenomena were observed, which indicate that further work is necessary to gain a more proper understanding of the dynamics and interactions of the open dimension type. Furthermore, local search produced superior results, particularly on more complex data sets. Adopting a local search operator may therefore improve many AIS models.

The results of exp. MR3 demonstrated that employing boosting in the AIS model yielded mixed outcomes on the benchmark classification sets. On one hand, boosting showed improved training accuracy, indicating its ability to effectively adapt and refine the immune system's response to fake news. However, the reduced testing accuracy suggested that overfitting occurred, where the model became overly specialised to the training data and struggled to generalise well to unseen instances.

Interestingly, when used in conjunction with high-dimensional sentence embeddings, boosting exhibited a distinct benefit. This finding might indicate that the method employed for leveraging embeddings was complemented by the boosting algorithm, mitigating the tendency towards overfitting observed on the benchmark sets.

While it does not serve as a direct contribution to the field, during the development of the model efficiency was an area of heavy focus. Accordingly, many optimisation areas of common AIS features were identified and implemented which should be transferable to other AIS models, thereby increasing their efficiency. The particulars of the implementations are beyond the scope of this thesis but can be found in the project's code repository.[1]

However, there are certain limitations to this study. The dynamics of the open dimension type were not fully explored, and a more rigorous understanding would likely result in a more effective classifier. Furthermore, local search was only implemented for the multiplier. Its beneficial effect here may not necessarily translate to other parts of the antibody. Boosting techniques were not fully explored either. Alternatives to the modified Adaboost scheme in the proposed model could prove more effective in avoiding overfitting.

Hopefully, the findings of this thesis bring the field one step closer to effectively combating the proliferation of fake news and promoting the dissemination of accurate and reliable information in an increasingly digital society.

6.3 Future Work

As the spread of fake news evolves, the main areas where people get their news are no longer discrete news articles but rather from a diverse array of sources, e.g. Twitter, TikTok and Facebook. The system described in this thesis may therefore not be applicable in the future. To keep pace with emerging fake news paradigms, an approach that leverages AISs to look at the spread of claims, and not just their contents would likely fare better.

Although the results did not reach the performance of Ellipsoidal-AIS [39], the proposed model boasts significantly greater efficiency. There might be room to amplify the model's performance by incorporating more attributes from Ellipsoidal-AIS, yet also focusing on their optimisation. For instance, redefining ellipsoidal rotations in the context of quaternions could potentially help tackle the extensive memory requirements tied to the rotational matrices, making it a more feasible alternative for an AIS working in high-dimensional environments.

[1] https://github.com/Tryxel-Industries/ais

6.3.1 Confidence of article sentence falsehood

The model could be refined to incorporate a system that annotates each sentence within an article with the model's confidence in its predictions. Such confidence scores could be derived from one of the mentioned evaluation methods. Given that the model separates articles into sentences for classification, this method of annotation could be adapted to accommodate datasets with multi-class labels, where individual sentences within an article could potentially be assigned different classes. This approach would offer system operators a more nuanced perspective on an article's authenticity, mirroring the real-world scenario where the truthfulness of information often extends beyond a simplistic binary false/true categorisation.

6.3.2 Grouping of similar antigens in the sentence embeddings

Approximately 10% of sentences across the datasets are alike among different articles. If a mechanism were developed to share a sentence antigen between multiple articles, and assign appropriate weights, an enhancement in computational efficiency could be expected. Such a mechanism could potentially reduce redundancy and optimise runtime performance.

6.3.3 Including the temporal aspect of the news

Incorporating the temporal dimension of news could allow the model to capture the evolving nature of fake news over time. This could allow the model to better account for the dynamic nature of information dissemination and the changes in language patterns that may occur. Furthermore, analysing temporal trends in fake news can help identify recurring patterns, topics, or narratives characteristic of deceptive information. By detecting and understanding these trends, the model can improve its ability to identify emerging fake news stories or variations of existing ones. This can be achieved by applying techniques such as time-series analysis or topic modelling to identify significant temporal patterns within the news data. By designing an approach leveraging the temporal component of news, the model could improve its performance over time and remain effective in the ever-changing landscape of deceptive information.

6.3.4 Memory cell boosting

The model proposed in this work draws from VALIS [26], and doesn't incorporate the use of memory cells. The incorporation of memory cells into the proposed model, following the example of CLONALG [8] and Ellipsoidal-AIS, could potentially enhance the model's precision and performance by preserving a set of appropriate antibodies associated with each antigen. Additionally, algorithms with memory cells might be more compatible with boosting techniques. Thus, the application of a boosting operator to such models could be another promising path for further investigation.

6.3.5 Exploration of the local search evaluator

The selection of the local search system's evaluation function was determined through initial testing. Delving deeper into the local search system could potentially yield substantial improvements to the model. For certain datasets where predicting the majority class proved problematic, employing a local search operator focused solely on identifying solutions with minimal errors may have mitigated the issue. However, it was deemed that exploring various evaluation types for the local search system was beyond the scope of this thesis.

6.3.6 Using general textual features and embeddings in concert for a more accurate fake news detector

Due to the limitations in time, an investigation into the optimal methods for integrating general textual features with news article embeddings was not conducted. Harnessing the strengths inherent in both approaches to create a comprehensive classification system is a promising path that may lead to a more effective classifier.

6.3.7 Feature selection for the general textual features

The thoroughness of the investigation and assessment of the semantic features was somewhat constrained by the time frame and focus of this study. Through more appropriate selection and refinement of the features utilised, considerable enhancements could be potentially achieved for the AIS model with general textual features.

Bibliography

[1] U. Aickelin, J. Greensmith, and J. Twycross, "Immune system approaches to intrusion detection – a review", in *Artificial immune systems: third international conference, ICARIS 2004, Catania, Sicily, Italy, September 13-16, 2004: proceedings*, 3239, 2004, pp. 328–341.

[2] S. Aldhaheri, D. Alghazzawi, L. Cheng, A. Barnawi, and B. A. Alzahrani, "Artificial immune systems approaches to secure the internet of things: A systematic review of the literature and recommendations for future research", *Journal of Network and Computer Applications*, vol. 157, p. 102 537, May 1, 2020.

[3] S. Atske. "News consumption across social media in 2021", Pew Research Center's Journalism Project. (Sep. 20, 2021), [Online]. Available: https://www.pewresearch.org/journalism/2021/09/20/news-consumption-across-social-media-in-2021/.

[4] E. Baug, P. Haddow, and A. Norstein, "MAIM: A novel hybrid bio-inspired algorithm for classification", in *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, Dec. 2019, pp. 1802–1809.

[5] S. Bradshaw, H. Bailey, and P. N. Howard, *Industrialized Disinformation: 2020 Global Inventory of Organized Social Media Manipulation.* Computational Propaganda Project at the Oxford Internet Institute, 2021.

[6] F. M. Burnet, *The clonal selection theory of acquired immunity.* Nashville: Vanderbilt University Press, 1959, 232 pp., Pages: 1-232.

[7] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter", in *Proceedings of the 20th international conference on World wide web*, Mar. 28, 2011, pp. 675–684.

[8] L. de Castro and F. Von Zuben, "Learning and optimization using the clonal selection principle", *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 3, pp. 239–251, Jun. 2002.

[9] H. Dai, Y. Yang, H. Li, and C. Li, "Bi-direction quantum crossover-based clonal selection algorithm and its applications", *Expert Systems with Applications*, vol. 41, no. 16, pp. 7248–7258, Nov. 15, 2014.

[10] B. Donnachie, J. Verrall, A. Hopgood, P. Wong, and I. Kennedy, "Accelerating cyber-breach investigations through novel use of artificial immune system algorithms", in *Artificial Intelligence XXXIX*, 2022, pp. 297–302.

[11] G. Dudek, “An artificial immune system for classification with local feature selection”, *IEEE Transactions on Evolutionary Computation*, vol. 16, no. 6, pp. 847–860, Dec. 2012.

[12] Y. Dun, K. Tu, C. Chen, C. Hou, and X. Yuan, “KAN: Knowledge-aware attention network for fake news detection”, *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 1, pp. 81–89, May 18, 2021.

[13] I. Dutt, S. Borah, and I. K. Maitra, “Immune system based intrusion detection system (IS-IDS): A proposed model”, *IEEE Access*, vol. 8, pp. 34 929–34 941, 2020.

[14] A. Eiben and J. Smith, *Introduction to Evolutionary Computing* (Natural Computing Series). Berlin, Heidelberg: Springer Berlin Heidelberg, 2015.

[15] K. Ethayarajh, “How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings”, in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 55–65.

[16] R. A. Fisher, *Iris*, 1936.

[17] G. Gravanis, A. Vakali, K. Diamantaras, and P. Karadais, “Behind the cues: A benchmarking study for fake news detection”, *Expert Systems with Applications*, vol. 128, pp. 201–213, Aug. 15, 2019.

[18] J. Greensmith and S. Cayzer, “An artificial immune system approach to semantic document classification”, vol. 2787, Sep. 1, 2003, pp. 136–146.

[19] E. Hart, “Not all balls are round: An investigation of alternative recognition-region shapes”, in *Artificial Immune Systems*, vol. 3627, 2005, pp. 29–42.

[20] B. Horne and S. Adali, “This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news”, *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, no. 1, pp. 759–766, May 3, 2017.

[21] Z. Ji and D. Dasgupta, “Real-valued negative selection algorithm with variable-sized detectors”, in *Genetic and Evolutionary Computation – GECCO 2004*, 2004, pp. 287–298.

[22] M. K. Johnson and C. L. Raye, “Reality monitoring”, *Psychological Review*, vol. 88, pp. 67–85, 1981.

[23] J. N. K, “Towards a network theory of the immune system”, *Ann.Immunol.*, vol. 125, pp. 373–389, 1974.

[24] M. Kahn, *Diabetes*, 1994.

[25] R. K. Kaliyar, A. Goswami, and P. Narang, “FakeBERT: Fake news detection in social media with a BERT-based deep learning approach”, *Multimedia Tools and Applications*, vol. 80, no. 8, pp. 11 765–11 788, Mar. 1, 2021.

[26] P. Karpov and G. Squillero, "VALIS: An evolutionary classification algorithm", *Genetic Programming and Evolvable Machines*, vol. 19, no. 3, pp. 453–471, Sep. 2018.

[27] J. Y. Khan, M. T. I. Khondaker, S. Afroz, G. Uddin, and A. Iqbal, "A benchmark study of machine learning models for online fake news detection", *Machine Learning with Applications*, vol. 4, p. 100 032, Jun. 15, 2021.

[28] B. Li, H. Zhou, J. He, M. Wang, Y. Yang, and L. Li, "On the sentence embeddings from pre-trained language models", in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Nov. 2020, pp. 9119–9130.

[29] D. Li, M. Gu, S. Liu, X. Sun, L. Gong, and K. Qian, "Continual learning classification method with the weighted k-nearest neighbor rule for time-varying data space based on the artificial immune system", *Knowledge-Based Systems*, vol. 240, p. 108 145, Mar. 15, 2022.

[30] J.-q. Li, Z.-m. Liu, C. Li, and Z.-x. Zheng, "Improved artificial immune system algorithm for type-2 fuzzy flexible job shop scheduling problem", *IEEE Transactions on Fuzzy Systems*, vol. 29, no. 11, pp. 3234–3248, Nov. 2021.

[31] S. Lotfi, M. Mirzarezaee, M. Hosseinzadeh, and V. Seydi, "Detection of rumor conversations in twitter using graph convolutional networks", *Applied Intelligence*, vol. 51, no. 7, pp. 4774–4787, Jul. 1, 2021.

[32] R. R. Mandical, N. Mamatha, N. Shivakumar, R. Monica, and A. N. Krishna, "Identification of fake news using machine learning", in *2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, Jul. 2020, pp. 1–6.

[33] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The stanford CoreNLP natural language processing toolkit", in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2014, pp. 55–60.

[34] J. S. Marshall, R. Warrington, W. Watson, and H. L. Kim, "An introduction to immunology and immunopathology", *Allergy, Asthma & Clinical Immunology*, vol. 14, no. 2, p. 49, Sep. 12, 2018.

[35] S. A. McCornack, K. Morrison, J. E. Paik, A. M. Wisner, and X. Zhu, "Information manipulation theory 2", *Journal of Language and Social Psychology*,

[36] C. McEwan and E. Hart, "Representation in the (artificial) immune system", *J. Math. Model. Algorithms*, vol. 8, pp. 125–149, Jun. 1, 2009.

[37] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality", in *Advances in Neural Information Processing Systems*, vol. 26, 2013.

[38] M. Nirav Shah and A. Ganatra, “A systematic literature review and existing challenges toward fake news detection models”, *Social Network Analysis and Mining*, vol. 12, no. 1, p. 168, Nov. 14, 2022.

[39] S. Ozsen and C. Yucelbas, “On the evolution of ellipsoidal recognition regions in artificial immune systems”, *Applied Soft Computing*, vol. 31, pp. 210–222, Jun. 2015.

[40] G. Pennycook and D. G. Rand, “The psychology of fake news”, *Trends in Cognitive Sciences*, vol. 25, no. 5, pp. 388–402, May 1, 2021.

[41] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, *Automatic detection of fake news*, Aug. 23, 2017. arXiv: 1708.07104[cs].

[42] M. Ph. D., A. Keya, M. A. Hamid, M. M. Monowar, and M. S. Rahman, “A comprehensive review on fake news detection with deep learning”, *IEEE Access*, vol. PP, pp. 1–1, Nov. 18, 2021.

[43] H. T. Phan, N. T. Nguyen, and D. Hwang, “Fake news detection: A survey of graph neural network methods”, *Applied Soft Computing*, vol. 139, p. 110 235, May 1, 2023.

[44] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein, *A stylometric inquiry into hyperpartisan and fake news*, Feb. 18, 2017. arXiv: 1702.05638[cs].

[45] N. Rai, D. Kumar, N. Kaushik, C. Raj, and A. Ali, “Fake news classification using transformer based enhanced LSTM and BERT”, *International Journal of Cognitive Computing in Engineering*, vol. 3, pp. 98–105, Jun. 1, 2022.

[46] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, “Truth of varying shades: Analyzing language in fake news and political fact-checking”, in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Sep. 2017, pp. 2931–2937.

[47] V. Raunak, V. Gupta, and F. Metze, “Effective dimensionality reduction for word embeddings”, in *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, Aug. 2019, pp. 235–243.

[48] B. Rizvi, A. Belatreche, and A. Bouridane, “A dendritic cell immune system inspired approach for stock market manipulation detection”, in *2019 IEEE Congress on Evolutionary Computation (CEC)*, Jun. 2019, pp. 3325–3332.

[49] Y. M. Rocha, G. A. de Moura, G. A. Desidério, C. H. de Oliveira, F. D. Lourenço, and L. D. de Figueiredo Nicolete, “The impact of fake news on social media and its influence on health during the COVID-19 pandemic: A systematic review”, *Journal of Public Health*, Oct. 9, 2021.

[50] V. L. Rubin, “On deception and deception detection: Content analysis of computer-mediated stated beliefs”, *Proceedings of the American Society for Information Science and Technology*, vol. 47, no. 1, pp. 1–10, 2010.

[51] S. J. Russell, P. Norvig, M.-w. Chang, *et al.*, *Artificial intelligence: a modern approach* (Pearson series in artificial intelligence), Fourth edition, global edition. Harlow: Pearson, 2022, 1166 pp.

[52] A. J. Saleh, A. Karim, B. Shanmugam, *et al.*, “An intelligent spam detection model based on artificial immune system”, *Information*, vol. 10, no. 6, p. 209, Jun. 2019.

[53] M. Shivers, C. Llanes, and M. Sherman, “Implementation of an artificial immune system to mitigate cybersecurity threats in unmanned aerial systems”, in *2019 IEEE International Conference on Industrial Internet (ICII)*, Nov. 2019, pp. 12–17.

[54] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, *FakeNewsNet: A data repository with news content, social context and spatialtemporal information for studying fake news on social media*, Mar. 27, 2019. arXiv: 1809.01286[cs].

[55] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, *Fake news detection on social media: A data mining perspective*, Sep. 2, 2017. arXiv: 1708.01967[cs].

[56] K. Shu, S. Wang, and H. Liu, “Beyond news contents: The role of social context for fake news detection”, in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, Jan. 30, 2019, pp. 312–320.

[57] S. Somasundaran, J. Ruppenhofer, and J. Wiebe, “Detecting arguing and sentiment in meetings”, in *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, 2007, pp. 26–34.

[58] M. F. Stefan Aeberhard, *Wine*, 1992.

[59] J. Su, J. Cao, W. Liu, and Y. Ou, *Whitening sentence representations for better semantics and faster retrieval*, Mar. 28, 2021. arXiv: 2103.15316[cs].

[60] S. Sverdrup-Thygeson, “An artificial immune system for fake news classification”, p. 148, 2021.

[61] R. G. Terry Sejnowski, *Connectionist bench (sonar, mines vs. rocks)*, 1988.

[62] Y.-W. Tseng, H.-K. Yang, W.-Y. Wang, and W.-C. Peng, “KAHAN: Knowledge-aware hierarchical attention network for fake news detection on social media”, in *Companion Proceedings of the Web Conference 2022*, Aug. 16, 2022, pp. 868–875.

[63] J. Twycross and S. Cayzer, “An immune-based approach to document classification”, in *Intelligent Information Processing and Web Mining*, 2003, pp. 33–46.

[64] U. Undeutsch, “Beurteilung der glaubhaftigkeit von aussagen”, *Handbuch der psychologie*, vol. 11, S 126 1967.

[65] S. W. V. Sigillito, *Ionosphere*, 1989.

[66] P. K. Verma, P. Agrawal, V. Madaan, and R. Prodan, “MCred: Multi-modal message credibility for fake news detection using BERT and CNN”, *Journal of Ambient Intelligence and Humanized Computing*, Jul. 27, 2022.

[67] S. Vosoughi, D. Roy, and S. Aral, “The spread of true and false news online”, *Science*, vol. 359, no. 6380, pp. 1146–1151, Mar. 9, 2018.

[68] W. Y. Wang, *"liar, liar pants on fire": A new benchmark dataset for fake news detection*, May 1, 2017. arXiv: 1705.00648[cs].

[69] Y. Wang and T. Li, "Local feature selection based on artificial immune system for classification", *Applied Soft Computing*, vol. 87, p. 105 989, Feb. 1, 2020.

[70] A. Watkins, J. Timmis, and L. Boggess, "Artificial immune recognition system (AIRS): An immune-inspired supervised learning algorithm", *Genetic Programming and Evolvable Machines*, vol. 5, no. 3, pp. 291–317, Sep. 1, 2004.

[71] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification", in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 1480–1489.

[72] H. Zhang, H. Xiao, S. Liu, *et al.*, "A relation b-cell network used for data identification and fault diagnosis", *Applied Soft Computing*, vol. 113, p. 107 921, Dec. 1, 2021.

[73] Y. Zhong, L. Zhang, B. Huang, and P. Li, "An unsupervised artificial immune classifier for multi/hyperspectral remote sensing imagery", *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 44, pp. 420–431, Mar. 1, 2006.

[74] X. Zhou, A. Jain, V. V. Phoha, and R. Zafarani, "Fake news early detection: A theory-driven model", *Digital Threats: Research and Practice*, vol. 1, no. 2, pp. 1–25, Jun. 30, 2020.

[75] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities", *ACM Computing Surveys*, vol. 53, no. 5, 109:1–109:40, Sep. 28, 2020.

[76] M. Zuckerman, B. M. DePaulo, and R. Rosenthal, "Verbal and nonverbal communication of deception", in *Advances in Experimental Social Psychology*, vol. 14, Jan. 1, 1981, pp. 1–59.

NTNU

Norwegian University of
Science and Technology