Otto-von-Guericke University Magdeburg

Faculty of Computer Science
Department of Digital Engineering

Master Thesis

Development of a Framework for the Extraction of Representative but Diverse Subset of Instances from a Dataset with Respect to an Instance of Interest

Author:
Stanley Chukwuemeka Umeh

November 04, 2023

Examiners:

First Examiner
Prof. Sylvia Saalfeld

Ilmenau University of Technology
Faculty of Computer Science and Automation
Ehrenbergstraße 29
98693 Ilmenau, Germany

Second Examiner
Prof. Christian Hansen

Faculty of Computer Science
Otto-von-Guericke University
Universitätsplatz 2
39106 Magdeburg, Germany

Umeh, Stanley Chukwuemeka:
Development of a Framework for the Extraction of Representative but Diverse Subset of Instances from a Dataset with Respect to an Instance of Interest
Master Thesis, Otto-von-Guericke University
Magdeburg, 2023.

Contents

Abstract

1 Introduction and Motivation
1.1 Motivation . . . 1
1.2 Aim of this thesis . . . 3
1.3 Structure of this thesis . . . 4

2 Background
2.1 Medical Background on Intracranial Aneurysms . . . 5
2.2 Background on Machine Learning Techniques . . . 9
2.2.1 Clustering . . . 10
2.2.2 Outlier Detection . . . 12
2.2.3 Instance Selection . . . 15

3 Related Work
3.1 Aneurysms . . . 17
3.2 Aneurysm Training Simulations . . . 20
3.3 Instance Extraction . . . 21

4 Methods
4.1 Problem Definition . . . 29
4.2 Concept . . . 29
4.3 Datasets . . . 30
4.3.1 Data Cleaning and Pre-processing . . . 32
4.4 Extraction Methodology . . . 39
4.4.1 Outlier Removal . . . 40
4.4.2 Clustering . . . 41
4.4.3 Prototyping . . . 46
4.5 Proposed Evaluation Approach . . . 47
4.5.1 Metrics . . . 47
4.5.2 Score . . . 49

5 Experiments and Evaluation
5.1 RIS Evaluation . . . 51
5.1.1 Main Dataset Evaluation . . . 52

5.1.2 Supplementary Dataset Evaluation 53
5.2 Instance Selection Adaptation Evaluation 56

6 Analysis of Results
6.1 RIS 59
6.1.1 Main Dataset 59
6.1.2 Supplementary Dataset 69
6.1.3 Summary of Deductions from Experiments 78
6.1.4 Challenges 79
6.2 Instance Selection Adaptation 80
6.3 Discussion of Research Questions 82
6.3.1 RQ 1. How do we define an extraction technique for the task of extracting representative but diverse samples? 82
6.3.2 RQ 2. How is representative but diverse defined? 83
6.3.3 RQ 3. What metrics will be used to evaluate the extracted set? 83
6.3.4 RQ 4. How do we distinguish between diverse cases and outliers? 84

7 Conclusions and Future Work
7.1 Conclusions 87
7.2 Limitations 88
7.3 Future Work 89

A List of Figures

B List of Tables

C Bibliography

Abstract

The aim of this thesis is to develop a framework for extracting a subset of instances from a given dataset or database that is similar to a given instance of interest but also diverse from each other. A framework like this will be very useful in the medical field for extracting instances for training simulations. Training simulation is a popular approach used, especially in surgical medicine, where doctors practice their skills by performing preparatory surgical operations before embarking on a human. Training simulations are also a proven way to expose and teach medical students the intricacies of their potential real-life patient encounters in various areas of medicine. It is also useful to stay sharp and prepared for surgeries, which are rarely performed, like surgical clipping of intracranial aneurysms. Hence, my objective is to develop this framework for extracting representative but diverse subsets using a dataset of intracranial aneurysms.

Although this task is similar to instance selection, which aims to extract a smaller samples from a dataset that effectively conveys all the information in the larger dataset, the introduction of an instance of interest makes this a novel task. For this reason, it is also necessary to develop methodologies for evaluating the quality of an extracted subset with respect to an instance of interest. Therefore, in this thesis I proposed a $ReverseInstanceSelection$ framework to extract instances in a bid to achieve this goal and I also developed a sequence of mathematical equations to measure the quality of an extracted subset with respect to a given instance. I also proposed an adaptation of this framework to cater for generic instance selection tasks.

The proposed framework consists of three major phases: outlier removal, clustering, and prototyping. Extensive experiments and analysis with various machine learning algorithms were performed on the given datasets to support the choices made for each phase of the framework. An evaluation of the proposed framework showed that while it is strongest for extracting smaller subsets of instances, the strength starts to dwindle as the size of an extracted set increases.

The proposed framework and equations are reusable and adaptable to other datasets in any domain, provided adequate analysis is performed to determine the appropriate choice of unsupervised algorithm to be used in each step of the framework, their hyper-parameters, and proper weights are assigned to each metric in the final equation.

Acknowledgements

I would like to express my heartfelt gratitude to my supervisors, Prof. Sylvia Saalfeld and Lena Spitz, for their tremendous assistance and support during this thesis. Their recommendations and views were invaluable in refining my study, and collaborating with them has been a really gratifying experience.

I am especially grateful to Lena Spitz for her constant support, prompt replies, and supply of necessary resources, which aided this work. I also extend my thanks to the collaborating medical doctors for making time to interact with me when necessary during the course of this work.

Finally, I am grateful to my friends and family for their unwavering encouragement and support throughout this program and the accompanying external challenges I faced during the course of this study. Their presence in my life is a continual source of inspiration and motivation for me, and I consider myself tremendously blessed to have them.

1 Introduction and Motivation

1.1 Motivation

An intracranial aneurysm (IA) is an atypical ballooning of a cerebral artery in a specific area due to weakened vessel walls. It is prevalent in about 2-5% of the population (BROWN und BRODERICK (2014); LIU et al. (2022); VERNOOIJ et al. (2007)), and is the primary cause of non-traumatic subarachnoid hemorrhage (SAH) when it ruptures (XU et al. (2019)). SAH resulting from aneurysm rupture is a severe neurological condition with high mortality and morbidity rates. Despite technological advancements in treatment and imaging, the mortality rate of SAH is between 27 - 50%, with approximately half of the survivors experiencing permanent neurological impairments (ETMINAN et al. (2019); NIEUWKAMP et al. (2009); ROKED und REDDY (2020)). As a result, early diagnosis and assessment of aneurysms are crucial for the treatment and prognosis of patients.

Technological advancements in medical imaging have improved diagnosis of IA's, which has also led to increase in data for IA-related research using machine learning (ML) techniques (ALWALID et al. (2022); MAUPU et al. (2022)). Some examples of work being done using ML techniques include rupture risk prediction (SPITZ et al. (2020)), discriminating feature analysis (TANG et al. (2022)), subgroup selection (RYTTLEFORS et al. (2008)), etc., which will be discussed extensively in the following chapters of this work. While some approaches have used a combination of hemodynamic, morphological, patient-specific features, etc. (AN et al. (2022); DETMER et al. (2019); TANIOKA et al. (2020) others used just one class of features (ABBOUD et al. (2017); DHAR et al. (2008)) or a mixture of both to build ML approaches to IA. Morphological features provide insights into the structural characteristics of the aneurysm, hemodynamic features provide informa-

tion about the blood flow dynamics that can affect its stability. A popular approach that uses patient-specific parameters is the PHASES score model (BACKES et al. (2015)), it leverages parameters such as population, hypertension, age, size of aneurysm, earlier SAH from another aneurysm, the site of the aneurysm to determine the rupture risk of an IA.

Subgroup selection approaches entail selecting a group of aneurysms similar to an aneurysm of interest (AOI). Presenting physicians with these results can aid clinical decisions as information from the selected samples can be leveraged to make decisions on the AOI. This approach can be extended to develop classification models such as case-based learning. Subgroup selection using similarity can be very useful to make preliminary decisions with respect to AOI's, but it can also be interesting to introduce diversity into the selections as closely similar instances might be lacking variance in the selections which can be useful for the medical training simulations.

Training simulation is an important aspect of surgical medicine, it provides an avenue to train students and young surgeons with useful experience before working on patients. It is also beneficial for experienced experts, to keep them sharp, especially for rare illnesses like IA's, because surgeons like every human can become deficient without practice (AGHA und FOWLER (2015)). Training simulations are an avenue for medical practitioners to engage in intentional practice, which helps in refining their skills and improving outcomes in real life situations (GORDON (2000)). There is a lot of research that emphasizes this importance in various use cases (AGHA und FOWLER (2015); ALLGAIER et al. (2022); SEIL et al. (2022)). In these training simulations, users may want to specialize their training to a set of very similar instances, or it may also be necessary to experiment with instances that have some diversity, which would be useful to practice how to navigate the nuances surrounding a particular type of case. While for the former, extracting the most similar instances with respect to a given AOI will be sufficient, that is not the case for the later. This is the problem I intend to tackle in this thesis.

Assuming we have a database of 5 aneurysms with 4 numerical features, given an AOI, we select three most similar IA's using the sum of absolute difference between each feature as a similarity measure. Table 1.1 shows the absolute difference per feature with respect to an AOI. Given the stated

condition, IA's 1, 3, and 5 will be selected as they have the lowest sum of differences, but it can also be interesting to consider IA 4, given that it is exactly the same in three features out of four, despite having the worst similarity score. This is the reason this work focuses on introducing the notion of diversity in the extraction process, I try to cover potential variances likely to be missed when extracting a subset solely based on similarity. Subsets extracted in this way can also be useful for training simulations among medical practitioners, where they try to ascertain the differences of similarities present in a set of aneurysms.

Table 1.1: Absolute Differences between AOI and Database of IA

IA ID	**Feature 1**	**Feature 2**	**Feature 3**	**Feature 4**	**Sum of Differences**
1	5	2	3	0	10
2	6	2	0	3	11
3	5	0	5	0	10
4	15	0	0	0	15
5	1	4	3	1	9

Introducing diversity into the selection can be modeled as an instance selection (IS) task. While the conventional IS tries to select the most representative samples that convey the most information about a larger database, what I am trying to do is to select instances that will convey the most information with respect to one sample. After extensive research in existing literature, there are no publications on IA's or other areas that approach the task in this manner. This would be called Reverse Instance Selection (RIS) for the rest of this work.

1.2 Aim of this thesis

There are existing techniques for IS that have been shown to select subset of instances that sufficiently explain the variance of a larger dataset or database to a reasonable extent. The aim of this thesis is to leverage the existing research to develop an extraction technique that introduces diversity to the selection of instances similar to an AOI. The idea is that selecting subgroups using just a similarity metric can lead to redundancy in the selection, and thus these selections might be lacking diversity, which may

be important for clinical training simulations. To achieve this task, I have formulated the following research questions:

RQ1 How do we define an extraction technique for this task?

RQ2 How is representative but diverse defined?

RQ3 What metrics will be used to evaluate the extracted set?

RQ4 How do we distinguish between diverse cases and outliers?

1.3 Structure of this thesis

The thesis outlines are as follows: Chapter 1, as discussed here, is an introduction to the work and the research questions I intend to answer, then chapter 2 gives the background and context of this work, such as intracranial aneurysms, their treatment procedures, etc., and machine learning approaches such as instance selection, clustering and outlier detection. Chapter 3 contains related works, a summary of the current approaches surrounding machine learning solutions to support IA's, instance selection, and the foundations of techniques explored to develop the RIS framework. Chapter 4 contains the concept, methodologies, and experiments for the proposed RIS framework. Chapter 5 contains the experimental setup for evaluation and the evaluation results for this novel approach. In chapter 6, I discuss and analyze the implications and meanings of the results in depth, the challenges faced in the course of evaluation and discussion of the results with respect to the research questions posed here. Finally, in chapter 7, conclusions are made at the completion of the work, I explore the limitations of the work and future works are extensively discussed.

2
Background

2.1 Medical Background on Intracranial Aneurysms

IA's are abnormal enlargements or bulges in the cerebral artery walls that may result in the development of a weak spot prone to rupture. The danger associated with this illness is high since aneurysm rupture can have devastating effects, such as subarachnoid hemorrhage (SAH) and neurological impairments. Therefore, effective management of cerebral aneurysms depends on early discovery, precise diagnosis, and appropriate therapeutic approaches. There are several arteries in the brain, any of which could have an IA. Figure 2.1 is a picture of the circle of Willis and its surrounding arteries.

IA's are often asymptomatic until rupture occurs. However, unruptured aneurysms may present with symptoms related to their mass effect, such as headaches, visual disturbances, cranial nerve palsies, or seizures. When an aneurysm ruptures, it leads to SAH, characterized by a sudden severe headache, neck stiffness, altered consciousness, and, in severe cases, coma (KEEDY (2006); TOTH und CEREJO (2018)) and death with a mortality rate of 27% - 50% (ROKED und REDDY (2020); STIENEN et al. (2018).

Aneurysms in the intracranial region are thought to affect 2-5% of people (BROWN und BRODERICK (2014); LI et al. (2022)). The prevalence rises with age and is slightly higher in women (FRÉNEAU et al. (2022)). IA's can form as a result of a number of risk factors, such as genetic susceptibility, smoking, hypertension, family history, connective tissue abnormalities, and specific systemic diseases.

The diagnosis and detection of cerebral aneurysms depends heavily on imaging. Non-invasive techniques that provide detailed vascular imaging

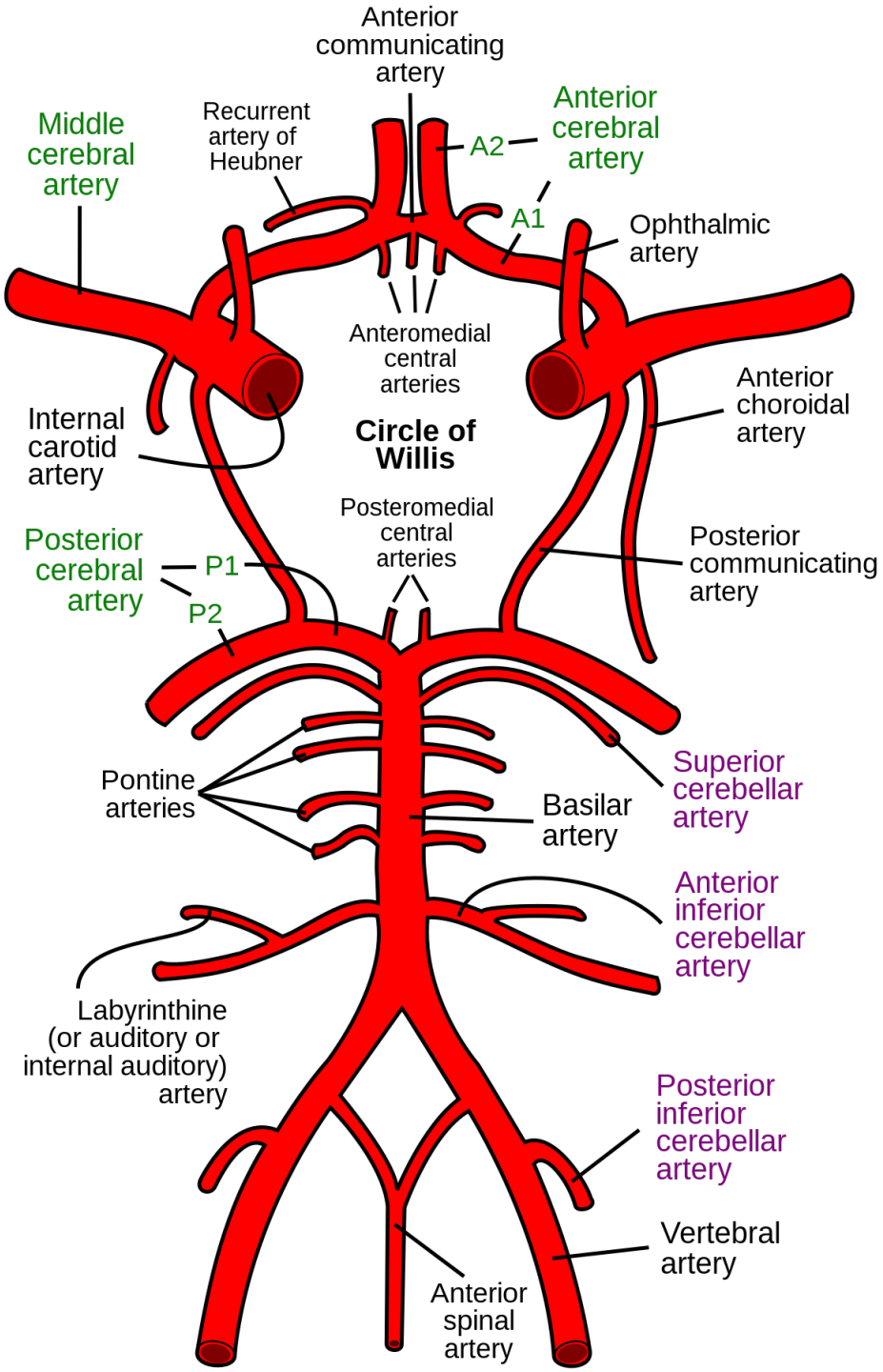

Figure 2.1: The circle of willis and surrounding arteries(FLANAGAN et al. (2015))

and help identify aneurysms include computed tomography angiography (CTA) and magnetic resonance angiography (MRA). THAKER et al. (2012) and GÖLITZ et al. (2014) show that digital subtraction angiography (DSA) is still the best method for accurately confirming and assessing aneurysm anatomy. The accuracy of diagnosis has increased thanks to developments in imaging technology including 3D rotational angiography (3DRA) and high-resolution imaging.

The management of IA aims to prevent rupture and subsequent hemorrhage. Treatment options include both surgical and endovascular techniques. Neurosurgical surgical treatments (NST), such as clipping and bypass procedures, involve accessing the aneurysm directly and securing it with a clip or graft. Endovascular treatment (EVT) approaches, such as coiling and flow diverters, involve navigating catheters and deploying devices to promote aneurysm occlusion. The choice of treatment depends on various factors, including aneurysm characteristics, location, patient age, and comorbidities. Figure 2.2 shows a pictorial representation of these approaches. Endovascular coiling is a minimally invasive procedure where platinum coils are inserted into an intracranial aneurysm via a catheter, promoting blood clotting and reducing the risk of rupture while neurosurgical clipping is a traditional surgical procedure involving the placement of a metal clip on the neck of an intracranial aneurysm to block blood flow and prevent rupture (BELAVADI et al. (2021), LINDGREN et al. (2018)).

Although JUVELA et al. (2013) points out that only 1-2% of unruptured IA's (UIA) will rupture, the importance of rupture risk assessment and analysis of IA's can not be overemphasised because it is important for treatment decision making. Medical practitioners constantly have to weigh the advantages and risks between treating an aneurysm and the potential rupture for each case, this is because even though its important and life saving, treatment of IA's can also present significant complications. Popular IA treatment procedures are EVT and NST, with each having their unique potential complications. Complications for EVT can be categorized into three: intraprocedural aneurysm rupture (IAR), thromboembolism (TE), and post-procedural early rebleeding (PER) (AHN et al. (2017); IHN et al. (2018)). Most poplar types of complications are the IAR's occurring in an estimated 1% to 5% of cases (BRISMAN et al. (2005); PIEROT et al. (2008))

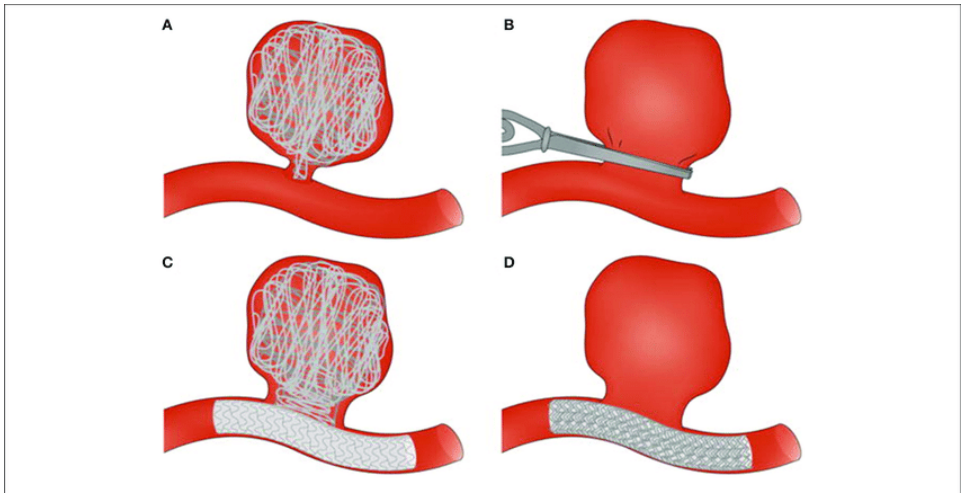

Figure 2.2: Endovascular and surgical treatments for IA. (A) Endovascular coiling of the aneurysm sac. (B) Surgical clipping of the aneurysm neck. (C) Endovascular treatment using coils and a stent. (D) Endovascular treatment using flow diverter. (PERRONE et al. (2015))

and TE reportedly within the range of 2% to 15% (IM et al. (2009); OISHI et al. (2012); PARK et al. (2005)). PER usually occurs in less than 1% of cases (AHN et al. (2017)).

AHN et al. (2017) analysed a database of 436 cases of saccular IA's between 2007 - 2015 treated with endovascular coiling and showed complications occurred in 61 cases (14%). MCLAUGHLIN und BOJANOWSKI (2004), conducted a on 143 patients treated using neurosurgical clipping over a 3 year period and found 29 patients (20.3%) suffered complications. Complications for these studies were defined using the Glasgow outcome scale (JENNETT et al. (1981)) scores.

Morphological features of aneurysms refer to the structural characteristics and appearance of the aneurysm. These features provide information about the shape, size, and location of the aneurysm. Some key morphological features include; orthogonal height, maximum diameter, aspect ratio, undulation index, etc. Hemodynamic features of aneurysms pertain to the blood flow patterns and forces acting within the aneurysm. Understanding the hemodynamics of an aneurysm is essential for evaluating its rupture risk. Some important hemodynamic features include; wall shear stress (WSS), flow velocity, pressure.

The evaluation of both morphological and hemodynamic features is crucial for assessing the risk of aneurysm rupture. Combining these features enables a more comprehensive understanding of aneurysm behavior and aids in clinical decision-making.

The PHASES score model (BACKES et al. (2015)) which is currently used by most medical facilities to determine the rupture risk of IA uses a few patient-specific parameters (population, hypertension, age, size of aneurysm, earlier SAH from another aneurysm, site of aneurysm). This does not leverage the rich vein of morphological and hemodynamic parameters which can improve the process of treatment decision making and estimating potential for aneurysm rupture (DETMER et al. (2019); DHAR et al. (2008); NIEMANN et al. (2018); XIANG et al. (2011)).

2.2 Background on Machine Learning Techniques

A branch of artificial intelligence known as "machine learning" focuses on creating algorithms and models that let computers learn from data and make predictions or judgments without having to be explicitly programmed. To enable systems to automatically learn from experience and improve, it requires the study of statistical techniques and computer models.

Instead of being explicitly written, machine learning algorithms understand patterns and correlations in data by examining samples. In order to find patterns and create a model, the algorithm uses training data, which comprises of input features and their corresponding desired outputs or labels (if available) (MITCHELL et al. (2007)).

Machine learning can be categorized into different types. In supervised learning, the algorithm learns from labeled data, where the desired outputs are provided (BURKART und HUBER (2021)). Unsupervised learning involves discovering patterns and structures in unlabeled data without explicit outputs (HAHNE et al. (2008)). Semi-supervised learning is a combination of supervised and unsupervised learning, where the algorithm learns from a small amount of labeled data along with a large amount of unlabeled data (VAN ENGELEN und HOOS (2020)). Reinforcement learning

focuses on an agent learning to interact with an environment to maximize a reward signal (KAELBLING et al. (1996)).

To learn the underlying patterns and correlations, machine learning models are trained using training data. The algorithm modifies its internal parameters based on optimization strategies during the training process to reduce errors or increase performance. Then test data are used to evaluate the trained models' generalization and prediction skills. Creating models that are good at generalizing to new data is one of the main objectives of machine learning (TAN et al. (2018)).

The books MITCHELL et al. (2007); TAN et al. (2018) covers several machine learning concepts and approaches in depth. For the rest of this chapter we give a brief overview of machine learning approaches and algorithms explored for this research.

2.2.1 Clustering

Clustering is a fundamental technique in unsupervised machine learning used to group similar data points together based on their inherent patterns, similarities, or proximity. It aims to identify meaningful structures within unlabeled data without any prior knowledge or guidance (MADHULATHA (2012)).

Clustering finds applications in various domains, including data analysis (DUBES und JAIN (1980)), customer segmentation (WU und LIN (2005)), image processing (DEHARIYA et al. (2010)), anomaly detection (PU et al. (2020)), and recommendation systems (AHUJA et al. (2019)). It enables exploratory data analysis, pattern discovery, and grouping similar instances together without any prior knowledge about the data.

It's critical to remember that clustering is an exploratory process, and the interpretation of the clusters requires human judgment and domain expertise in order to give the results significance.

K-Means

K-means clustering is a popular unsupervised machine learning algorithm used for partitioning data points into distinct groups or clusters. It is widely employed in various fields, including data analysis, pattern

recognition, image segmentation, and customer segmentation. The algorithm aims to minimize the intra-cluster variance while maximizing the inter-cluster variance, resulting in well-separated clusters (JIN und HAN (2010); MACQUEEN (1967)).

K-means clustering is sensitive to the initial selection of centroids, as it can converge to suboptimal solutions. To mitigate this, the algorithm is often run multiple times with different initializations, and the clustering solution with the lowest overall within-cluster variance is selected.

K-means clustering has several advantages, including simplicity, efficiency, and scalability, making it suitable for large datasets. However, it also has some limitations. The algorithm assumes that the clusters are spherical and of similar size, which may not hold true for all datasets. It is also sensitive to outliers, and the determination of the optimal number of clusters (k) can be challenging.

Overall, K-means clustering is a powerful technique for discovering inherent patterns and structures in unlabeled data, providing insights and enabling further analysis in various domains.

DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a popular density-based clustering algorithm used to group data points based on their density and proximity. Unlike K-means, DBSCAN does not require the number of clusters to be predefined and can discover clusters of arbitrary shapes. It is particularly effective when dealing with datasets containing clusters of different densities or with noise/outliers (SCHUBERT et al. (2017)). The main parameters of DBSCAN are: epsilon (ϵ), representing the maximum distance between two points to consider them as neighbors, and minPts, the minimum number of points required to form a dense region.

DBSCAN provides a number of advantages. It can find clusters of all forms and sizes and is not affected by the original point setup. It can successfully find outliers in datasets with varied densities. However, choosing optimal parameter values (ϵ and $minPts$) can be difficult, and because to the curse of dimensionality, the method may struggle with high-dimensional datasets.

To summarize, DBSCAN is a robust and adaptable clustering algorithm capable of detecting clusters based on density and proximity. It is well-suited for datasets with different densities and complicated structures, making it a useful tool for exploratory data analysis, anomaly detection, and pattern discovery in different domains.

OPTICS

OPTICS (Ordering Points To Identify Clustering Structure) is a density-based clustering technique that builds on DBSCAN ideas. It gives a hierarchical representation of the data's clustering structure and more flexibility in recognizing clusters of varied densities. OPTICS solves some of DBSCAN's drawbacks, such as the requirement to establish a specified distance threshold (ANKERST et al. (1999)).

OPTICS has a number of benefits over DBSCAN. It can handle datasets with varied densities of clusters and does not require a preset distance threshold to be established. The clustering structure's hierarchical representation provides for a more thorough comprehension of the data. However, OPTICS may be computationally costly, especially for big datasets, and parameter selection (such as $minPts$) is still important for achieving the best results.

OPTICS is a density-based clustering method that gives a hierarchical representation of clusters based on density connection. It allows for greater flexibility in detecting clusters of varied densities and allows for more in-depth investigation of the clustering structure. OPTICS is especially beneficial in applications where cluster density changes.

2.2.2 Outlier Detection

Outlier detection, also known as anomaly detection, is a technique used to identify data points or instances that deviate significantly from the norm or the majority of the data (HAWKINS (1980)). Outliers are observations that exhibit unusual behavior, differ significantly from the expected patterns, or represent rare events in the dataset.

Outlier detection finds applications in various fields, including fraud detection, network intrusion detection, sensor data analysis, health mon-

itoring, and anomaly-based intrusion detection systems. By identifying unusual or anomalous instances, outlier detection helps in identifying potential problems, anomalies, or outliers that require further investigation or intervention.

Outlier detection is a challenging task as the definition of outliers and the appropriate detection method may vary depending on the context, domain, and dataset. Careful consideration of the data characteristics and domain knowledge is essential to select the most suitable technique for outlier detection.

DBSCAN Based Approach

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is primarily a clustering algorithm but can also be used for outlier detection. DBSCAN detects outliers as data points that do not belong to any cluster or form sparse regions in the dataset. DBSCAN identifies outliers based on the concept of density. Density refers to the closeness of data-points in a specific area, this can vary from area to area in a dataset. DBSCAN classifies data points as core points, border points, or outliers. Core points are densely surrounded by other points, while border points are on the outskirts of dense regions. Outliers are points that do not meet the density requirements to be classified as core or border points (SCHUBERT et al. (2017)).

DBSCAN's ability to identify outliers stems from its density-based nature. It can effectively detect outliers that form sparse regions or do not fit well into any cluster. However, it is important to note that DBSCAN's primary focus is clustering, and its outlier detection capabilities may not be as robust or specialized as dedicated outlier detection algorithms.

When using DBSCAN for outlier detection, it is crucial to interpret the results carefully, considering the specific characteristics of the dataset and the domain knowledge. DBSCAN can be a valuable tool for identifying outliers in datasets with varying densities and complex structures, providing insights into anomalous or sparse data regions.

Autoencoders

Autoencoders are neural network models that can be utilized for outlier detection. Originally designed for dimensionality reduction and data reconstruction tasks, autoencoders can also be leveraged to identify anomalies or outliers in datasets (GOODFELLOW et al. (2016)).

Autoencoders offer a flexible and data-driven approach to outlier detection. They can capture complex patterns and non-linear relationships in the data, enabling the detection of subtle anomalies. However, it's important to note that autoencoders are sensitive to the choice of architecture, hyperparameters, and the quality and representativeness of the training data.

Autoencoders are just one approach among various outlier detection techniques. Their effectiveness depends on the specific characteristics of the dataset and the ability of the model to capture the normal patterns while discerning outliers. Careful training, validation, and interpretation of results are crucial for successful outlier detection using autoencoders.

Isolation Forest

Isolation Forest is an ensemble-based outlier detection algorithm that uses an innovative approach based on random forests (LIU et al. (2012)). It is designed to efficiently identify anomalies or outliers in datasets, particularly in high-dimensional spaces.

Isolation Forest has several advantages, it can handle high-dimensional data efficiently, as it randomly selects features for splitting, thereby avoiding the curse of dimensionality. It does not require any assumptions about the data distribution and is less sensitive to the presence of irrelevant attributes. The algorithm can also provide an anomaly score, which represents the normalized average path length. This score can be used to rank instances according to their abnormality level.

Isolation Forest is a popular choice for outlier detection due to its ability to handle high-dimensional data, its efficiency, and its ability to capture anomalies by isolating them in the trees. However, it is important to note that like any algorithm, Isolation Forest has its limitations and may not be optimal for all types of datasets or outlier patterns. Appropriate parame-

ter tuning and careful interpretation of results are necessary for effective outlier detection using Isolation Forest.

Local Outlier Factor

The Local Outlier Factor (LOF) is an unsupervised outlier detection algorithm that measures the local density deviation of data points compared to their neighbors. LOF identifies anomalies by considering the local density of instances, allowing for the detection of outliers in datasets with varying densities (BREUNIG et al. (2000)).

LOF is effective in identifying outliers in datasets with varying densities and complex structures. It is capable of capturing local anomalies that may be missed by global outlier detection methods. However, LOF requires setting appropriate parameters, such as the number of neighbors (k), and is sensitive to the choice of distance metric.

LOF is a valuable tool in outlier detection, particularly in scenarios where the density of normal instances varies or when local anomalous patterns are of interest. Careful consideration of the dataset characteristics and parameter tuning is essential for successful outlier detection using LOF.

2.2.3 Instance Selection

IS is a data preprocessing technique involving the selection of a subset of instances from a given dataset while preserving data quality and representativeness. IS aims to reduce dataset size by removing redundant, irrelevant, or noisy instances, potentially enhancing efficiency and effectiveness in subsequent data analysis tasks (GARCÍA et al. (2015)). IS can be performed using various strategies, including optimisation of objective functions(MA et al. (2017)), clustering-based selection(LIN et al. (2017), genetic algorithms(TSAI et al. (2013) and ensemble-based selection(PAN et al. (2005)). The choice of strategy depends on the characteristics of the data, the analysis goals, and the available resources.

The effectiveness of an IS task can be evaluated by comparing the results of the analysis using the full dataset versus the selected subset (HUANG et al. (2021); LIN et al. (2017)). Evaluation metrics such as accuracy, precision, recall, or computational efficiency can be used to assess the impact

of instance selection on the analysis task. Instance selection involves a trade-off between data size reduction and information loss. Aggressive IS can lead to significant reduction in dataset size but may result in loss of valuable information. It is important to strike a balance between reducing the dataset size and preserving the integrity and representativeness of the data.

IS is a useful technique in scenarios where data size, computational resources, or processing time are limiting factors. It reduces duplication, noise, and outliers while focusing on a representative selection of cases. However, careful thought and review are necessary to guarantee that the selected instances sufficiently represent dataset features while maintaining the acceptable data quality for the task at hand.

3
Related Work

Reverse instance selection (RIS) is a novel approach, after thorough research, there is no publication that directly explores solutions to problems in any domain using this approach. However, this is similar to conventional IS. Intuitively, criteria that determine a good IS should also hold for RIS. As with IS, we do not want to extract instances that are outliers, redundant, or not representative of the given instance, so for the rest of the section, I discuss related work that influenced my approach to develop the RIS extraction technique applied to a dataset of IA instances.

For the section on aneurysms, I discuss some of the relevant applications of machine learning to IA related tasks, it covers rupture risk prediction, subgroup analysis, and discriminating feature analysis using patient-specific, morphological and hemodynamic features. I also briefly discuss aneurysm training simulations; this section talks about their advantages and current state-of-the art approaches to training simulations for IA's. The section on instance selection discusses some IS extraction techniques, metrics that were developed and used to judge how representative selected instances are with respect to a larger database, and the changes in classification performance of some IS approaches. These existing researches in IS guided the RIS approach discussed in subsequent chapters of this work, i.e. the novel RIS approach introduced in this work is an adaptation of IS to suit our task.

3.1 Aneurysms

Features of IA's can be broadly classified into three categories; Patient-specific features (examples are age, hypertension, aneurysm location),

Morphological features (examples are aneurysm area, orthogonal height, aspect ratio) and Hemodynamic features (examples are wall shear stress, oscillatory shear index). Some of the most popular risk factors associated with rupture of IA have been studied extensively, these include hypertension, smoking, a history of SAH, presence of multiple aneurysms, the location and the size of the aneurysm (JUVELA et al. (2013), SONOBE et al. (2010), WIEBERS (2003), MORITA et al. (2005), WERMER et al. (2007), INVESTIGATORS (2012), BOULOUIS et al. (2017)).

Significant scientific efforts have gone into IA related tasks, ranging from rupture risk prediction (WEIR et al. (2002), AN et al. (2022)), identifying discriminating features (XIANG et al. (2011)), subgroup selection (WANG et al. (2021), NAGGARA et al. (2012), ZHANG et al. (2022)) etc. These approaches has been applied to both image and tabular datasets using several combinations of features for training or building the ML model, depending on the type of learning approach to be employed.

WEIR et al. (2002) examined the risk of rupture in intracranial aneurysms based on size, location, and patient age. A retrospective database of 945 aneurysm patients treated between 1967 and 1987 was analyzed. 86% of the patients had ruptured aneurysms. Ruptured aneurysms were mostly smaller than or equal to 10 mm, located on the anterior cerebral artery or anterior communicating artery, and less commonly on the middle cerebral artery. The average size of ruptured aneurysms (10.8 mm) was significantly larger than unruptured ones (7.8 mm). Patient age did not show a significant impact on aneurysm size although they may have significant impact on rupture risk. Symptomatic unruptured aneurysms tended to be larger than incidental unruptured aneurysms. The study concluded that aneurysm location, patient age, and size can influence the likelihood of rupture.

To investigate discriminating parameters for IA's, XIANG et al. (2011) and TANG et al. (2022) approached the task of identifying important morphological and hemodynamic parameters that are associated with ruptured IA's. These experiments were performed by training multivariate regressions on the parameters and evaluating them using the area under the curves (AUCs) of their results using receiver operating characteristics (ROCs) within a given statistical significance. The former found the size ratio to be an important morphological parameter and the important

hemodynamic parameters to be wall shear stress (WSS) and oscillatory shear index (OSI), while the later found bleb formation, neck width and size ratio to be important parameters.

AN et al. (2022) introduce a novel semiautomatic prediction model for estimating the risk of aneurysm rupture. The model utilizes 110 datasets with 128 annotated aneurysms provided by the cerebral aneurysm detection and analysis (CADA) challenge. It incorporates multidimensional feature fusion, feature selection, and classification methods. Four types of features (morphological, radiomics, clinical, and deep learning) are extracted and combined into a feature set. Different deep learning features are analyzed using a feature extractor. Five classification models are constructed, with the k-nearest neighbor classifier performing the best, achieving an F1-score of 0.789 for aneurysm rupture risk estimation. The study demonstrates that leveraging multidimensional feature fusion enhances the accuracy of aneurysm rupture risk assessment, outperforming other methods based on CADA challenge 2020.

SPITZ et al. (2020) developed a tool for case-based reasoning support of rupture risk prediction using morphological parameters. This tool holds a reference database and outputs the most similar aneurysms with respect to an aneurysm of interest. Similarity was calculated using Euclidean distances, where smaller distances represent the most similar aneurysms, the rupture status was summarized using three K-nearest neighbor (KNN) classifiers with different constraints. Evaluation of this work was done via a questionnaire with six highly experienced physicians, and they evaluated this tool positively. This tool was extended in SPITZ et al. (2021) by removing the restriction to just morphological parameters and allowing classification to be done for arbitrary parameters.

ABBOUD et al. (2017) also showed that the morphology of aneurysms has an independent predictive value for aneurysm rupture. They conducted experiments on 420 patients to compare ruptured with unruptured aneurysm by classifying aneurysm morphology into single-sac aneurysms with smooth margin, single-sac aneurysm with irregular margin, aneurysms with a daughter sac and multiboluted aneurysms. The analysis was done using logistic regression, PHASES score (BACKES et al. (2015)) features, and Fisher's exact test (UPTON (1992)) which is used to study correlation between morphological features.

One of the most popular study on subgroup analysis of IA is by RYTTLEFORS et al. (2008), they proposed a course of treatment for some aneurysms by conducting a subgroup analysis on 278 elderly patients aged 65 years and older, they proposed that EVT should be the optimal choice for ruptured internal carotid and posterior communicating artery aneurysms while, NST is beneficial for patients with ruptured middle cerebral artery aneurysms.

NAGGARA et al. (2011) emphasize the problem of subgroup analysis on ruptured IA's, pointing out spurious effects and how detrimental they are to prescribing courses of treatment, and explicitly stating that the work of RYTTLEFORS et al. (2008) should not be used for clinical decision-making because it doesn't meet some prudent criteria such as clinical plausibility and replication of results in other studies. They advised that results from subgroup analysis should be used as a hypothesis for another trial.

3.2 Aneurysm Training Simulations

IA's are a complicated, potentially fatal disease that requires precise surgical intervention. Neurosurgeons must get extensive training due to the delicate nature of these procedures, and surgical simulators have proven to be a useful tool in this regard (AGHA und FOWLER (2015); SEIL et al. (2022)). Generally, surgical simulators offer a learning environment independent of the hazards associated with patient care, allowing students and established practitioners to make mistakes in judgment and execution without having devastating effects (ISSENBERG und SCALESE (2008); KOCKRO et al. (2007)). By emphasizing significant research and developments in the area, this section seeks to give an overview of the state of the art in surgical training simulators for IA's.

Personalized simulations based on patient-specific anatomical data have been developed. Through the use of these simulations, surgeons may practice procedures on digital models that closely mimic the anatomy of actual patients, improving the precision and customization of surgical planning and execution. The case of an 8-year-old kid with a fusiform cerebral aneurysm with recorded progressive growth is discussed by MCGUIRE et al. (2021). The boy was successfully treated after the authors practiced

the installation of a flow diverter using a replicator system model made specifically for him by 3D printing.

For neurovascular surgical simulations, RYAN et al. (2016) created a novel manufacturing process. The procedure uses patient-derived anatomic data and three-dimensional (3D) printing to create a practical, dimensionally correct model for aneurysm clipping. The model was created with reproducibility and flexibility for new patient geometries in mind. A patient-derived, modular medical simulator was created so that medical students could practice aneurysmal clipping. A geometrically precise model of the human cranium and vascular tree with nine patient-derived aneurysms were created using a variety of 3D printing techniques. To create a patient-derived brain model, 3D printing and elastomeric casting were used. A qualitative follow-up research offers the possibility of improving present educational programs, and evaluations back up the effectiveness of the dummy.

The use of virtual reality (VR) simulators for training in IA surgery is growing. They provide experience for various surgical techniques and realistic three-dimensional environments with haptic feedback. According to the studies (ALLGAIER et al. (2022); KOCKRO et al. (2007), VR-based simulations enhance trainee performance and confidence in actual surgical situations.

ALLGAIER et al. (2022) proposed an immersive VR training simulation system where the aneurysm neck can be treated with a certain microsurgical clip. The affected area is seen to determine the clip position before the clip is closed and the vessels are deformed. Their qualitative assessment of two neurosurgeons with varying degrees of experience reveal advantages including heightened motivation, presence, and the opportunity to test out various tactics. Nevertheless, several surgical procedures can be modified to boost realism and learning impact, and interactions can be further enhanced.The proposed training method gains from trial-and-error learning in an enjoyable setting, resulting in an enhanced training experience.

3.3 Instance Extraction

REINARTZ (2002) described a unifying approach to instance selection which consists of the following steps: sampling, clustering and proto-

typing. A set of examples is drawn from the database using a sampling technique in the sampling phase, then these are clustered to group them into smaller subsets, and the prototyping phase selects the final representatives from the clusters. GARCÍA et al. (2015) divided instance selection strategies into two, namely: prototype selection (PS) and training set selection (TSS) (Figures 3.1 and 3.2). PS is mainly used for instance based learners, where instances are selected from the training set which maximizes the classification accuracy of the test set, while TSS strategy is simply selecting instances which are used to train a machine learning algorithm in order to obtain a model that can be used on a test set.

IS approaches are very useful for ML training with medical datasets because of the benefits they bring, these benefits include reducing the size of training set by selecting the most useful examples for training, this consequently leads to less memory usage and less time required for modeling. There has been extensive research on IS for medical datasets. HUANG et al. (2021) worked on improving the performance on various instance selection algorithms like the Decremental Reduction Optimization Procedure 3rd version (DROP3) and Instance Based framework version 3 (IB3) by introducing a divide-and-conquer based IS framework (DCIS) where the dataset is divided into subsets and a specific IS algorithm is applied on a combination of the subsets and the selections are then combined at the end. This methodology was evaluated using various small and large scale medical datasets from the UCI Machine Learning Repository and the results showed that using DCIS gave better classification performance than using the individual IS algorithm alone. HUANG et al. (2018) proposed an approach for handling missing value imputation that uses a combination of IS algorithms (DROP3, GA, IB3) and conventional imputation algorithms (K-nearest neighbor imputation (KNNI), Multi-layer perceptron (MLP), Support Vector Machines (SVM)), because the estimations of conventional imputation algorithms can be influenced by outliers. The results showed that the IS approach gave positive results of numerical and mixed medical datasets, but there was no positive impact on the categorical counterparts. The authors also used data from the UCI Machine Learning Repository.

Clustering based approaches are one of the most popular methods to instance selection, many authors have applied this to the instance selection

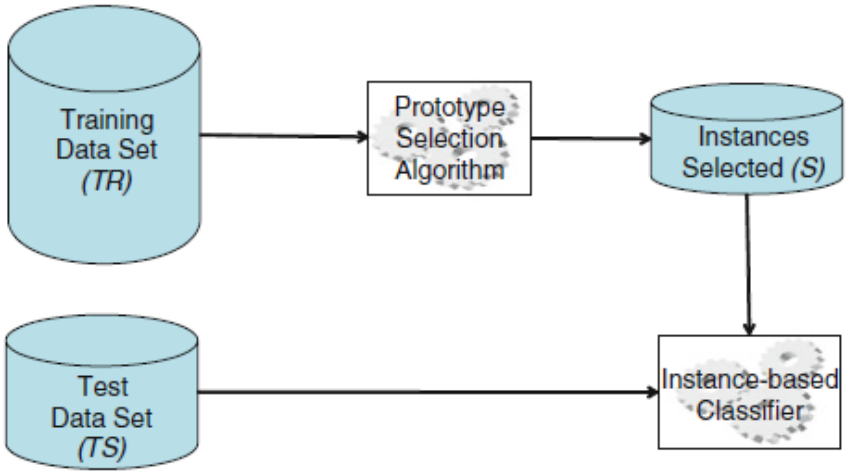

Figure 3.1: Prototype selection strategy. (GARCÍA et al. (2015))

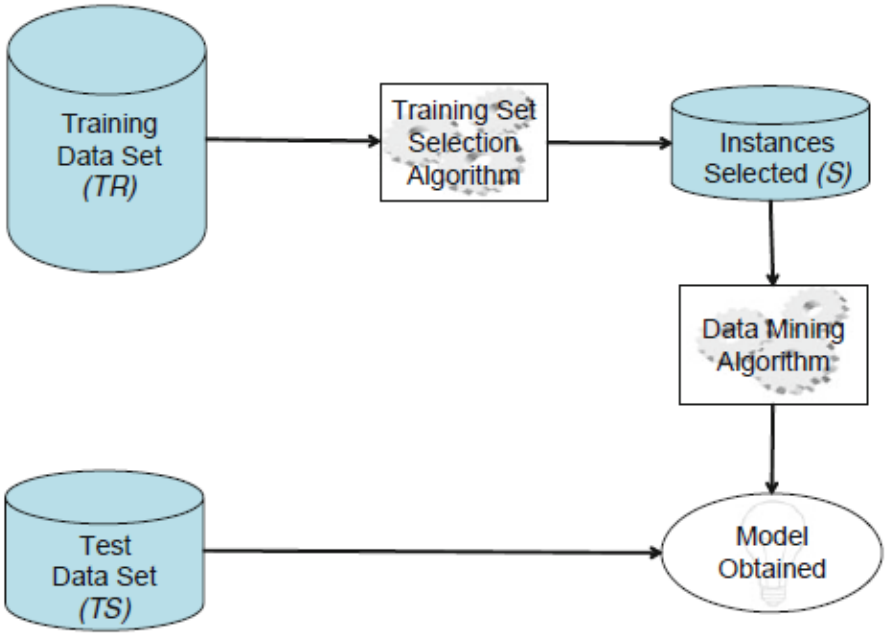

Figure 3.2: Training set selection strategy. (GARCÍA et al. (2015))

problem, using various design configurations to suit different tasks. LIN et al. (2017) proposed a a clustering based selection approach for the class imbalance problem. After splitting the data into the the train and validation sets, the authors clustered the majority class of training set using the number of minority class entries as the number of clusters, instances are then selected from these clusters using the centroids or entries closest to the centroids. This yields a balanced dataset which is used for training a model. CZARNOWSKI und JĘDRZEJOWICZ (2018) also used a clustering approach for imbalanced data classification, their approach was augmented with a population learning method that was used for the prototype selection phase.

IS play a primary role in machine learning on medical datasets as they most times have problems like class-imbalance and lack of representative data. KIM et al. (2017) achieved improvement in recall for the minority class after pre-processing with instances selected using a clustering based approach on medical datasets. The datasets are the labeled i2b2/VA 2010 shared task corpus, and the unlabeled MIMIC II clinical database. Figure 3.3 shows the performance of the instances selected with their Labeled Data Counterparts (LDC) Selection algorithm compared to supervised learners.

MALHAT et al. (2020) proposed two class label based approaches to instance selection: GDIS (Global density-based Instance Selection) and EGDIS (Enhanced Global Density-based Instance Selection) which uses K-Nearest Neighbors and a relevance or irrelevance function with respect to class labels to output a reduced set of the original dataset.

PAN et al. (2005) extracted representative samples using two requirements (high coverage and low redundancy), the objective function consisted of an equally weighted combination of mutual information and relative entropy. The former was used to evaluate the coverage, while the later was used to evaluate redundancy. They employed a greedy algorithm that selects new samples which optimizes this objective function at each step. This algorithm was simplified to reduce computational complexity for large datasets. They evaluated their work by comparing the performance of their algorithm on the Mushroom and 20 News Group datasets using MaxCover (HOCHBAUM und PATHRIA (1998)) and a random selection as baselines, the coverage and clustering accuracy metrics that they designed

Relation type	Recall		Precision		F_1 score		True positive	TP Gain (%)	
Minority classes									
TrIP	*38.9	(+7.1)	60.6	(-3.0)	*47.4	(+5.0)	77	14	(+22.2)
TrWP	*11.9	(+7.7)	50.0	(+7.1)	*19.2	(+11.6)	17	11	(+183.3)
TrCP	*65.1	(+12.8)	*44.9	(-14.6)	53.1	(-2.5)	289	57	(+24.6)
TrNAP	23.6	(-1.6)	*58.4	(+9.0)	33.6	(+0.3)	45	-3	(-6.3)
TeCP	*57.7	(+12.6)	*61.4	(-10.0)	*59.5	(+4.2)	339	74	(+27.9)
Majority classes									
TrAP	*80.8	(+0.9)	71.3	(+0.2)	75.8	(+0.5)	2,009	23	(+1.2)
TeRP	*88.5	(-1.8)	*85.0	(+2.4)	*86.7	(+0.4)	2,682	-55	(-2.0)

Figure 3.3: Experiment results from KIM et al. (2017). In the rightmost column, the Recall, Precision, and F1 outcomes of LDC instance selection are displayed, together with the total counts of true positives (TP) and the number and percentage of true positive increases (in comparison to supervised learning). In the Recall, Precision, and F1 columns, the numbers in parentheses reflect the difference between the supervised classifier and the LDC technique. Asterisks (*) indicate results that are significantly different from supervised learning at the 95% confidence level.

were used as performance metrics. Figure 3.4 shows the results for one of the experiments conducted by the authors.

ZHUANG et al. (2008) also use a greedy algorithm that optimizes an objective function which comprises measures of representativeness, anomaly and diversity. Their approach employs a clustering based measure where similarities or dissimilarities are computed with respect to cluster centroids. This approach was applied to extract the most representative selection of N posts that sufficiently profile a blog based on all topics covered by the all entries posted on the blog. Distance measures were used to account for representativeness, diversity and anomaly.

MA et al. (2017) developed a sequence of heuristics that culminated in the development of $FastCov_{c+s-Select}$ method for extracting samples which tries to optimize the objective function developed by MA und WEI (2012). This function combines a pairwise similarity $Cov_{c-Select}$ that checks for similarity and an entropy $Cov_{s-Select}$ part which handles structure. The same authors extended their work (CHEN et al. (2018)) where they expanded on various metrics which were grouped into the closeness aspect and the duplication aspect. Metrics in the former are content coverage,

$\|R\|$	Representative Set	MaxCover	RandomPick
2	100%	50%	70%
3	100%	100%	95%
4	100%	100%	90%
5	100%	100%	100%
10	100%	100%	100%

(a) Representative Coverage

$\|R\|$	Representative Set	MaxCover	RandomPick
2	67.9%	51.7%	48.3%
4	75.1%	71.0%	63.5%
8	89.0%	89.2%	79.3%
20	96.3%	96.4%	90.7%
30	100%	96.3%	93.7%

(b) Clustering Accuracy

Figure 3.4: Experiment results from PAN et al. (2005). (a) shows the scores of the coverage metric while (b) shows the score for clustering accuracy metric. R represents the number of instances in the extracted set

structure coverage and consistency, while those in the later are redundancy and compactness. Choice of distance function for the metrics that require the calculation of distances can be changed based on needs and domain.

Some authors such as TSAI et al. (2013) and CANO et al. (2003) have explored the use of genetic algorithms for instance selection, modeling the instance selection task as a search task. They used concepts such as cross-over, mutation, representation and a fitness function that combines classification rate and reduction rate of the selected subset with respect to the complete dataset. LEE et al. (2021) proposed a method for handling redundancy in representative set selection comprising of two major components: $EXACTSUBSTR$ and $NEARDUP$. $EXACTSUBSTR$ compares two documents and handles duplication by removing exact long substrings from one document if they exist in both documents, while $NEARDUP$ handles duplication by checking for approximate matches. Approximate matches are estimated using the Jaccard coefficient between the two sub-texts. Although this method has great advantages, it employs string matching techniques that are suitable for language models.

Although some of these approaches are useful, they have some limitations to this task. While popular instance selection techniques entail selecting subset with respect to a larger dataset, I have to select a subset with respect to an example. Measures which are centroid-based, density-based, etc., are not sufficient for this task. Also, some of the approaches are in the Natural Language Processing domain, where ideas like string matching are not useful for this task.

Majority of IS (AJMAL et al. (2023), LIN et al. (2015), WILSON und MARTINEZ (2000), SONG et al. (2017), OLVERA-LÓPEZ et al. (2010), KIM (2006)) related works are evaluated either by comparing the classification accuracy of the generated subsets with that of the entire dataset, or comparing how the presented method performance with respect to other existing measures of extraction, but these approaches are not suitable for this task because I am finding subsets that are similar to one instance. Hence our work will be evaluated by using randomly selected samples from the database and most similar instance to a given instance as baseline.

4
Methods

4.1 Problem Definition

The task of selecting representative but diverse samples of aneurysms from a database of aneurysms can be approached as an IS task. The idea is that while IS aims to select a subset of examples that sufficiently describes a larger database, our task aims to select a subset from a larger database that describes a single example sufficiently.

For this task, the selected representative subset should not be the most similar because the diversity of selections is pivotal. The variance in the selected subset can be useful for medical training simulation as it provides the medical practitioner or trainee with potential differing nuances with respect to an AOI, hence we are not interested in a subset most similar to the given example, but in a subset that offers variance in-line with similarity.

The task can be formalized as such: given an AOI ($\boldsymbol{A}$) and a database of aneurysms ($\boldsymbol{D}$), extract a subset ($\boldsymbol{S}$) from $\boldsymbol{D}$ such that the entries in $\boldsymbol{S}$ are both representative and diverse with respect to $\boldsymbol{A}$.

4.2 Concept

Figure. 4.1 shows a general overview of the RIS framework. We develop an extraction methodology to generate a subset of similar and diverse instances from a larger dataset with respect to an AOI, this subset is then evaluated using metrics, which would be defined in subsequent sections. While instance selection is the process of selecting smaller samples with respect to a database with more samples, our task tries to select more sam-

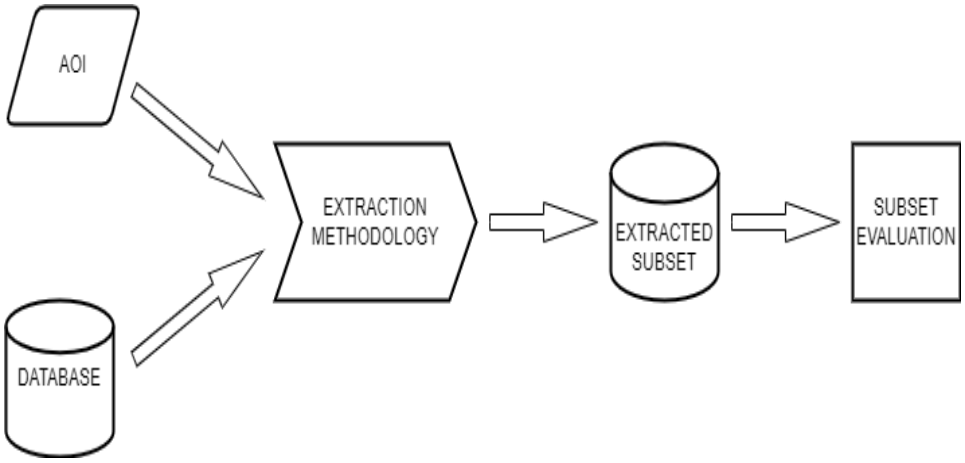

Figure 4.1: Low-level overview of the proposed RIS framework

ples with respect to one sample, hence we call this Reverse Instance Selection (RIS)(See Table 4.1).

Table 4.1: IS Vs RIS

	IS	**RIS**
Given	Dataset	Dataset
Focusing	Dataset	An Instance
Extract	Subset of Dataset	Subset of Dataset
Logic	From More to Less	From Less to More

The following software and libraries shown in table 4.2 was used for analysis, design, experimentation and evaluation of methods and techniques used to build this framework.

4.3 Datasets

Main Dataset: The main dataset used in this study comprise 76 IA's from 54 patients. This dataset was curated from three medical facilities in Germany. Table 4.3 shows the cooperating medical facilities and a summary of the IA's amassed from them. This dataset was presented to me in an excel $.xlsx$ format. I also had access to the 3D segmentation of every IA instance in the $.xlsx$ document.

Table 4.2: Software used for Development and Experiment

Software	**Version**	**Use Case**
Python	3.10.12	Programming Language
Pandas	1.5.3	Data Wrangling and Analysis
Numpy	1.23.5	Data Wrangling
Sci-kit Learn	1.2.2	ML Algorithms
Scipy	1.11.2	Analysis and Experimentation
Matplotlib	3.7.1	Visualisation
Seaborn	0.12.2	Visualisation
Sdv	0.17.2	Synthetic Data Generation
Prince	0.6.2	Dimensionality Reduction
Tensorflow	2.13.0	ML Algorithm
Xgboost	1.7.6	ML Algorithm

Table 4.3: Co-operating Medical Facilities for Main Dataset

Medical centre	**Number of patients**	**Number of IA's**
University Hospital Magdeburg	18	30
Kiel	25	25
Hannover	11	21

These IA's were diagnosed using several imaging modalities, such as digital substration angiography (DSA), 3D rotational angiography (3DRA) and magnetic resonance angiography (MRA) (Figure 4.2 shows an example of multiple IA's from one patient). Morphological features were extracted from these images using techniques stated in SAALFELD et al. (2018), while hemodynamic features were extracted using techniques described in CEBRAL et al. (2011). Table 4.4 shows the feature type and the amount present in the dataset after the extraction of morphological and hemodynamic features from the images.

Table 4.4: IA feature type and the number present in dataset

Feature Type	**Number of Type**
Patient-specific	7
Morphological	23
Hemodynamic	7
Sum	37

This dataset will be used for training and evaluating the model.

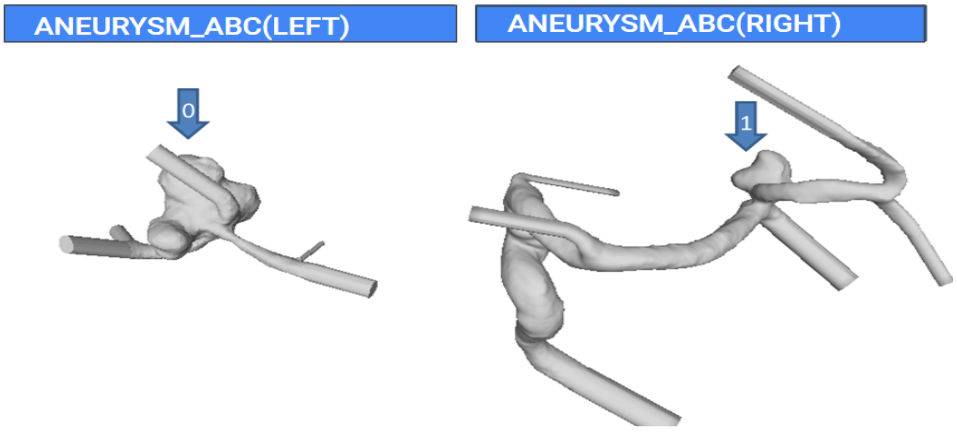

Figure 4.2: Image of multiple aneurysms from a patient

Supplementary Dataset: The supplementary dataset is from an IA database contained 406 cases (178 females, 69 men, and 159 unclassified) ranging in age from 17 to 93. The data was collected over a 5 year period with collaborating clinics. Each patient was initially described by 106 qualities, the majority of which were numerical and categorical. This database was collected and cleaned in a previous research. After the previous cleaning, the final supplementary dataset used in this work is a tabular dataset of aneurysms containing data for 351 aneurysms and 12 features, 4 of the features are categorical (Rupture Status, Multiple Aneurysm, Aneurysm Location, Side), while 8 were numeric (Width, Neck, Parent Vessel, Maximum Height, Size Ratio, Vessel Angle, Inclination Angle, Inflow Angle). This data set is in an excel $.xlsx$ format.

This dataset will be used for further evaluation of the model.

4.3.1 Data Cleaning and Pre-processing

Main Dataset

Necessary data cleaning and pre-processing were done because the dataset contained some common problems such as missing values, typographical errors, and regular expressions. The cleaning and pre-processing techniques employed are briefly discussed subsequently.

Data Correction: Firstly, I corrected some obvious typographical errors in the data and also removed meaningless regular expressions which were a product of data transformations. Entries that were only senseless expressions were also replaced with *NaN*.

Missing Value Imputation or Removal: While it was adequate to fill some missing features or aneurysms intuitively from analysis, others had to be removed either because there was no intuitive way to impute them or over 50% of its entries were missing. An example of a data imputation technique I employed were labeling the few aneurysms from the Kiel clinic as a bifurcation aneurysm of the middle cerebral artery (MCA-bif) by examining the images of the aneurysm in reference to whole circle of willis (Figure 2.1). My conclusions were also supported by the fact that every other aneurysms from the Kiel clinic which was labelled were also MCA-bif.

Feature Aggregation: Some morphological features, such as the ostium area, orthogonal height and aspect ratio were measured twice. These features highly correlated with each other, so they were combined into one feature by taking the average of both values.

Table 4.5 shows the features of the dataset, their description, and properties after the feature aggregation phase of the pre-processing steps.

Data Augmentation: I tried to increase the samples by augmenting with synthetic data, but this approach was discarded because synthetic data could not be evaluated qualitatively. To evaluate qualitatively, images of the AOI and its extracted set should be presented to doctors, who will provide expert opinions on the suitability of the extracted set with respect to the AOI. If synthetic data is used to build a model, there will be no images to represent these data points, unlike the real data points. Furthermore after clustering the synthetic dataset in conjunction with the real dataset, DBSCAN produced 14 clusters. This is a significantly larger number of clusters in comparison to that of the real dataset which produced 3 clusters. Figure 4.3 and 4.4 shows the clustering of the dataset with 500 synthetic data points and the 70 real data points.

Categorical and Boolean Encoding: Categorical and Boolean features such as rupture status, aneurysm location, etc., were encoded to enable processing using popular machine learning libraries. Encoding involves converting each categorical variable into distinct Boolean variables (also

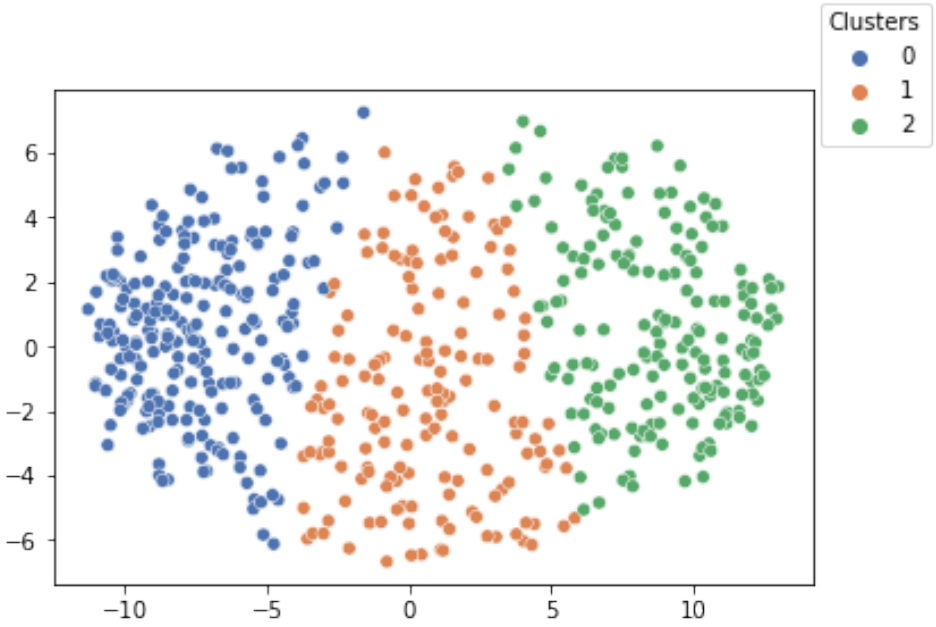

Figure 4.3: K-means clustering of data augmented with 500 data points

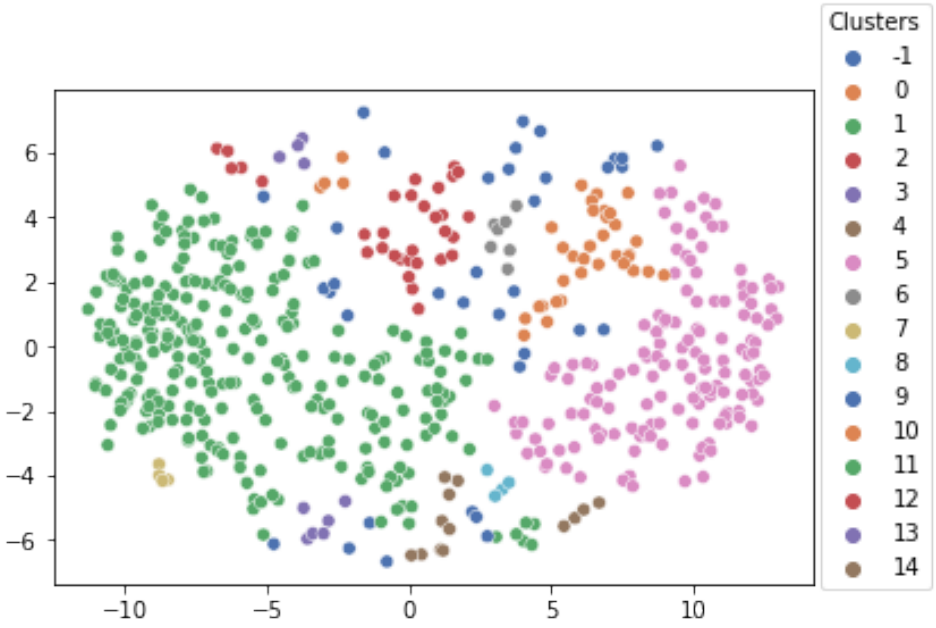

Figure 4.4: DBSCAN clustering of data augmented with 500 data points

Table 4.5: Description of the Features of the Main Dataset After Pre-processing

Feature	**Description**	**Feature Type**
D_{max}	Maximum diameter of aneurysm	Morphological
N_{avg}	Average ostium diameter	Morphological
N_{max}	Maximum ostium diameter	Morphological
AneurysmArea	Surface area of the aneurysm sac	Morphological
AneurysmVolume	Volume of the aneurysm sac	Morphological
H_{max}	Maximum height of aneurysm	Morphological
W_{max}	Maximum width perpendicular to H_max	Morphological
W_{ortho}	Maximum width parallel to the projected ostium plane	Morphological
convexHullVolume	Volume of the convex hull of the aneurysm sac	Morphological
convexHullSurface	Surface area of the convex hull of the aneurysm sac	Morphological
EI	Ellipticity index	Morphological
NSI	Non-sphericity index	Morphological
UI	Undulation index	Morphological
alpha	Larger angle between centerline and dome	Morphological
beta	Smaller angle between centerline and dome	Morphological
gamma	Angle at the aneurysm dome	Morphological
H_{ortho}	Height perpendicular to the ostium center	Morphological
OstiumArea	Area of the ostium	Morphological
AspectRatio	$((H_{ortho}/N_{max}) + (H_{ortho}/N_{avg}))/2$	Morphological
RuptureStatus	Rupture status of aneurysms	Patient-Specific
MultipleAneurysms	Presence of multiple aneurysms	Patient-Specific
$AWSS_{mean}$	Mean wall shear stress	Hemodynamic
OSI_{max}	Oscillatory shear index	Hemodynamic
MeanNeckInflowRate	Mean inflow rate into the aneurysm neck	Hemodynamic
ICI_{mean}	Inflow concentration index	Hemodynamic
SCI	WSS concentration index	Hemodynamic
LSA	Low wall shear area	Hemodynamic

known as dummy variables) that take values of 0 or 1, indicating whether or not a category is present in an observation. For example, if we have a categorical feature "fruit" with the values "apple," "orange," and "cherry," we can create three additional features called "apple," "orange," and "cherry." These features will be set to 1 if the observation is of the specified color and 0 otherwise. The initial categorical feature is then dropped at the end of this process.

Scaling: The data is transformed using the standard scaler to have a mean of 0 and a standard deviation of 1. This procedure is essential because it equalizes the variables. In the learning process, some variables may dominate over others when features have differing scales. For instance, if one feature has values between 0 and 1, but another has values between 0 and 1000, the latter will have a significantly greater impact on the performance of the model. The features are scaled to ensure that they all contribute equally to training process. The mathematical expression for standard scaling is giving in equation 4.1

$$\textit{Standardized Value}(z) = \frac{x - \mu}{\sigma} \tag{4.1}$$

In this formula, x represents the original value of the data point. μ represents the mean (average) of the feature across the entire dataset. σ represents the standard deviation of the feature across the entire dataset.

Dimensionality Reduction: After the above preprocessing steps, the dimension of the data was 70 aneurysms with 29 features. Some ML algorithms struggle with high dimensional data, hence it was necessary to use a dimension reduction technique. This will give me the opportunity to analyse more ML approaches for this task. Principal component analysis (PCA) was used to reduce the dimension. PCA is a dimensionality reduction approach that is frequently used to decrease the dimensionality of big data sets by reducing a large collection of variables into a smaller one that still includes the majority of the information in the large set.The number of principal components (PC) selected when using PCA should be determined by the explained variance covered by each component. Explained variance is a measure of how much information load each PC covers with respect to the dataset. JOLLIFFE und CADIMA (2016) recommend a total explained variance of 70% is sufficient to determine how many PC's should

be used. Table 4.6 shows the explained variance of the first 3 PC's of the dataset.

Table 4.6: Explained Variance of Principal Components

Principal Component	**Explained Variance**
PC1	55%
PC2	22%
PC3	11%
Sum	88%

I also attempted to do the dimensionality step using factor analysis of multidimensional data (FAMD) because of its ability to implicitly handle a dataset with categorical and numerical features without encoding, but the results were difficult to cluster. Figure 4.5 shows the DBSCAN clustering of the data points after dimensionality reduction using FAMD. DBSCAN assigned all the points to one cluster (orange points) except 2 points which where outliers (blue points). This is why PCA was chosen as the technique for dimensionality reduction for this work.

After the application of all cleaning and pre-processing techniques, the dimension of the dataset is 70 aneurysms and 3 features.

Supplementary Dataset

This dataset was already cleaned and pre-processed from a previous research, but further pre-processing was necessary to suit the pipeline I created for RIS. Table 4.7 shows the features, the type of feature, and their descriptions for the supplementary dataset.

The following are the pre-processing step:

- Categorical and Boolean Encoding
- Scaling
- Dimensionality reduction

These were the last 3 steps of the pre-processing steps used on the main dataset, the reasons given in the previous subsection for the main dataset, also holds for the supplementary dataset.

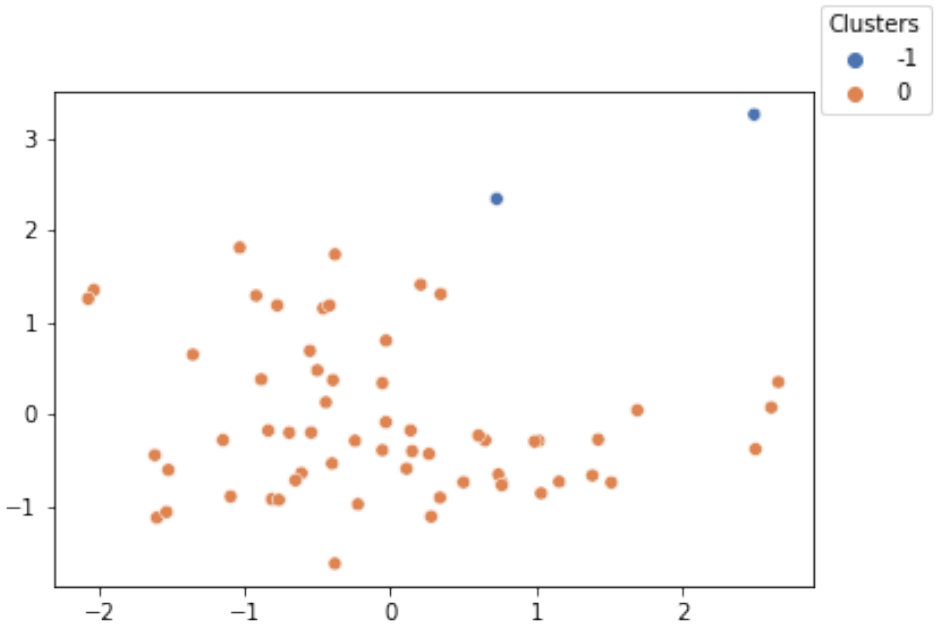

Figure 4.5: DBSCAN clustering of data after FAMD dimensionality reduction

Table 4.7: Description of the Features of the Supplementary Dataset After Preprocessing

Feature	**Description**	**Feature Type**
RuptureStatus	Rupture status of aneurysms	Patient-Specific
MultipleAneurysms	Presence of multiple aneurysms	Patient-Specific
Localisation	Location of the aneurysm	Patient-Specific
Side	Hemisphere in which the aneurysm is located	Morphological
Neck(D)	Diameter of aneurysm neck	Morphological
$W_{\max}$	Maximum width perpendicular to $H_{\max}$	Morphological
ParentVessel(T)	Parent vessel diameter	Morphological
$H_{\max}$	Maximum height of aneurysm	Morphological
SizeRatio	$H_{\max}/T$	Morphological
VesselAngle	Parent vessel inlet angle relative to the aneurysm	Morphological
InclinationAngle	Angle at which the IA is tilted with respect to the incoming flow	Morphological
InflowAngle	Angle of incoming flow	Morphological

4.4 Extraction Methodology

The goal of this thesis is to use current research to create an extraction approach that adds variation to the selection of instances that are similar to an AOI. The concept is that picking subgroups using only a similarity metric might result in repetition, and hence these instances may lack diversity, which is necessary for clinical training simulations. For this task, similar to IS, the criteria for a representative set selection apply. The following criteria are crucial to ensuring quality selection:

- **Diversity:** It is important to select varied entries in order to enhance the information contained in the chosen entries. This is important to prevent choosing the duplicate instances or instances, which are very similar to each other, such that they convey no (or little) variance between them. This will render the selections redundant.
- **Similarity:** Although diversity is important, we also want to have selections that share a reasonable relationship with the AOI. Similarity criteria is necessary to ensure instances different from the AOI are not selected.
- **Inliers:** The final criteria is inliers. In a bid to achieve variation in the selection, anomalies or outliers should not be selected because those entries are not a representation of the database.

The strategies for extraction utilized in this process were carefully selected to ensure that only the most relevant and representative entries were included. Outlier detection and removal was done to ensure only inliers are selected by the extraction methodology. Clustering was employed to group similar instances together, reducing redundancy, and ensuring a diverse selection by extracting the most similar instances from each cluster as opposed to selecting the top k most similar instances from the dataset. In the prototyping phase, similarity criteria was used to select instances most similar to the AOI, further refining the extraction process. Figure. 4.6 shows a visual representation of the extraction process, including the ML techniques explored for each step of the process. Overall, this approach ensures that only high-quality instances are included in the extracted set providing valuable insights with respect to an AOI.

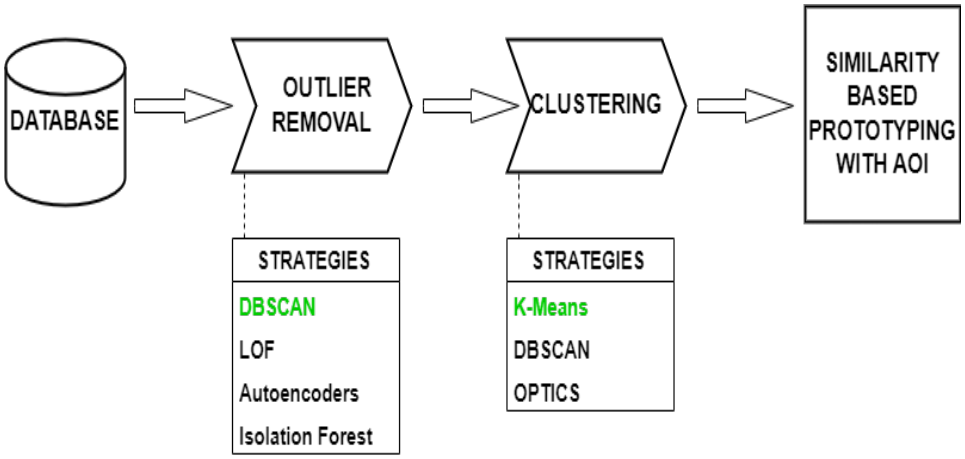

Figure 4.6: Steps for instance extraction with strategies explored for each step. Strategies highlighted in green were used after experiments.

4.4.1 Outlier Removal

To fulfill the inlier criterion, an outlier removal step was added to the extraction methodology. Experiments were conducted using 4 outlier detection algorithms, namely; autoencoders, isolation forest, DBSCAN and local outlier factor (LOF). These models were trained using Scikit-Learn's library except the autoencoder model, this was developed using the keras wrapper of tensorflow.

I designed the autoencoder architecture, which was made up of 5 layers excluding the input and output layers, these 5 layers constituted of 2 encoding layers that map the input, a middle layer that holds the mapped input in a lower dimensional space, and finally, 2 decoding layers that try to reconstruct the input to the output layer through the middle layer. The neurons in the 5 hidden layers are (128, 64, 32, 64, 128) respectively, which all used the rectified linear Unit (ReLu) activation function. This network was trained using mean squared error as loss function with an Adam optimizer.

SANDER et al. (1998); SCHUBERT et al. (2017) recommends using $2 \times N$ where N is the number of dimensions for $MinPts$. Furthermore RAHMAH und SITANGGANG (2016) proposed a technique that uses the pre-determined $MinPts$ to estimate eps. With k equal to the $MinPts$ value you chose, this method determines the average distance between

each point and its k-nearest neighbors. On a k-distance graph, the average k-distances are then shown in ascending order. The ideal estimate for eps will be at the moment of maximal curvature (i.e., the graph's steepest slope). This procedure gives $MinPts$ of 6 and eps of 2.15 for the dataset. Using these values for $MinPts$ and eps, both DBSCAN and OPTICS classified all the points as one cluster. Following multiple experiments, DBSCAN was developed with $MinPts$ as 4, eps as 1.6 and OPTICS with $MinPts$ as 4, eps as 1.74, which were found empirically, and distance metric was euclidean distance.

For the isolation forest model, the number of estimators and contamination were set as 100 and 0.06 respectively. 0.06 was used to constrain the model to select the same number of outliers as DBSCAN which assigns outliers automatically without having to specify a contamination factor.

The following are reasons that motivated the choice of DBSCAN as the algorithm for this task.

- It also doesn't require hyper-parameters like contamination rate used in the isolation forest algorithm, which explicitly determines how many data points are selected as outliers.
- LOF favors datasets with clearly defined clusters but varying outlier distances with respect to the clusters. The dataset used for this experiment doesn't share this trait.
- Neural networks like autoencoders require the tuning of several hyper-parameters unlike DBSCAN with only 2. Performing this hyper-parameter tuning can be time-consuming.

In addition to this, upon visual inspection of the outlier points selected by the 4 models as shown in Figure. 4.7, outliers spotted by DBSCAN were more understandable visually in comparison to the other algorithms.

4.4.2 Clustering

This step in the extraction is to account for the diversity of selections, I separate the points into different clusters, so similar aneurysms are grouped into the respective clusters. Experiments were conducted using three clus-

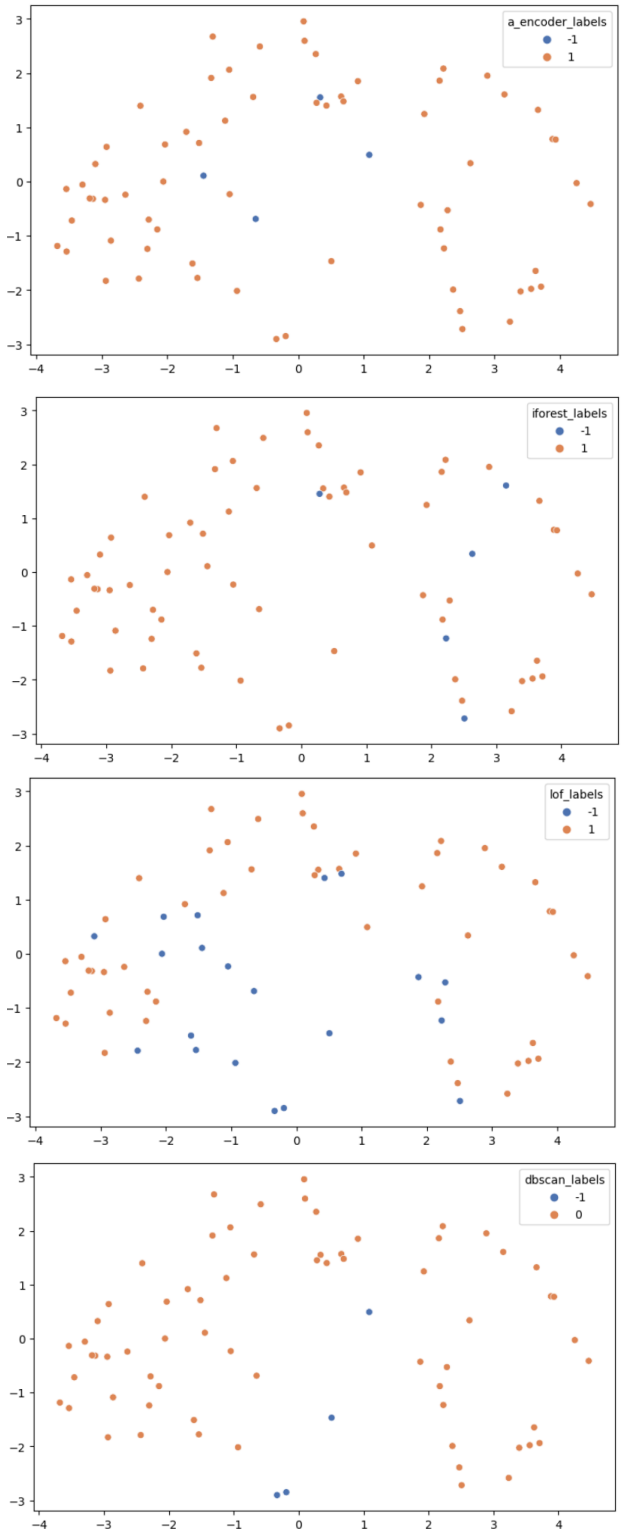

Figure 4.7: Outlier detection scatter-plots showing assignment of outliers by autoencoders (top), isolation forest (second), LOF (third), DBSCAN (bottom). X and Y axis of the scatter-plots are the first 2 principal components used only for visualisation while the assignments was done on the 3 principal components datasets

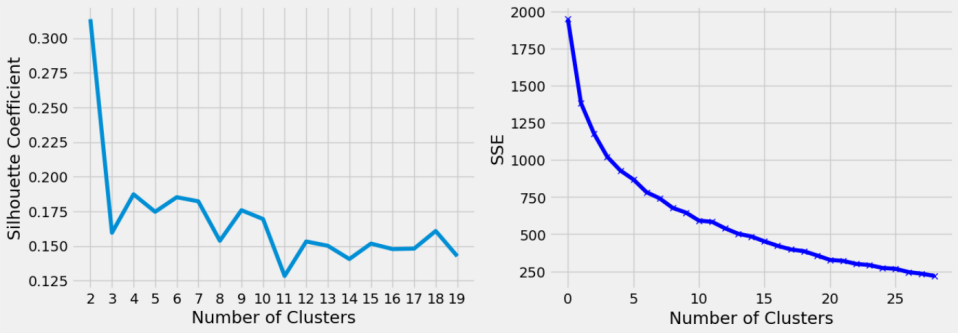

Figure 4.8: Selecting number of clusters using plots of silhouette coefficient against number of clusters (left) and SSE against number of clusters (right)

tering techniques, namely; DBSCAN, OPTICS and k-means. The Scikit-Learn library was used for all three algorithms.

Hyperparameters used for DBSCAN and OPTICS were $MinPts$ as 4 and distance metric as euclidean distance, while eps was set to 1.6 for DBSCAN, as explained in section. 4.4.1.

K-means was implemented using K=3, this value was determined by visualizing the plots of the sum of squared error (SSE) against the number of clusters and Silhouette coefficient against number of clusters and selecting the knee of the graph as K. As can be seen from Figure. 4.8, the knee of the SSE plot is between 3-5, while that of the silhouette coefficient was spurious, but there is an obvious dip when the number of clusters is 3, hence this was selected as the value for k. Figure. 4.9 shows the clustering results of the three techniques.

I also performed clustering on the dataset without dimensionality reduction. While the clustering produced by K-means was similar to that produced by the 3 dimension dataset, DBSCAN and OPTICS unsurprisingly did not produce sensible clusters as they labeled all the points as outliers, these algorithms do not usually do well on high-dimensional datasets. Although our analysis for the development of RIS was done using the dataset produced after 3-dimensional PCA, selecting k-means as the algorithm for the clustering step was influenced by the fact that the k-means clustering result produced on the dataset before PCA was very similar to the k-means clustering result produced after PCA. Figure. 4.10 and Figure.

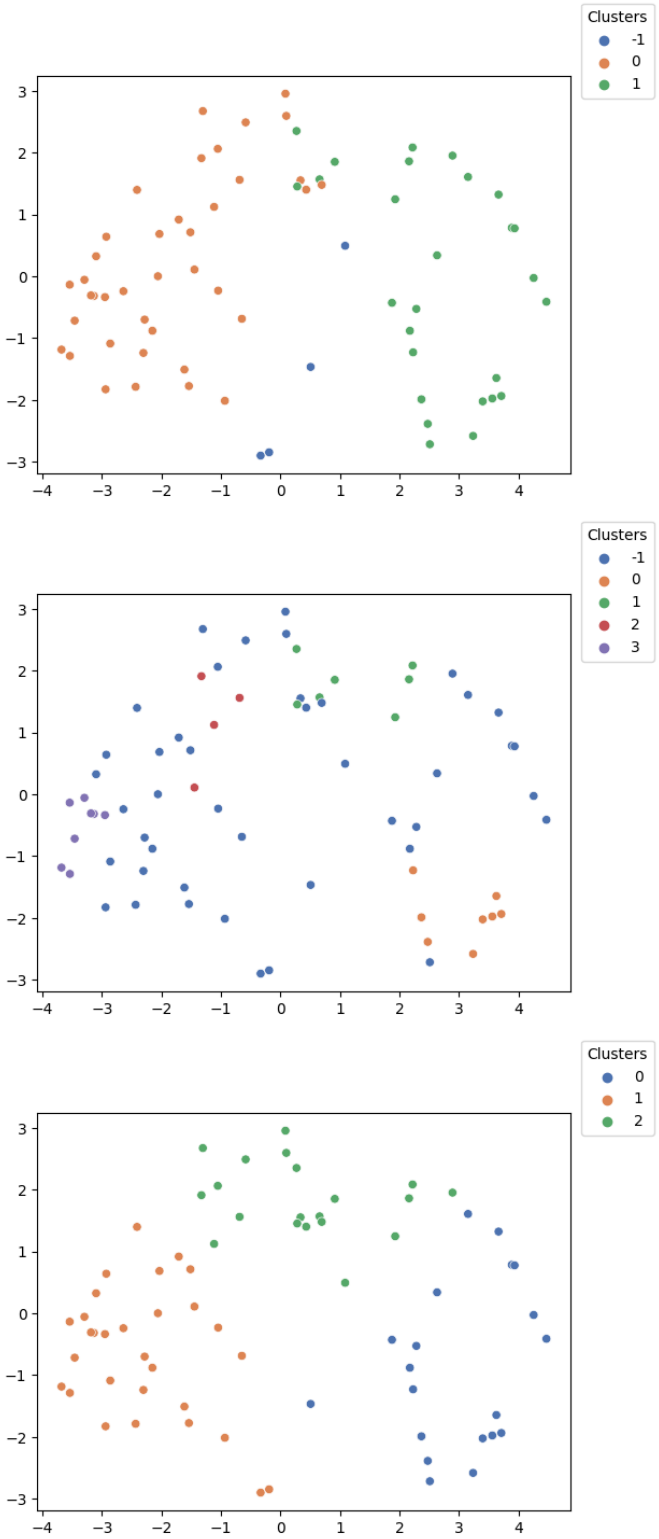

Figure 4.9: Cluster assignments: DBSCAN assignments (top) with 2 clusters and 4 points as outliers (0: 39, 1: 27, -1: 4), OPTICS assignments (middle) with 4 clusters and 43 points as outliers (-1: 43, 3: 8, 0: 8, 1: 7, 2: 4), Kmeans assignments (bottom) with 3 clusters (1: 30, 0: 20, 2: 20)

4.11 shows the K-means clustering on the full and 3-dimensional datasets visualized using the first two principal components for both plots.

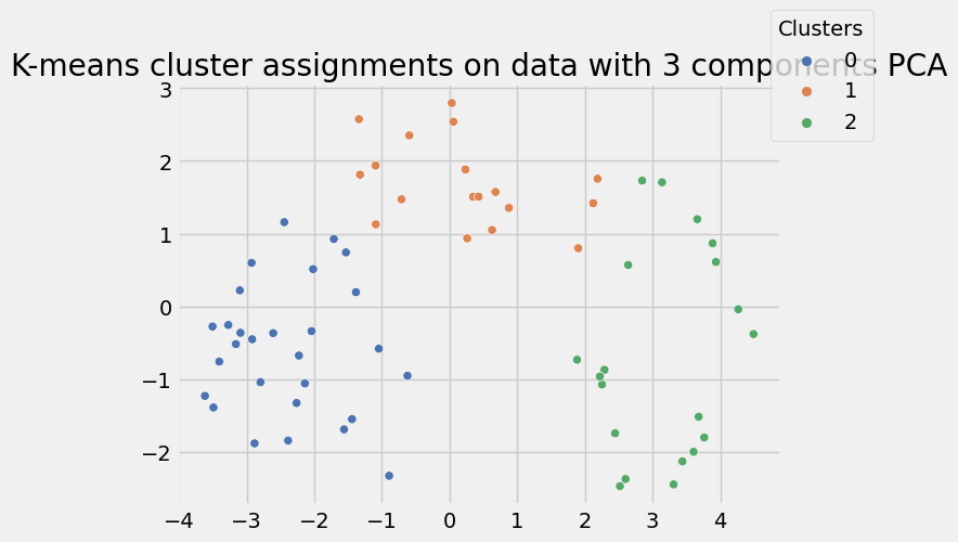

Figure 4.10: K-means clustering of data with 3 components of PCA visualised using the first 2 PC's as the X and Y axis respectively

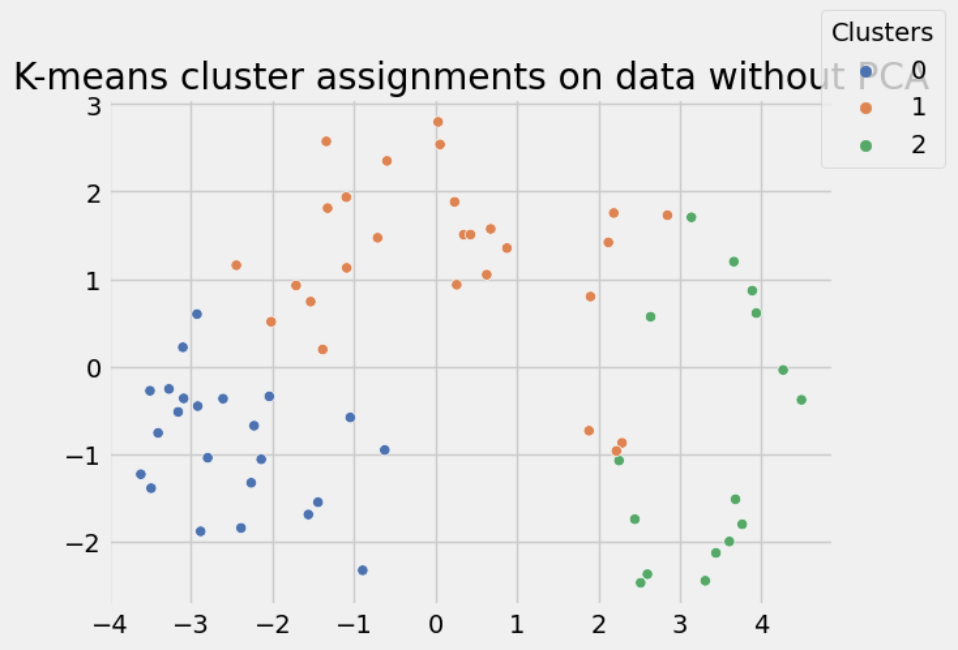

Figure 4.11: K-means clustering of full dataset without PCA visualised using the first 2 PC's as the X and Y axis respectively

4.4.3 Prototyping

At this step of the process, the candidate selection is done by extracting instances most similar to the AOI from each cluster iteratively until the desired number of samples is achieved. This subset represents the instances that are similar to the AOI, but also diverse from each other. For the RIS model, the minimum number of instances that can be extracted is equal to the number of clusters in the clustering step while the maximum number of instances that can be extracted is equal to the number of instances in the dataset.

A key advantage of this proposed RIS framework is the flexibility it offers with respect to ML algorithms. Depending on various factors such as domain, type of dataset, sparsity of dataset, etc, various aspects can be adapted to suit the needs of the user. Parameters such as the number of clusters can be changed if there is existing knowledge of the number of groups present in a given dataset. Also, it allows for flexibility in ML algorithms, depending on the needs of the designer and the dataset, this framework can be adapted using various ML techniques for the outlier removal, clustering, and prototyping phases.

Instance Selection Adaptation

The proposed extraction methodology can be adapted to suit IS tasks by changing the prototyping phase. Instance selection is the process of selecting a subset of samples from a database that sufficiently conveys all the nuances of the complete dataset. This task does not require an AOI, therefore changing the similarity-based prototyping to a centroid-based prototyping makes this framework suitable for instance selection.

Centroid-based prototyping is done by iteratively selecting instances closest to the clustering centroids until the desired number of instances for the extracted subset is reached.

4.5 Proposed Evaluation Approach

4.5.1 Metrics

To evaluate the extracted subset, we use the following metrics: similarity, redundancy, and anomaly. These metrics are explained in the following subsections.

Similarity

Distance based measures are popular ways of calculating similarity between vectors. The idea is built on the notion that points that are more similar to each other have smaller distances between them and vice-versa. There are different types of distance measures (SHARMA und KUMAR (2016)), for this work, we use Euclidean distance.

The Euclidean distance is widely used as a measure of similarity or dissimilarity between points in space in many domains, including mathematics, physics, computer science, and data analysis. It serves as a foundation for many algorithms, such as clustering, nearest neighbor search, and dimensionality reduction techniques, which are used extensively in this work. I decided to adopt this as the distance function for this reason.

Euclidean distance is a measure of the straight-line distance between two points in Euclidean space. It is derived from the Pythagorean theorem and is commonly used to calculate the distance between two points in a multi-dimensional space. Given two vectors (p, q) with n dimensions, the Euclidean distance $(dist(p,q))$ between them can be defined as shown in 4.2.

$$dist(p,q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \tag{4.2}$$

Equation 4.2 is modified for similarity as shown in equation 4.3 (SEGARAN (2007)). To avoid division errors and ensure that the maximum value is 1 when $(dist(p,q))$ is 0, 1 is added to the denominator, thus $0 \leq (sim(p,q)) \leq 1$.

$$sim\big(p,q\big) = \frac{1}{1+dist\big(p,q\big)} \tag{4.3}$$

For this task, to calculate the similarity of an AOI (A) with respect to an extracted subset of aneurysms S, we calculate the average of the similarity of every element in S with A as shown in equation 4.4.

$$sim\,(A,S) = \frac{\sum_{s_i \in S} sim\,(a,s_i)}{|S|} \tag{4.4}$$

Redundancy

Given two objects, s_1 and s_2, the extent to which s_1 is redundant with respect to s_2 can be estimated using sim(s_1, s_2) from equation 4.3. Since s_1 has replicated some of the information about s_2 and vice versa.

Given an extracted set S, the degree to which an object in the set s_1 is redundant with respect to the extracted set S can also be determined using this concept, as shown in equation 4.5. $\frac{\sum_{s \in S} sim(s_1,s)}{|S|-1}$ is the average similarity of s_1 with respect to S, this is then subtracted from 1 to get $Red\,(s_1,S)$.

$$Red\,(s_1,S) = \left(1 - \frac{\sum_{s \in S} sim\,(s_1,s)}{|S|-1}\right) \tag{4.5}$$

From equation 4.5, to calculate the redundancy in an extracted subset S, I take the average redundancy of every element in the set with respect to the subset as shown in equation 4.6.

$$Red(S) = \frac{\sum_{s_i \in S} Red\,(s_i,S)}{|S|} \tag{4.6}$$

Anomaly

The degree of anomaliness (or outlierness) of a point from an extracted set will be measured using local reachability distance (LRD). LRD is a measure used in the field of data mining and outlier detection to quantify the local density of a data point with respect to its neighbors. It is commonly associated with the Local Outlier Factor (LOF) algorithm (BREUNIG et al. (2000); LI et al. (2022)).

The LRD of a data point P is calculated based on the average reachability distance of P to its k-nearest neighbors. The reachability distance between two points, denoted as $reach-dist_k(P,Q)$, is defined as the maximum of the Euclidean distance between P and Q or the k-distance of Q. The k-distance of a point Q is the distance to its k_{th} nearest neighbor.

The formula for calculating the LRD of a data point P with respect to its k-nearest neighbors can be expressed as shown in equation 4.7.

$$LRD_k(P) = \left(\frac{1}{\frac{1}{k} \sum_{Q \in N_k(P)} reach-dist_k(P,Q)} \right) \tag{4.7}$$

The LRD measures the local density of a data point by considering the average reachability distance to its k-nearest neighbors. A low average reachability distance indicates that the point is located in a dense region, while a high average reachability distance suggests that the point is located in a sparse or outlier region. It is a useful metric in outlier detection algorithms like the Local Outlier Factor (LOF). The value for k for a dataset is determined by the same empirical analysis used for all density based approaches in this work. This is explained in detail in section 4.4.1.

After LRD has been calculated for all instances in the dataset, we normalise the data using equation 4.8 where x is the LRD of an instance in the dataset, so the value range is between [0, 1].

$$\mathrm{x}_{\mathrm{scaled}} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{4.8}$$

The degree of anomaly in an extracted set S is the average LRD for all instances s_i in the set.

$$Anom_k(S) = \frac{\sum_{s_i \in S} LRD_k(s_i)}{|S|} \tag{4.9}$$

4.5.2 Score

The metrics proposed in the previous subsection 4.5.1 can be aggregated to form a single score, which defines a rating for an extracted subset. The metrics in this aggregation can be weighted with a parameter w depending on various factors such as the sparsity of the dataset, domain knowl-

edge of the designer, priority for the metrics e.t.c. The equation can be adapted to suit the required needs.

$$Score = w(sim(A,S)) + w(Red(S)) + w(Anom_k(S)) \tag{4.10}$$

5

Experiments and Evaluation

5.1 RIS Evaluation

To evaluate the performance of the RIS model, I compared the quality of the extracted set with respect to the extracted sets of a similarity based extraction and randomly selected subset from the dataset. These sets were then evaluated using the equations proposed in 4.5. The process of this evaluation is described below for each instance to be evaluated:

- Select an instance randomly from the dataset
- User defined number of instances N to be extracted
- Extract a subset using RIS
- Extract the N most similar subset
- Select N instances randomly
- Determine the quality of extractions using equations in section 4.5

This evaluation approach is similar to that used in the research of HOCHBAUM und PATHRIA (1998).

Given that redundancy and similarity are the most important metrics for this task, the weights for $sim(A,S)$, $Red(S)$, $Anom_k(S)$ have been set as 0.4, 0.4, and 0.2, respectively for equation 4.10. Also, the dataset is sparse and this can significantly affect the score for anomaly, although the points may not be outliers. This can give a significant advantage to the random model when it selects points from denser areas.

5.1.1 Main Dataset Evaluation

Table 5.1 shows the performance of RIS in comparison to the similarity based model which optimizes for similarity, and a random model that should ideally optimize for redundancy on the main dataset. For the first experiment, I used $N = 3$ for the size of the extracted set, which represents the number of clusters determined from the clustering step, I also used 6 randomly selected instances as AOI. As seen from the table, RIS has the best $finalscore$ in 5 (1 tied score with similarity-based model) of the 6 AOI, similarity-based model has the best $finalscore$ in 2 (1 tied score with RIS model) out of 6, while the random model did not win for any instance.

Instances	**Model**	**Metrics**			
		Sim	Red	$Anom$	$FinalScore$
1	**RIS**	0.371	0.656	0.270	**0.465**
	Similarity	**0.450**	0.536	**0.286**	0.452
	Random	0.178	**0.837**	0.265	0.455
2	**RIS**	0.425	**0.807**	0.273	**0.550**
	Similarity	**0.669**	0.468	**0.478**	**0.550**
	Random	0.160	0.788	0.204	0.421
3	RIS	0.418	0.741	0.161	0.496
	Similarity	**0.625**	0.473	0.351	**0.509**
	Random	0.177	**0.781**	0.200	0.422
4	**RIS**	0.370	0.718	**0.100**	**0.454**
	Similarity	**0.457**	0.541	0.084	0.416
	Random	0.287	**0.785**	0.075	0.444
5	**RIS**	0.365	0.781	0.298	**0.518**
	Similarity	**0.457**	0.342	**0.569**	0.494
	Random	0.188	**0.774**	0.245	0.433
6	RIS	0.447	0.758	0.411	0.564
	Similarity	**0.751**	0.316	**0.923**	**0.612**
	Random	0.362	**0.890**	0.417	0.556

Table 5.1: Evaluation Results $N = 3$

I further experimented by comparing the performances using a higher value for N ($N = 5$) and the same AOI's used in the previous experiment. Table 5.2 shows the results of this experiment. As can be seen from the $finalscore$ column on this table, the similarity-based model has the best

score for 4 of 6 instances, RIS is the best for the remaining 2 instances. Again, the random model is not the best for any of the 6 instances.

Instances	Model	Metrics			
		Sim	Red	$Anom$	$FinalScore$
1	**RIS**	0.336	0.727	0.014	**0.454**
	Similarity	**0.420**	0.558	**0.248**	0.441
	Random	0.273	**0.738**	0.150	0.435
2	RIS	0.407	0.744	0.322	0.525
	Similarity	**0.605**	0.482	**0.601**	**0.560**
	Random	0.287	**0.782**	0.439	0.515
3	RIS	0.402	0.718	0.183	0.485
	Similarity	**0.546**	0.529	0.300	**0.490**
	Random	0.233	**0.796**	**0.379**	0.487
4	**RIS**	0.367	0.705	0.100	**0.448**
	Similarity	**0.419**	0.600	0.084	0.424
	Random	0.148	**0.798**	**0.241**	0.423
5	RIS	0.393	0.680	0.352	0.500
	Similarity	**0.482**	0.443	**0.524**	**0.515**
	Random	0.274	**0.759**	0.142	0.442
6	RIS	0.445	0.717	0.460	0.557
	Similarity	**0.700**	0.389	**0.832**	**0.600**
	Random	0.253	**0.789**	0.305	0.478

Table 5.2: Evaluation Results $N = 5$

5.1.2 Supplementary Dataset Evaluation

I conducted further evaluation of the RIS model using the supplementary dataset described in section 4.3.1. After pre-processing and cleaning, I conducted the same experiment on this data using RIS, a similarity-based model and a random selection model.

Table 5.3 shows the scores of these models using $N = 3$ and 7 randomly selected samples as AOI's. As can be seen from the table, RIS performs better than the other extracting the best subset of instances for all 7 AOI's. This outcome is similar to the results achieved on the main dataset for $N = 3$ (Table 5.1).

Table 5.4 shows the results of the evaluation on the supplementary dataset with $N = 5$ using the same sample of AOI's. Unlike the results for $N = 5$

Instances	**Model**	**Metrics**			
		Sim	Red	$Anom$	$FinalScore$
1	**RIS**	0.442	0.671	**0.671**	**0.579**
	Similarity	**0.534**	0.540	0.638	0.557
	Random	0.229	**0.867**	0.619	0.562
2	**RIS**	0.389	0.685	**0.613**	**0.552**
	Similarity	**0.527**	0.440	0.480	0.482
	Random	0.200	**0.872**	0.583	0.544
3	**RIS**	0.347	0.744	0.552	**0.547**
	Similarity	**0.435**	0.444	0.501	0.452
	Random	0.145	**0.878**	0.687	**0.547**
4	**RIS**	0.370	0.785	0.604	**0.583**
	Similarity	**0.598**	0.389	**0.720**	0.539
	Random	0.192	**0.788**	0.638	0.520
5	**RIS**	0.334	0.733	**0.668**	**0.560**
	Similarity	**0.483**	0.645	0.517	0.555
	Random	0.204	**0.829**	0.622	0.538
6	**RIS**	0.499	0.585	**0.721**	**0.578**
	Similarity	**0.592**	0.451	0.714	0.560
	Random	0.237	**0.760**	0.642	0.527
7	**RIS**	0.469	0.717	0.674	**0.609**
	Similarity	**0.674**	0.311	**0.759**	0.546
	Random	0.141	**0.879**	0.500	0.508

Table 5.3: Evaluation Results Supplementary Dataset $N = 3$

on the main dataset (Table 5.2) RIS performs the better for $finalscore$ compared to the other two models on the supplementary dataset for N = 5. RIS is has the best $finalscore$ for 5 (3 ties with the similarity-based model) of 7 instances, similarity-based model is the winning model for 3 (3 ties with the RIS model), and the random model is best for 1 instance.

Instances	Model	Metrics			
		Sim	Red	$Anom$	$FinalScore$
1	**RIS**	0.427	0.644	0.651	**0.559**
	Similarity	**0.493**	0.595	0.585	0.552
	Random	0.142	**0.861**	**0.655**	0.532
2	**RIS**	0.392	0.692	0.577	**0.549**
	Similarity	**0.486**	0.515	0.538	0.508
	Random	0.019	**0.794**	**0.642**	0.522
3	**RIS**	0.301	0.741	0.525	**0.522**
	Similarity	**0.393**	0.533	0.539	0.478
	Random	0.136	**0.869**	**0.555**	0.513
4	**RIS**	0.388	0.714	0.629	**0.566**
	Similarity	**0.556**	0.512	**0.699**	**0.566**
	Random	0.122	**0.855**	0.633	0.517
5	RIS	0.321	0.758	0.651	0.562
	Similarity	**0.437**	0.616	0.596	0.541
	Random	0.226	**0.808**	**0.753**	**0.564**
6	**RIS**	0.480	0.616	0.700	**0.578**
	Similarity	**0.551**	0.543	**0.705**	**0.578**
	Random	0.185	**0.698**	0.539	0.539
7	**RIS**	0.457	0.672	0.680	**0.588**
	Similarity	**0.630**	0.423	**0.747**	**0.588**
	Random	0.192	**0.792**	0.634	0.520

Table 5.4: Evaluation Results Supplementary Dataset N = 5

From these results, it can be assumed that the larger the dataset, the better the performance of RIS because there is a larger search space to select samples. Also, the outlier removal step will be more useful because while RIS handles outliers, the other two models do not. A larger dataset is likely to emphasize the importance of this step because it contains more outliers and the effect of this can be seen in the improved $Anom$ scores for RIS in the supplementary dataset in contrast to the main dataset. This effect

on more data samples on the performance of *RIS* should be an expected outcome given that for most ML algorithms, more data is usually better.

5.2 Instance Selection Adaptation Evaluation

Most instance selection tasks are evaluated by comparing the classification accuracy of a ML model trained on an extracted subset and one trained on the complete dataset (HUANG et al. (2021); LIN et al. (2017); PAN et al. (2005)). I also use this approach to evaluate the performance of the proposed IS adaptation part of this work.

To evaluate this task, I first train a classification model using XGBoost on the full dataset to ascertain its accuracy, I then train several models using various percentages of a representative subset of samples extracted from the training set using the proposed *RIS* adaptation. This model is also evaluated by training multiple models extracted by random sampling from the training set.

The datasets used for this evaluation are the iris dataset and the supplementary dataset introduced in section 4.3.1. I split it into train and test set using the 80/20 ratio. Samples are then drawn from the training set using the proposed *RIS* adaption and also randomly selected samples. To score the performance of the model trained using a randomly selected subset, for every percentage threshold of subset size, 5 models are built using random samples drawn 5 different times with replacement, the average performance of the models trained using the randomly selected samples is then taken as the score for a specific threshold. Rupture state was used as the class label for the supplementary dataset.

All models built for this test used the same hyper-parameters to avoid bias. Table 5.5 and table 5.6 show the results of this experiment on the iris and supplementary datasets respectively.

On the iris dataset, the proposed *RIS* adaptation has a better classification accuracy on the experiment with smaller percentages of extracted representative set, while the random model is the best for all percentages of the representative set on the supplementary dataset.

Representative Set (%)	**RIS adaptation**	**Random Sampling**
12.5	**83.3**	77.34
30	**96.7**	94.68
50	96.7	**99.28**

Table 5.5: Evaluation results of proposed IS adaptation on iris dataset. The table shows the accuracy on the test data of the models trained using different percentages of representative set extracted from the training set.

Representative Set (%)	**RIS adaptation**	**Random Sampling**
12.5	45.07	**49.86**
30	53.52	**56.34**
50	54.93	**56.62**

Table 5.6: Evaluation results of proposed IS adaptation on supplementary dataset. The table shows the accuracy on the test data of the models trained using different percentages of representative set extracted from the training set.

6 Analysis of Results

6.1 RIS

6.1.1 Main Dataset

In this section, I take a deep dive into the analysis of the results shown in the RIS evaluation section (section 5.1). As shown in section 5.1.1, with $N = 3$ and $N = 5$ for the main dataset, the RIS model and the similarity-based model were the respective winning models with respect to the $finalscore$ (see table 5.1 and table 5.2 respectively). This dataset is a sparse dataset with 70 instances. Table 6.1 is the summary of the results of all experiments performed on the main dataset, the full results are shown in table 5.1 and table 5.2. This summary is the performance of the models on all instances, this calculated by taking the mean scores for each metric for every model.

Extraction of 3 samples ($N = 3$)

For this experiment, I selected $N = 3$ as the size of the extracted set, which is also the number of clusters from the k-means clustering of the dataset shown in section 4.4.2. From figure 6.1, for the similarity score, the similarity-based model, as expected is consistently the best for every instance, while the random model is consistently the worst (Also see 6.1, average similarity score for RIS, Similarity-based model and Random model are 0.4, 0.568 and 0.225 respectively). RIS on the other hand remains consistently second best for the similarity metric, this is a desirable outcome because, for the purpose of this task, the objective is not to ex-

Table 6.1: Experiment Summary on Main Dataset

Model	Main Dataset: $N = 3$			
	SIM	*RED*	*ANOM*	*FINALSCORE*
RIS	0.400	0.744	0.252	**0.508**
Similarity	**0.568**	0.446	**0.449**	0.506
Random	0.225	**0.809**	0.235	0.455
Model	Main Dataset: $N = 5$			
	SIM	*RED*	*ANOM*	*FINALSCORE*
RIS	0.392	0.715	0.239	0.495
Similarity	**0.529**	0.500	**0.431**	**0.505**
Random	0.245	**0.777**	0.276	0.463

tract the most similar instances with respect to an AOI because diversity in selection is also desirable, as explained in previous chapters.

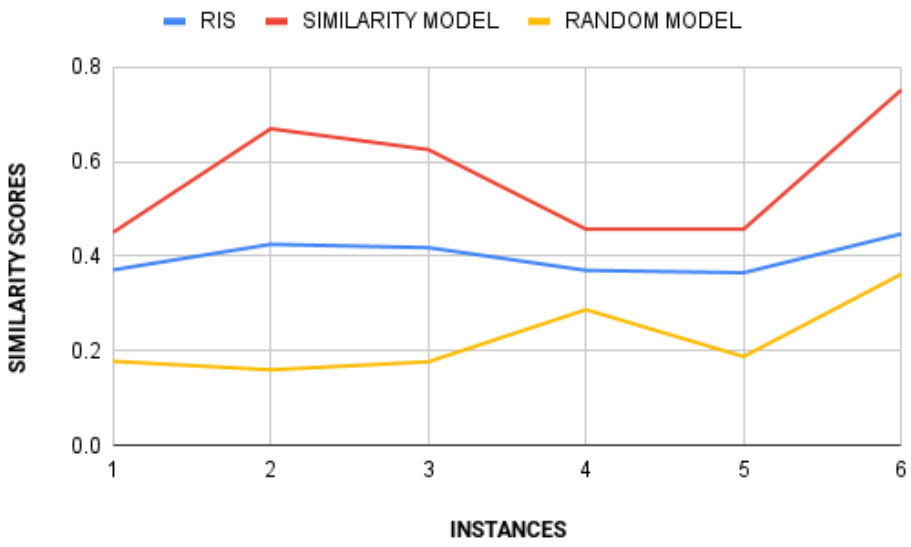

Figure 6.1: Line plot of the performance of each model on the selected AOI instances on the main dataset for $N = 3$ for similarity metric.

As shown in figure 6.2 for the redundancy metric, which scores for diversity, the random model is consistently better for all instances except the two (instance 2 and 5) while the similarity-based model is usually the worst model by a considerable amount for all instances (See also 6.1,

average redundancy scores for RIS, Similarity-based model and Random model are 0.744, 0.446, and 0.809 respectively). Interestingly, RIS performs very well for this metric, this is a good outcome as diversity of the extracted set is a criteria for selecting a representative and diverse subset.

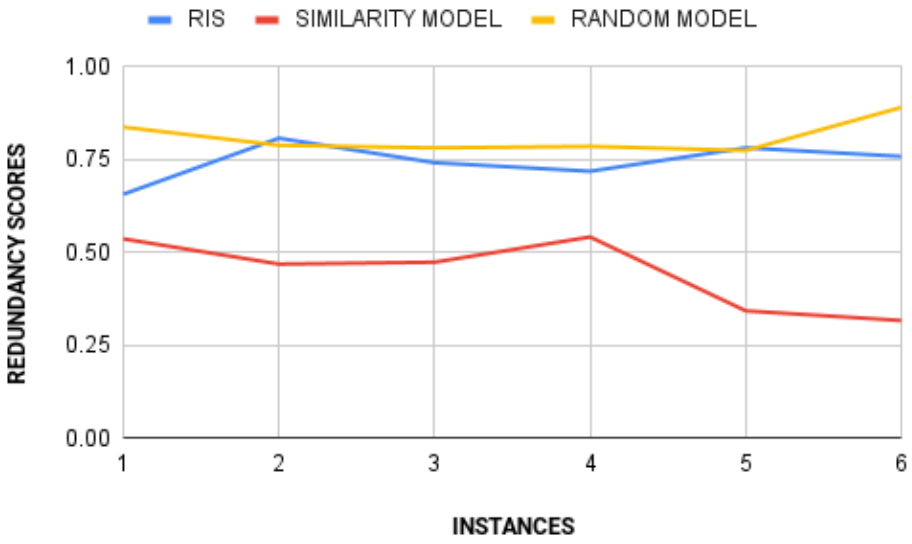

Figure 6.2: Line plot of the performance of each model on the selected AOI instances on the main dataset for $N = 3$ for redundancy metric.

The similarity-based model dominates the anomaly metric compared to the other two models, as seen in figure 6.3 and table 6.1 (Anomaly scores for RIS, Similarity-based and Random models are 0.252, 0.449 and 0.235 respectively). The similarity-based model always selects the closest points to the AOI, this implicitly optimizes the LRD which calculates anomaly scores based on the distance of a point to its k-closest points. The sparsity of the dataset amplifies the effects of this extraction technique because most points will be farther from each other, thus selecting the most similar in a certain area will improve the LRD depending on the value of k and the size of the extracted set. Despite selecting points having no pattern to its selection process, the Random model scores are similar to that of RIS, with an average of 0.235 and 0.252 respectively across all AOI (See table 6.1.

The final score plot (Figure 6.4) shows that while the random model is clearly the worst performing, the other two models are relatively close

Figure 6.3: Line plot of the performance of each model on the selected AOI instances on the main dataset for $N = 3$ for anomaly metric.

with RIS slightly better (Average scores for RIS, Similarity-based and random models are 0.508, 0.506, 0.455 respectively). This is because of the dominance of the similarity based model for the anomaly metric and the weighting of each metric. A lower weight for the anomaly would amplify the performance of RIS compared to the similarity based model.

RIS does not consistently perform terribly on the individual metrics $sim(A,S)$, $Red(S)$, $Anom_k(S)$ unlike the other models; like the similarity based model which consistently have the worst $Red(S)$ scores for all 6 instances or the random selection model which have the worst $sim(A,S)$ scores for all 6 instances.

Extraction of 5 samples ($N = 5$)

This experiment was performed to examine how the models perform when the size of the desired extracted set is greater than the number of clusters. For this experiment, $N = 5$. From figure 6.5, figure 6.6, and figure 6.7 we see that the similarity, redundancy and anomaly plots respectively, follows the same pattern as their corresponding counterparts in the experiment with $N = 3$ which have been discussed in section 6.1.1.

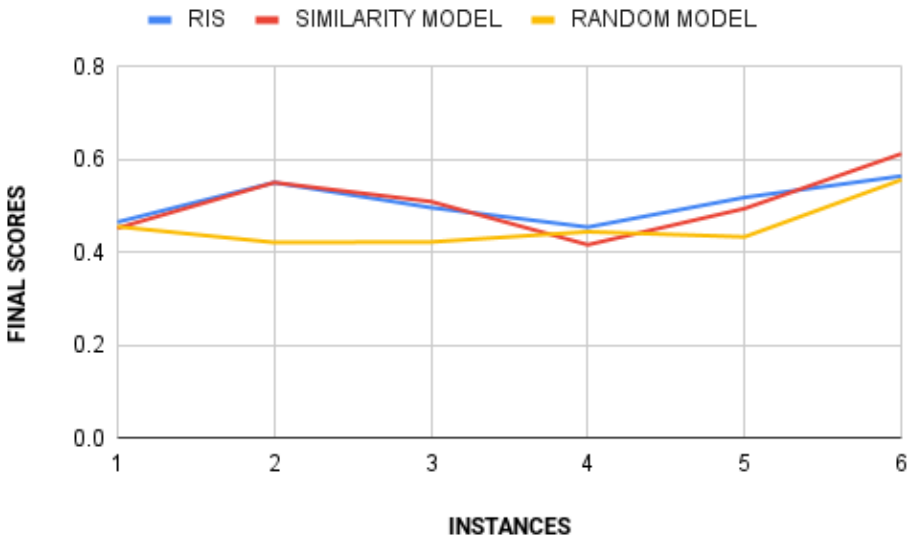

Figure 6.4: Line plot of the performance of each model on the selected AOI instances on the main dataset for $N = 3$ for final score.

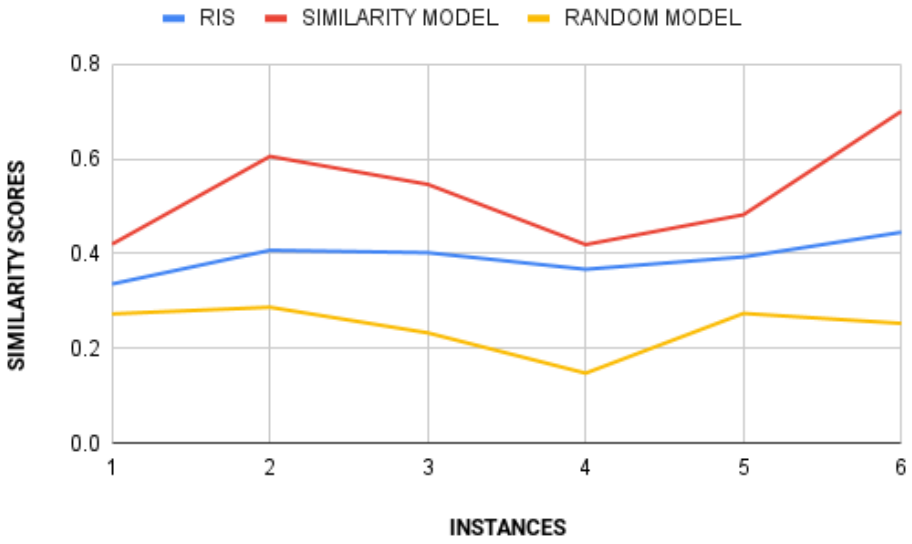

Figure 6.5: Line plot of the performance of each model on the selected AOI instances on the main dataset for $N = 5$ for similarity metric.

Although the results for redundancy for this experiment also follow the pattern from the experiment with $N = 3$, it can be seen that the redun-

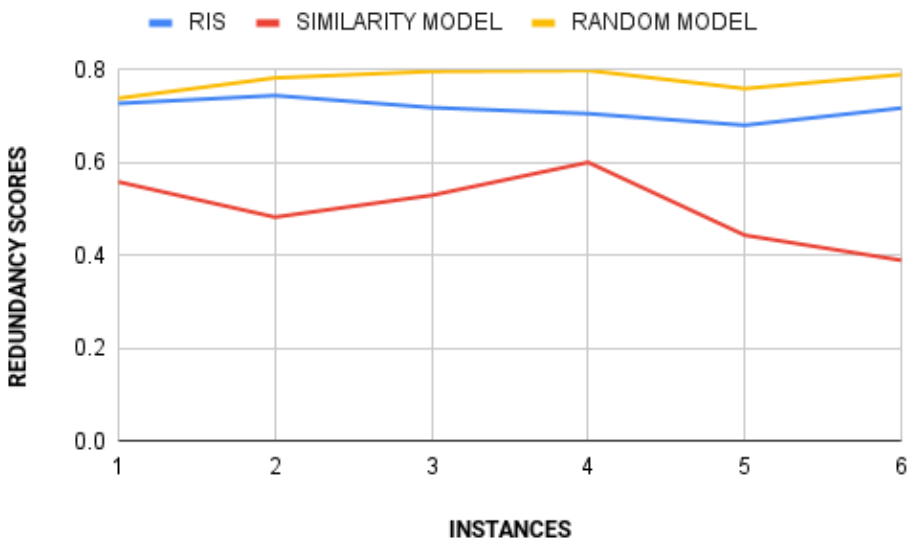

Figure 6.6: Line plot of the performance of each model on the selected AOI instances on the main dataset for $N = 5$ for redundancy metric.

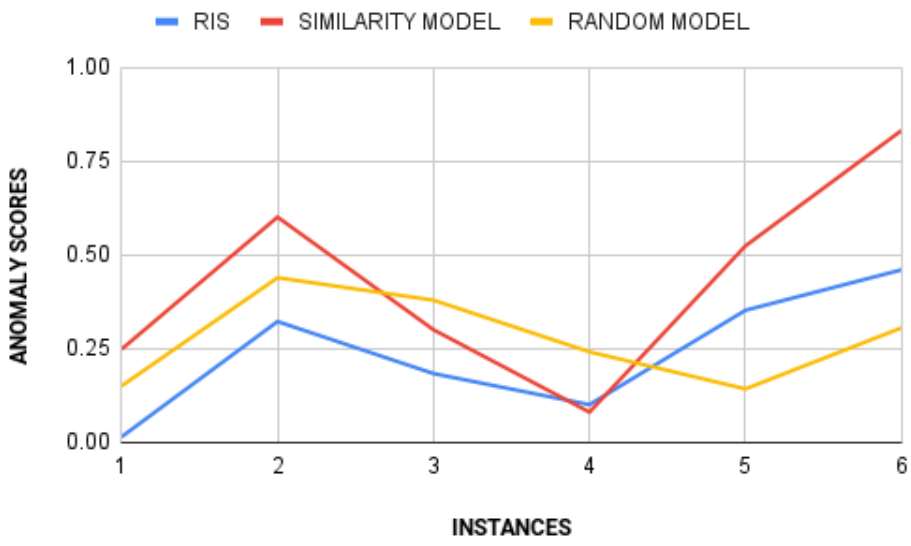

Figure 6.7: Line plot of the performance of each model on the selected AOI instances on the main dataset for $N = 5$ for anomaly metric.

dancy scores for the similarity-based model (Average redundancy score is 0.446 and 0.5 for $N = 3$ and $N = 5$ respectively, see 6.1) have improved

with increase in the size of the extracted set. The dominance of the RIS and the random model have reduced (Table 6.1 shows the respective values). This is because an increase in the size of the extracted set means the similarity-based model is forced to extract less similar instances. Given that the dataset is sparse, there is a higher probability it selects less similar instances thus improving its redundancy.

Similar to the experiment with $N = 3$, RIS still maintains a good balance on the Sim and Red metrics, it does not consistently have the worst scores for these metrics (See table 6.1 for the average scores of all models for the experiments), unlike the random model which has the worst Sim score for all the instances and the similarity-based model with the worst Red scores for all 6 instances.

The improvement in redundancy for the similarity-based model translates to an improvement in the final score compared to the random model and RIS (See figure 6.8). The similarity based model has the best $finalscore$ for 4 out of the 6 instances for this experiment on a sparse dataset. I analyse the change in scores for all metrics and the final score in section 6.1.1

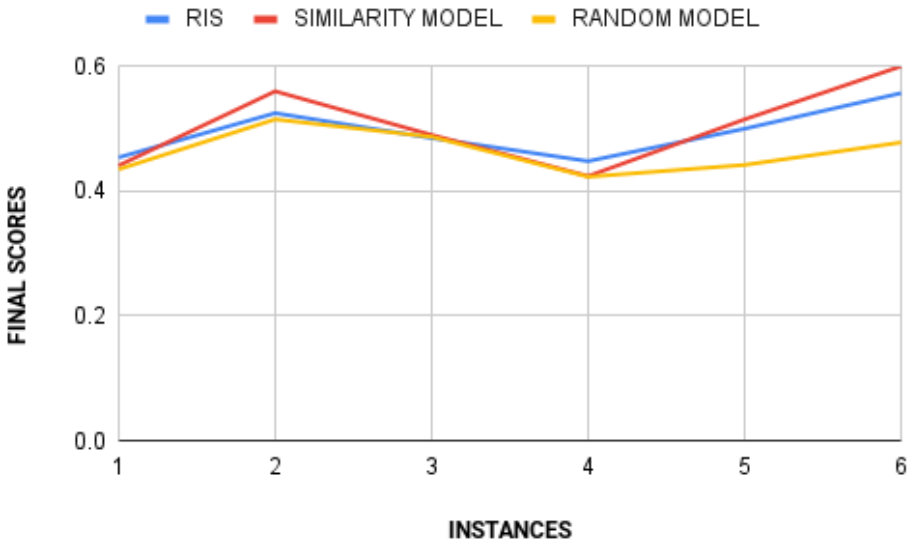

Figure 6.8: Line plot of the performance of each model on the selected AOI instances on the main dataset for $N = 5$ for final score.

The RIS model was not the best model for extracting instances greater than the number of clusters in this dataset using the weights assigned to

the final score equation discussed earlier. It was the best in only 2 out of 6 instances. However, the final score can be improved by using another method to extract the excess instance when N is greater than the number of clusters; for example, when N is greater than number of clusters, the excess instances can be extracted by using a hybrid model which uses a similarity based approach for extracting remaining instances or extracting them from the same cluster as the given AOI.

Examining the Change in Scores ($N = 3$ - $N = 5$)

In this section, I analyze the difference in scores for both experiments to determine why there were different winners for the experiments conducted on the same dataset. This was calculated by subtracting the scores of the experiment conducted with $N = 5$ from those of the experiment conducted with $N = 3$ ($N = 3$ - $N = 5$). A positive result implies that there is a drop in scores from $N = 3$ to $N = 5$, while a negative result implies the opposite. This is necessary to analyse the how the size of N influences the scores for the respective models.

In Figure 6.9, from the change in similarity plot, we can see that there is a drop in similarity scores for the RIS and similarity-based models for 5 out of the 6 instances. Intuitively, the larger the size of the extracted set, the higher the potential of selecting less similar instances, this is the reason why there is a drop in the similarity scores for the RIS and similarity-based models. The drop for similarity-based model is significantly larger, this is because when the size of the extracted set is smaller, the selections are the closer to the AOI and hence higher similarity scores. When the size of the extracted set increased and less similar instances are extracted, this yields a drop in the score.

As can be seen from the change in redundancy plot in figure 6.10, there is a significant increase in the redundancy score for the similarity-based for all 6 instances, this can also be explained by the inclusion of less similar instances to an increased extracted set. The inclusion of less similar instances to the extracted set introduce diversity to the extracted set and led better scores for redundancy. On the other hand, the RIS model had a drop in redundancy scores for 5 out of the 6 instances, since the algorithm now have to transverse each cluster multiple times when the value for N

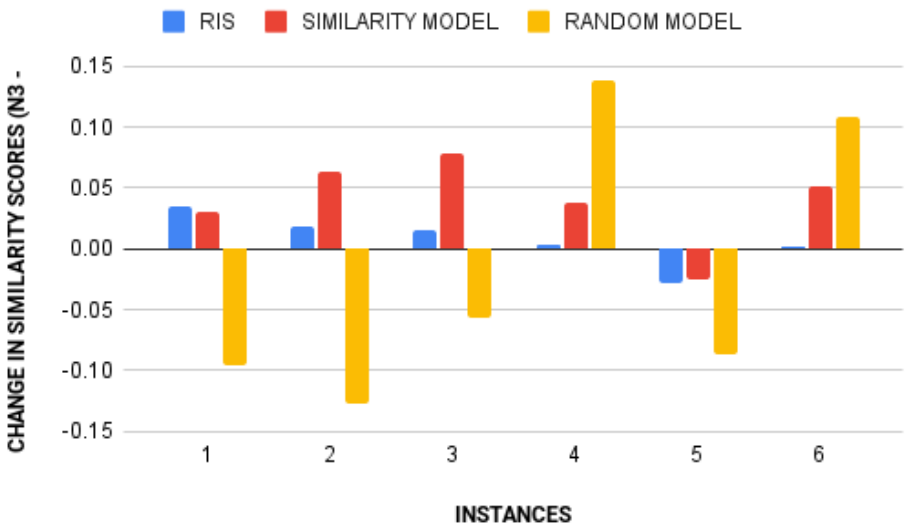

Figure 6.9: Bar plot of the change in scores of similarity metric for each instance on the main dataset for all models.

is greater than the number of clusters, it selects instances similar to an already selected instance. This leads to a decrease in the redundancy score.

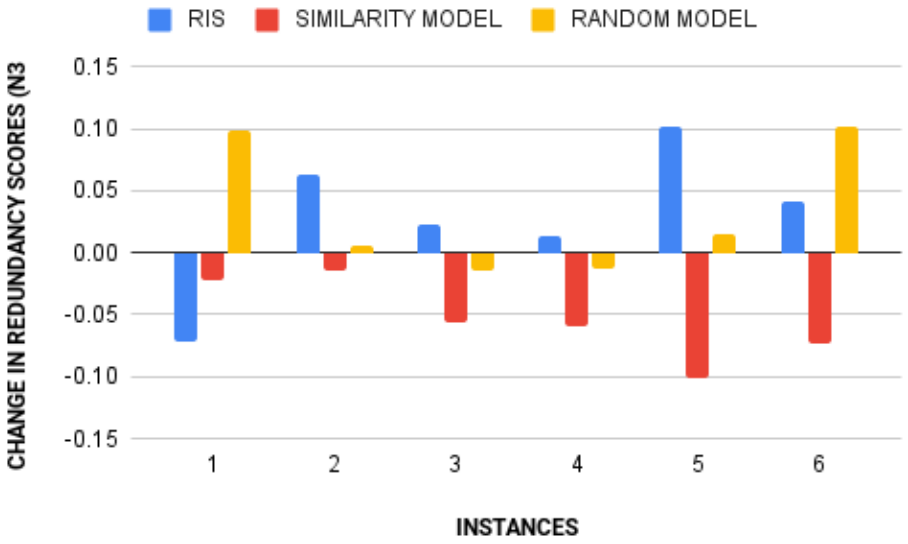

Figure 6.10: Bar plot of the change in scores of redundancy metric for each instance on the main dataset for all models.

There is a small increase in the anomaly scores for the RIS model on 4 of the 6 instances as can be seen from the change in anomaly score plot in figure 6.11. This means the model selects samples from points in dense area. However, there is a significant reduction in the anomaly score for instance 1, the model selected points in sparse regions when the size of the extracted set was increased for this instance.

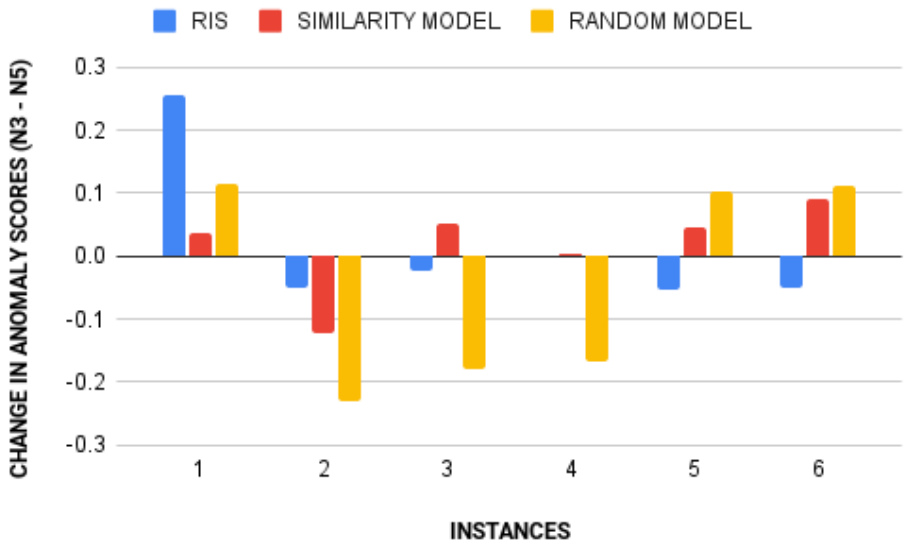

Figure 6.11: Bar plot of the change in scores of anomaly metric for each instance on the main dataset for all models.

The accumulation of these changes in each metric led to the change in the final score as can be seen on the final score plot in figure 6.12. From this analysis, it may be interesting to change the weights for the final score with an increase in the size of N. For example, intuitively, the larger the size of N, the greater the chance of having diversity in the extracted set, thus, it can be useful to reduce the weights for redundancy when the size of N is larger than a certain threshold. The decision on this threshold is dependent on the dataset, it's characteristics and the domain.

It is not useful to analyze the random model in this context because there is no pattern to its selection. It can extract a completely different subset for several iterations with the same instance as AOI.

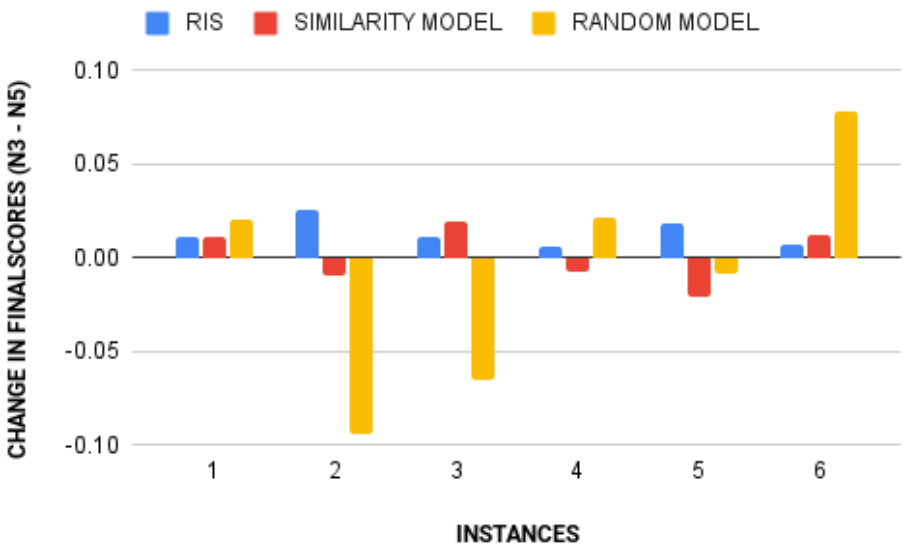

Figure 6.12: Bar plot of the change in scores of final for each instance score on the main dataset for all models.

Across all experiments performed for $N = 3$ and $N = 5$ on the main dataset, RIS model was the best for 6 of the 12, while the similarity-based model was the best for the other 6. RIS dominated the $N = 3$ experiment while the similarity-based model dominated the $N = 5$ experiment. Thus for sparse datasets, RIS is better for extracting representative but diverse samples when the size of N is small, as N starts to increase, it can be beneficial to extract based on similarity.

6.1.2 Supplementary Dataset

In this section, I do an in-depth analysis of the results shown in the RIS evaluation section (section 5.1). As shown in section 5.1.2, with $N = 3$ and $N = 5$ for the supplementary dataset, the RIS model was the best model with respect to the $finalscore$ for both experiments (see table 5.3 and 5.4). This supplementary dataset contains more instances (351 instances), it is necessary to see how the models perform on a larger dataset denser than the the main dataset (70 instances). As discussed in section 6.1.1, table 6.2 shows a summary of all experiments performed with the supplementary as shown in chapter 5 with table 5.3 and table 5.4.

Table 6.2: Experiment Summary on Main Dataset

Model	Supplementary Dataset: $N = 3$			
	SIM	*RED*	*ANOM*	*FINALSCORE*
RIS	0.407	0.703	**0.643**	**0.573**
Similarity	**0.549**	0.46	0.618	0.527
Random	0.193	**0.839**	0.613	0.535
Model	Supplementary Dataset: $N = 5$			
	SIM	*RED*	*ANOM*	*FINALSCORE*
RIS	0.395	0.691	0.630	**0.561**
Similarity	**0.507**	0.534	0.630	0.544
Random	0.146	**0.811**	0.630	0.530

Extraction of 3 samples ($N = 3$)

Figure 6.13 shows the scores of the similarity metric for 7 AOI's for the 3 models with $N = 3$. As can be seen from this figure, it follows the same pattern seen on the two experiments conducted on the main dataset (Figure 6.1 and figure 6.5) where the similarity-based model has the best scores across all instances, the *RIS* model is second across all instances, and the random model is consistently the least performing model (See also table 6.2, the average similarity scores for *RIS*, Similarity-based model and Random model are 0.407, 0.549, and 0.193 respectively). The patterns for the similarity metric is clearly not affected by the size of the dataset. The *RIS* model getting the second best similarity scores (behind the similarity-based model which optimizes for similarity) across all instances is a desirable outcome because we do not want the extracted set to contain the most similar instances which makes some instances in the extracted set redundant.

Again, the same patterns noticeable in the in the two experiments on the main dataset (Figure 6.2 and figure 6.6) is also seen in figure 6.14. For all AOI's the random model have the highest scores for the redundancy metric, second is *RIS* and the similarity-based model has the worst scores (Average scores for redundancy are 0.703, 0.46, and 0.839 for *RIS*, Similarity-based model and Random model respectively. See table 6.2). It shows that the higher the similarity scores the worse the redundancy scores. This

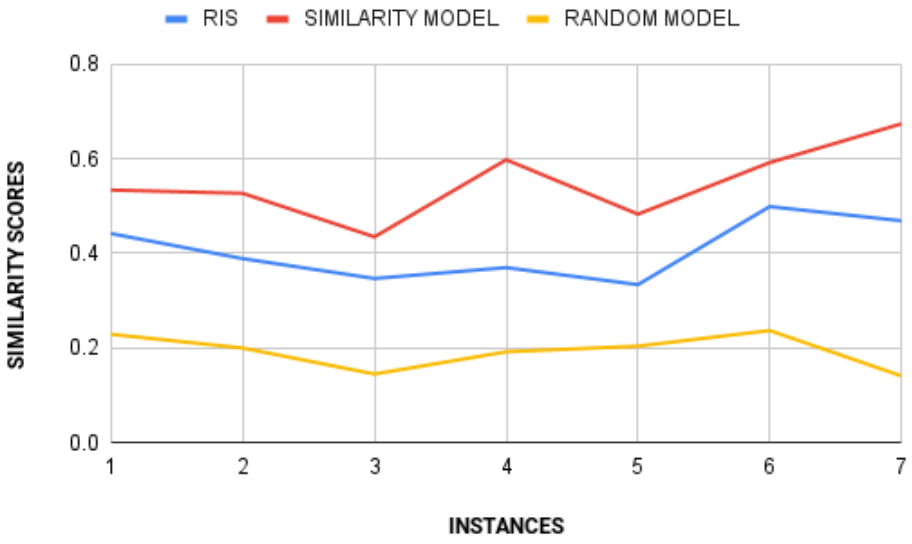

Figure 6.13: Line plot of the performance of each model on the selected AOI instances on the supplementary dataset for $N = 3$ for similarity metric.

metric is also not dependent on the size of the dataset as it shows same patterns for both a sparse and dense dataset.

Figure 6.14: Line plot of the performance of each model on the selected AOI instances on the supplementary dataset for $N = 3$ for redundancy metric.

Unlike the other two metrics, there is a clear change in pattern for the anomaly metric when the dataset is dense (See figure 6.15). All the models gets relatively higher scores for this dataset compared to the sparse main dataset and there are no clearly discernable patterns in these scores. This is because for a denser dataset more points will get better LRD scores as there will be more points close together, unlike a sparse dataset where points will be farther from each other. (Average scores for anomaly are 0.643, 0.618, and 0.613 for RIS, Similarity-based model and Random model respectively. See table 6.2).

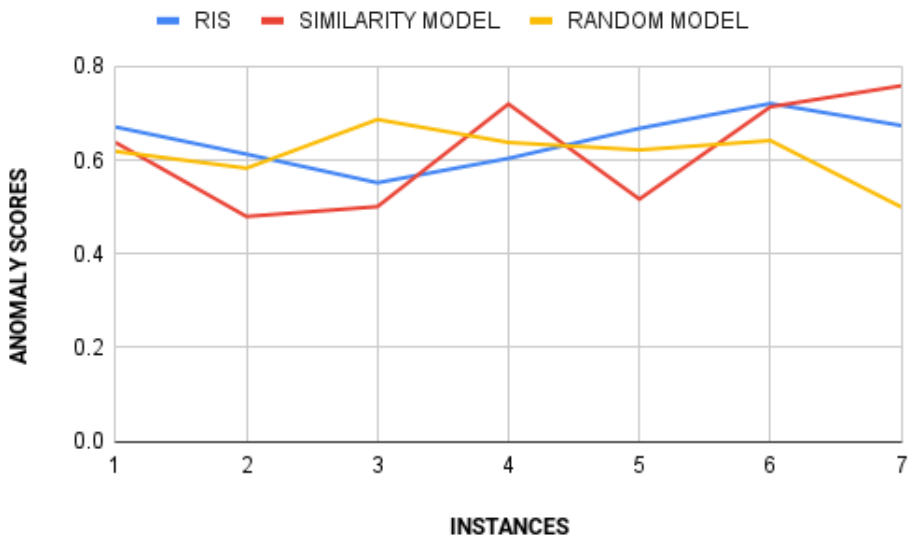

Figure 6.15: Line plot of the performance of each model on the selected AOI instances on the supplementary dataset for $N = 3$ for anomaly metric.

The final scores for this experiment shown in figure 6.16 also follows the patterns from the experiments on the main dataset where the scores are tight between the models. However, RIS is consistently the best model for all 7 instances using the weights assigned for each metric. LIke the experiment on the main dataset, RIS is best again when the size of N is 3.

Extraction of 5 samples ($N = 5$)

I also performed the experiment with a larger size of the extracted set ($N = 5$) to compare how the models react to a dense dataset when N is greater

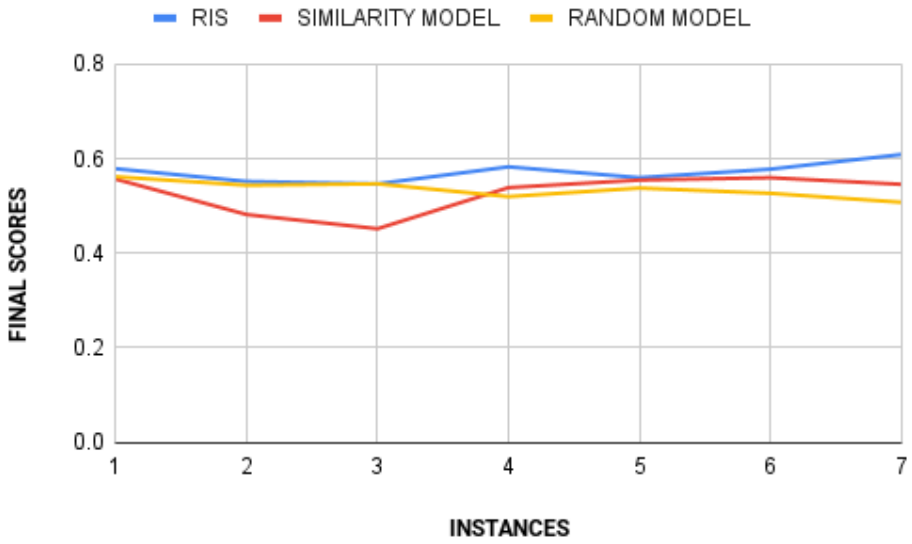

Figure 6.16: Line plot of the performance of each model on the selected AOI instances on the supplementary dataset for $N = 3$ for final score.

than the number of clusters. Figure 6.17 shows the scores for the model for 7 randomly selected AOI's. This follows the same patterns as the previous 3 experiments and the reasons discussed in previous sections hold here.

The redundancy scores for the models also follow similar patterns to the previous 3 experiments, however there seems to be an improvement on the scores for the similarity-based model when $N = 5$. The change in scores between the two experiments on the supplementary dataset is discussed in subsequent sections for all metrics.

Figure 6.19 shows the results of the anomaly metric for all models with $N = 5$. It further confirms the results from the experiment with $N = 3$ that anomaly scores are relatively higher for dense datasets. I also compare the difference in anomaly scores for both experiments on the supplementary dataset in subsequent sections.

The final scores for all models in this experiment (See figure 6.20) are also relatively close with respect to each AOI, this is similar to previous experiments. However, unlike the experiment with $N = 5$ on the main dataset, RIS is the best for all instances, although, it ties thrice with the similarity-based model, and once with the random model. It can be inferred that

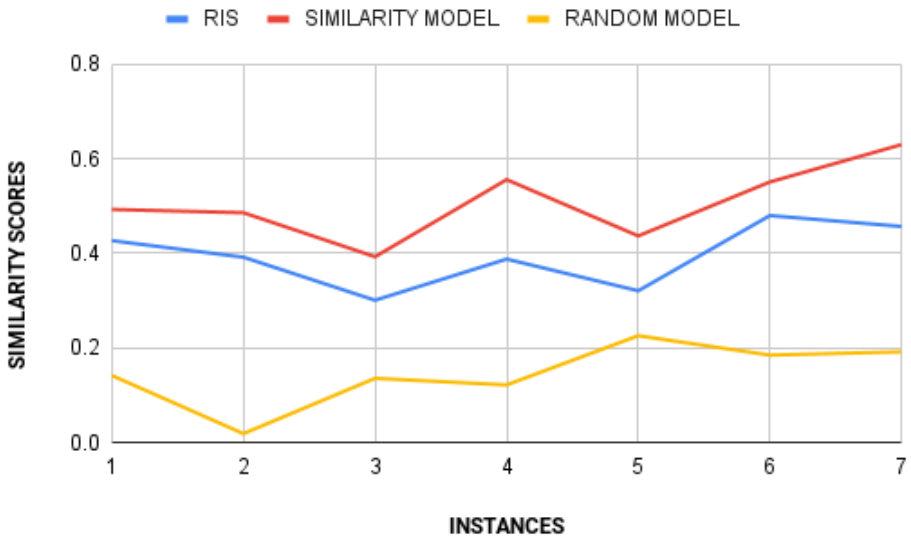

Figure 6.17: Line plot of the performance of each model on the selected AOI instances on the supplementary dataset for $N = 5$ for similarity metric.

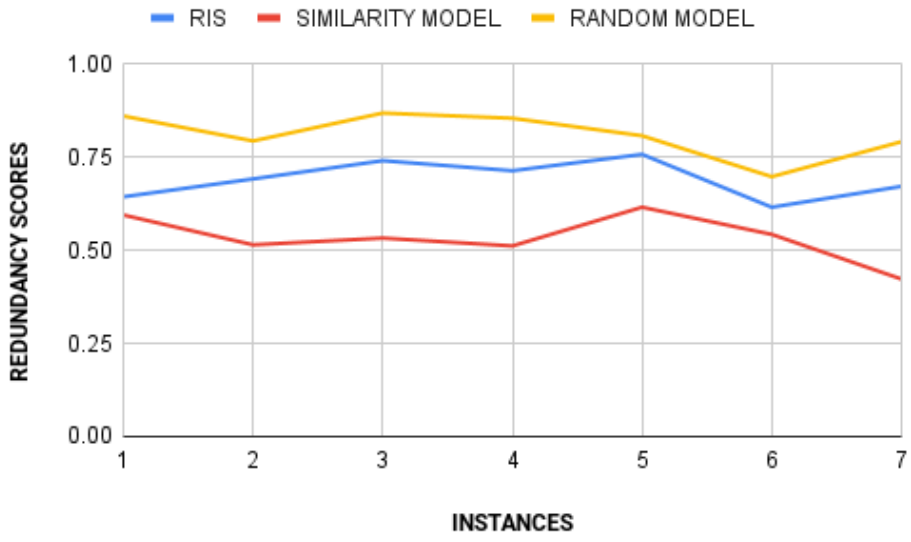

Figure 6.18: Line plot of the performance of each model on the selected AOI instances on the supplementary dataset for $N = 5$ for redundancy metric.

RIS gets better results for the task of extracting representative but diverse

Figure 6.19: Line plot of the performance of each model on the selected AOI instances on the supplementary dataset for $N = 5$ for anomaly metric.

samples with N greater than the number of clusters when the dataset is dense compared to when the dataset is sparse like the main dataset.

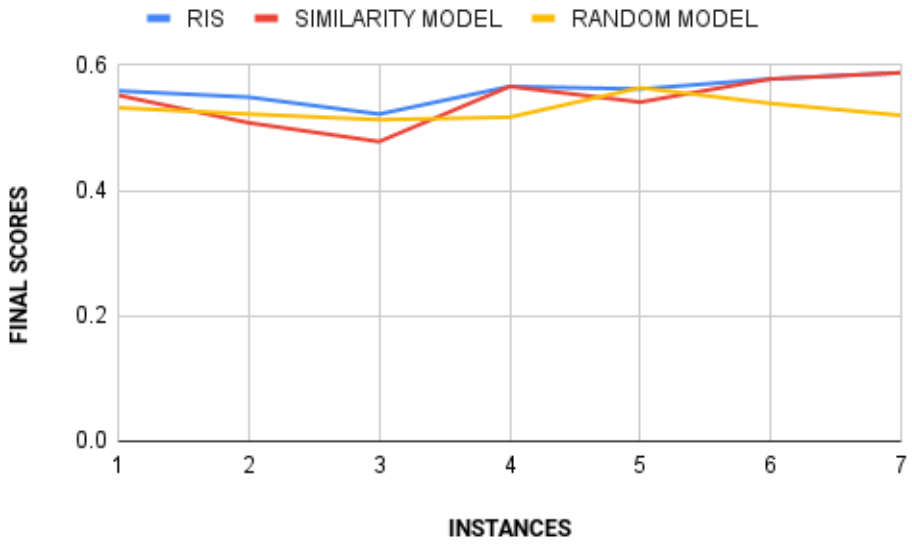

Figure 6.20: Line plot of the performance of each model on the supplementary dataset for $N = 5$ for final score.

Examining the Change in Scores ($N = 3$ - $N = 5$)

In this section, I also examine the difference between the $N = 3$ and $N = 5$ on the supplementary dataset. This was also estimated by subtracting the scores of the experiment conducted with $N = 5$ from those of the experiment conducted with $N = 3$ ($N = 3$ - $N = 5$). A positive result implies that there is a drop in scores from $N = 3$ to $N = 5$, while a negative result implies the opposite. For this analysis, there is no need to deeply analyze the change in the random model as there is no clearly defined pattern to its extraction, hence comparing it to a previously result is not interesting.

In Figure 6.21, from the plot for change in similarity, we can see that the drop in similarity scores for the RIS and similarity-based models from $N = 3$ to $N = 5$ for the supplementary dataset is consistent with the drop in the main dataset (See figure 6.9). As explained in section 6.1.1, it holds that the larger the size of the extracted set the lower the similarity scores between the extracted set and the AOI.

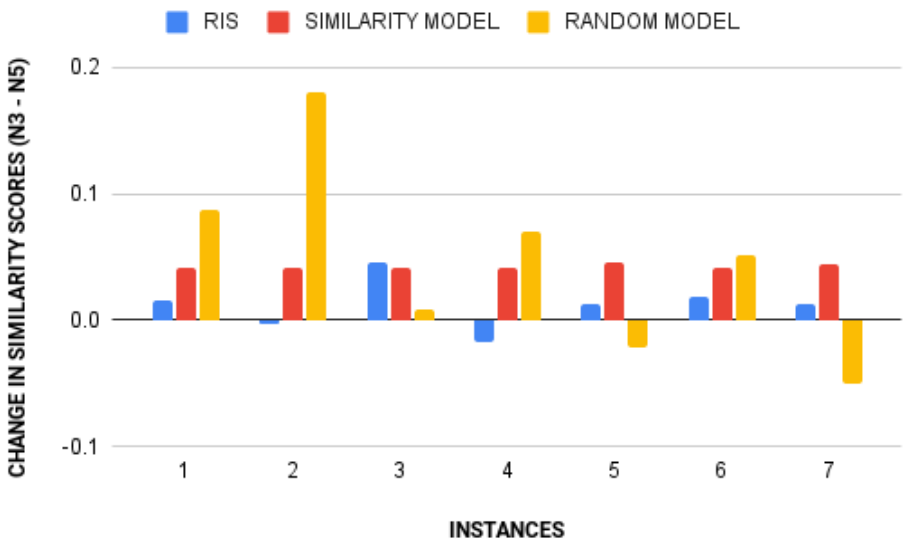

Figure 6.21: Bar plot of the change in scores of similarity metric for each instance on the supplementary dataset for all models.

The change in redundancy plot in figure 6.22 also follows the pattern of its main dataset counterpart (Figure 6.10). There is a significant increase in the redundancy score for the similarity-based for all 7 instances, by the ad-

dition of less similar instances to a larger extracted set. Diversity in the extracted set is improved by selecting more instances leading to better scores for redundancy. On the other hand, the RIS model had worse redundancy scores for 5 out of the 7 instances, this is because algorithm transverses each cluster multiple times when the value for N is greater than the number of clusters, it selects instances similar to an already selected instance. This leads to a decrease in the redundancy score.

Figure 6.22: Bar plot of the change in scores of redundancy metric for each instance on the supplementary dataset for all models.

From the plot for the change in anomaly scores shown in figure 6.23, there is no consistent change in anomaly scores across the 7 instances used as AOI's for the similarity-based model, 4 instances had a reduction in anomaly scores, while 3 had improved scores for the $N = 5$ experiment. The RIS model on the other hand, had a reduction in anomaly scores for 5 instances and an improvement on 2 instances.

As can be seen from figure 6.24 there was a significant improvement in the redundancy scores of the similarity-based model for $N = 5$ on the supplementary dataset, as was also observed on the main dataset (See figure 6.12). However, this change was not enough to alter the dominance of RIS as the winning model for this experiment on a dense dataset like it did on the sparse main dataset.

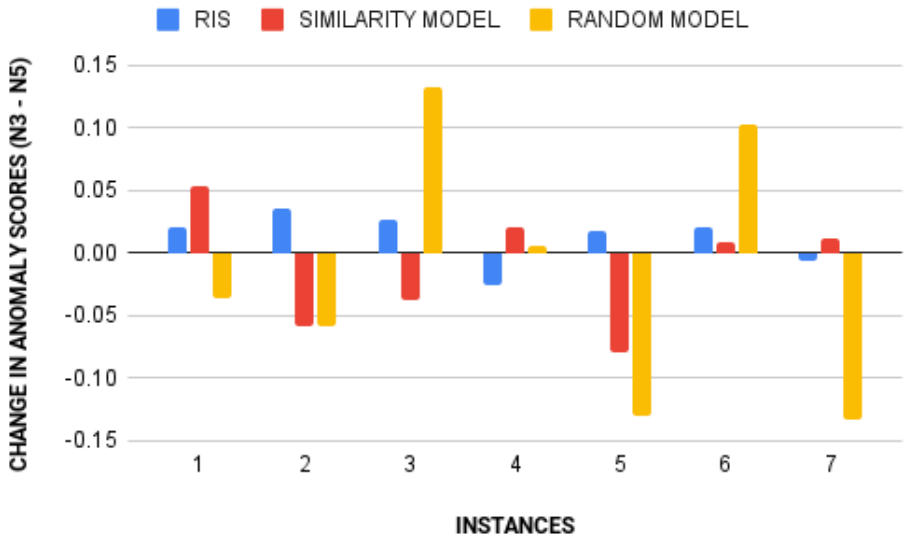

Figure 6.23: Bar plot of the change in scores of anomaly metric for each instance on the supplementary dataset for all models.

On the supplementary dataset, across all experiments performed for $N = 3$ and $N = 5$ on the dataset, RIS model was the best for 13 out of 14, of the 13 there were ties on 4 while it was the clear winner on 9. The performance of RIS on denser datasets (RIS won 69% of experiments) is better when compared to the sparse main dataset where RIS is better for 6 of 12 experiments (RIS won 50% of experiments).

6.1.3 Summary of Deductions from Experiments

In this section, I outline a summary of deductions from the results of the experiment. The deductions are as follows;

- Regardless of the density of the dataset and the size of the extracted set (N), the redundancy and the similarity metrics follow the same patterns. While the similarity-based model had the best scores for the similarity metric, the random model always had the best score for redundancy. RIS model consistently has the second best scores for both metrics for all experiments.

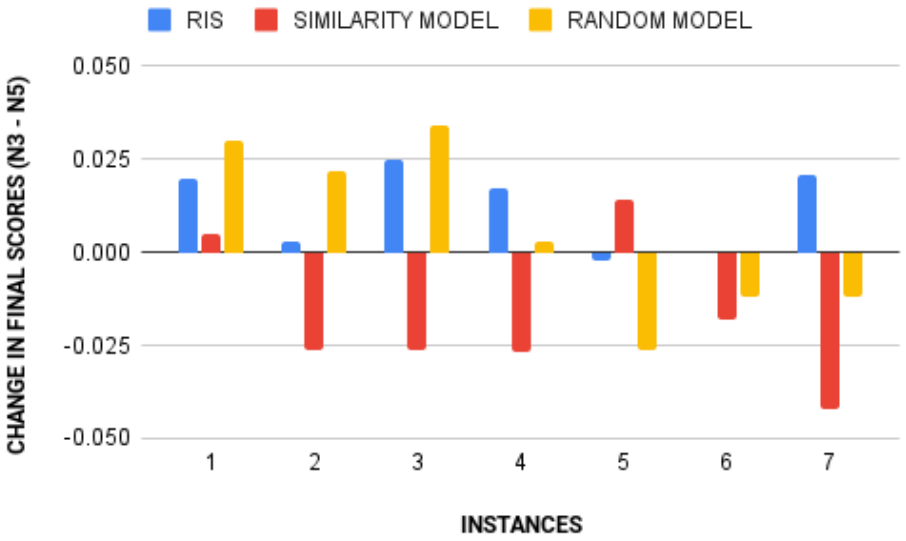

Figure 6.24: Bar plot of the change in scores of final score for each instance on the supplementary dataset for all models.

- The density of the dataset is a major factor for the anomaly metric. The anomaly scores for the models are higher for denser datasets and lower for spare datasets.

- Across all experiments, it is also noticeable that when the similarity score is low for a model, the redundancy score for its extracted set is high and vice-versa.

- RIS is the best model for extracting representative but diverse samples when the value of N is small regardless of the density of the dataset. However, when N is big, it is beneficial to consider the similarity-based models, especially for sparse datasets.

- RIS appears to perform better on dense datasets in comparison to sparse datasets especially when a small N is desired

6.1.4 Challenges

Ideally, a qualitative evaluation is necessary for this task, but some challenges that are discussed in this section made this impossible.

The RIS idea introduced in the research is a novel idea, and hence a robust evaluation proved to be difficult. Also, discussions with medical doctors and extensive research proved that aneurysm location and type are important features for the treatment of IA's. The size of the main dataset posed problems with evaluating this qualitatively because there was not enough samples per IA location or type to facilitate localized extraction of representative and diverse subset for the IA's. This is why only a quantitative evaluation was explored.

The notion of a representative and diverse selection of samples with respect to IA's have to be localized to the aneurysm location and the aneurysm type as these are already established discriminators of IA's (THOMPSON et al. (2015); ZHAO et al. (2018)). After preprocessing of the given dataset as explained in section 4.3.1, there were 70 instances in the dataset. After the inclusion of the features for IA location and type, the instances would have reduced to 57 instances after pre-processing because the other instances have at least one of the features missing. Furthermore, grouping the instances based on location yielded very small amount of samples for each location and this would not be sufficient for this research.

I explored the idea of augmenting this data with synthetic data to facilitate the training of localized models for IA's, but this idea proved insufficient because synthetic data can't be evaluated qualitatively by medical experts. Evaluating synthetic data would be impossible because when synthetic samples are extracted, there will be no images for a medical expert to judge the quality of the extracted samples.

Furthermore, a more robust quantitative evaluation proved challenging because, after thorough research there was no existing work to compare the proposed approach against, there was also no existing dataset that labeled pairs of instances using the notion of representative and diverse in any domain.

6.2 Instance Selection Adaptation

The experiment explained in section 5.2 was done to evaluate the performance of the proposed IS adaptation of RIS discussed in section 4.4.3. Fig-

ure 6.25 shows the performance of the proposed approach in comparison with subsets extracted by random sampling on the iris dataset. The figure shows the accuracy of models trained using subsets extracted by both approaches on the test set. From the figure, the RIS adaptation has better accuracy's using models trained on 12.5% and 30% of the training set, while the random model is better with 50% of the training set. This results aligns with the evaluation for RIS (section 5.1) where we see RIS perform better for smaller sizes of extracted set compared to other models. The accuracy of a model trained using the complete training set had an accuracy of 100%.

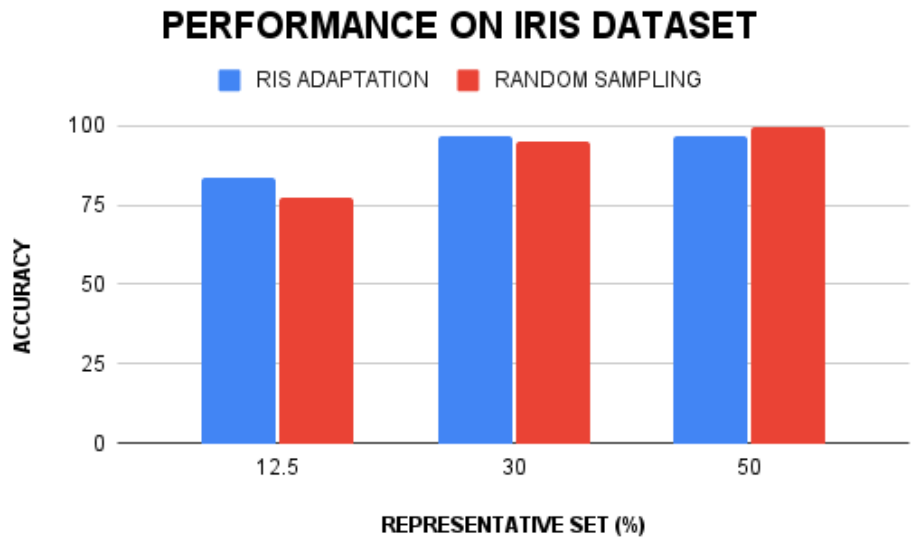

Figure 6.25: Accuracy of models trained with different percentages of representative sets extracted from the training set of the iris dataset

The second experiment to determine the efficacy of the proposed adaptation was performed using the supplementary dataset. The result is shown in figure 6.26. The accuracy of a model trained using the complete training set had an accuracy of 57.75%. Unlike the experiment with the iris dataset, the random sampling extracted sets posted the best accuracy's for all 3 sizes of the representative set. This deviation in pattern may be because of the size of these datasets, while the iris dataset has 150 instances, the supplementary dataset has 351 instances. Therefore, after holding out 20% of the respective datasets for testing, every percentage of representative set

extracted from the training set of the supplementary dataset is 2.34 times bigger than its iris counterpart. This may not be small enough to emphasize the strength of the proposed *RIS* adaptation on the supplementary dataset.

Furthermore, the class labels can also be a reason for this deviation, while the iris dataset has 3 classes, the rupture state feature of the supplementary dataset has 2 classes. The probability of misclassification on the iris dataset is 66% while it is 50% on the supplementary dataset.

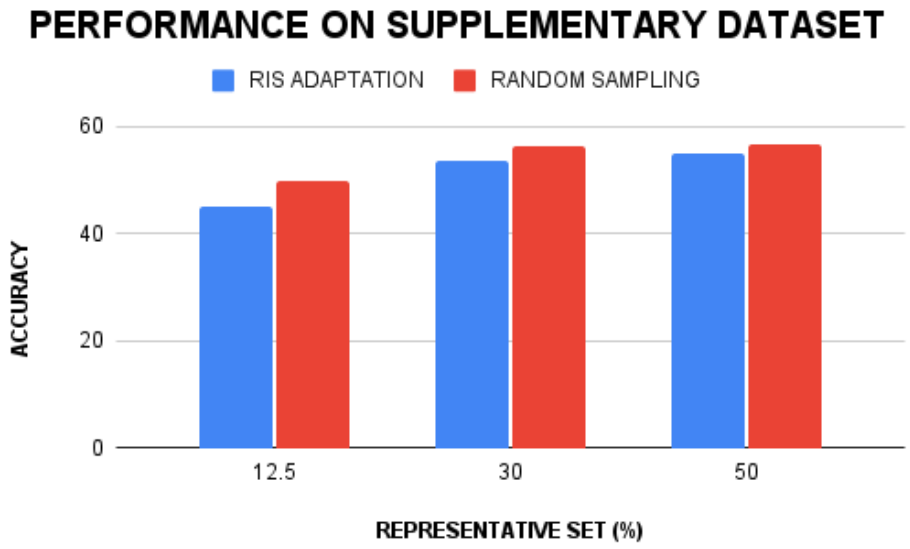

Figure 6.26: Accuracy of models trained with different percentages of representative sets extracted from the training set of the supplementary dataset

6.3 Discussion of Research Questions

6.3.1 RQ 1. How do we define an extraction technique for the task of extracting representative but diverse samples?

The proposed *RIS* technique (explained in detail in section 4.6) in this work fulfills the task of selecting representative but diverse subset. This methodology has mainly three steps, namely;

- Outlier detection and removal

- Clustering
- Prototyping

The outlier removal step is used to drop anomalies from the dataset to ensure they are not in the extracted set, the clustering step groups similar instances together so that the extracted set does not consist of instances that are alike. Finally the prototyping step is the technique with which the extracted subset is generated after the completion of the initial two steps.

6.3.2 RQ 2. How is representative but diverse defined?

The term "representative" denotes similarity, while "diverse" represent difference. Therefore a representative but diverse subset with respect to a given instance of interest shouldn't be the most similar instances in the dataset only because variety in selections is essential. We are not interested in a subset that is most similar to the supplied example, but in a subset that offers variance in line with similarity.

A representative but diverse subset can then be defined as a set of instances which are similar to an instance of interest but different from each other. The novel RIS algorithm was designed to extract this kind of subset from a larger dataset using various unsupervised machine learning techniques and algorithms. The steps used to achieve this are outlined in RQ1.

6.3.3 RQ 3. What metrics will be used to evaluate the extracted set?

A metric suitable for evaluating this task have to fulfill 3 conditions, this metric must:

- measure the degree of similarity between the each of the instance in the extracted set and the instance of interest
- measure the degree of similarity between the instances in the extracted set
- penalise the extracted set if outliers are in the set

The equation proposed in section 4.5.2 satisfies the above conditions. The similarity part Sim, measures the representation of the instances in the extracted set with respect to the instance of interest, the redundancy part Red, measures the degree of similarity within the extracted set and penalises a lack of diversity and the anomaly part $Anom$ checks and penalises the inclusion of outliers in the extracted set.

This equation is a weighted aggregation and these weights can be used to emphasize different parts of the equation depending on the importance of a metric to the user.

6.3.4 RQ 4. How do we distinguish between diverse cases and outliers?

The outlier detection step in the RIS framework is to ensure that outliers are not selected when the algorithm attempts to optimise for diversity. For this reason, I developed the $Anom$ metric (Equation 4.7) which penalises the selection when an outlier is present in the extracted set.

This metric is based on LRD which calculates the distance of a point to its k nearest neighbours. Table 6.3 shows the scores for the outliers dropped by the outlier removal step of the RIS algorithm and the scores of the least scoring instances in the dataset. As can be seen, there is a relationship between DBSCAN predictions and the LRD scores, the lowest scoring 6 points from the total of 70 sufficiently captures the outliers found by DBSCAN. Thus the scores from this metric can sufficiently ensure outliers are not mistaken for diversity.

Table 6.3: LRD Scores and DBSCAN Outliers

Scores	**DBSCAN Evaluation**
0.000000	Outlier
0.028132	Inlier
0.028295	Inlier
0.030359	Outlier
0.030802	Outlier
0.036230	Outlier

Although LRD can sufficiently handle outliers, the spread of it's scores for inliers can also adversely affect the scoring of the aggregated metric. Since

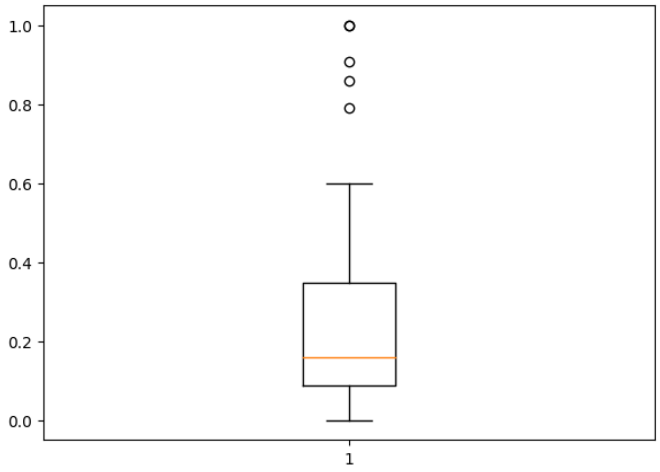

Figure 6.27: Distribution of LRD score for the dataset

data points can either be inliers or outliers, there no justification for a huge disparity in scores between inliers. Figure 6.27 is a boxplot of the LRD scores for the points in the dataset with a mean of approximately 0.18. When points with a higher value (for example 0.8) for LRD are selected, this can have an unfair impact on the aggregated scores despite being an inlier like a point with a value of 0.1.

7

Conclusions and Future Work

7.1 Conclusions

In this thesis, I developed a novel framework for extracting a representative but diverse subset from a larger database with respect to an AOI using a dataset of IA's. I also developed metrics used to evaluate the quality of an extracted set to ascertain the degree to which the subset fits the AOI.

The framework, which I have called RIS, consists of three major steps; the outlier detection and removal step, the clustering step and the prototyping step. For each of these steps, experiments were conducted using several unsupervised machine learning algorithms. For the outlier removal, Autoencoder, Isolation Forest, DBSCAN and Local Outlier Factor were tested. Likewise, K-means, DBSCAN, and OPTICS were tested for clustering. Upon analysis of the suitability of these algorithms to the given IA dataset, DBSCAN, K-means, and an iterative similarity-based prototyping approach was used for each of these steps respectively.

I also developed metrics for evaluating an extracted subset by exploiting existing research in IS and adapting these equations to suit our task. The three developed metrics are Similarity (Sim) which measures representativeness, Redundancy (Red) which measures diversity, and Anomaly ($Anom$) which penalizes the inclusion of outliers in the extracted set. These individual metrics were aggregated to form a final equation, which measures the degree of representativeness and diversity of an extracted subset with respect to a given instance. This equation can be weighted to emphasize the importance of a particular metric depending on the user, the use case, the domain of the data, etc.

Our experiments showed that for the given dataset, when the size of the extracted set is equal to the number of clusters, RIS performs better than a similarity based model which optimizes for representativeness and a random selection approach which optimizes for diversity. However when the size of the of the extracted subset was increased, the performance of RIS began to decline. I also saw that while the similarity based model gets high scores for the similarity, its redundancy scores were low. The random selection approach on the other hand has high redundancy scores and low scores for similarity. RIS tries to optimize the weaknesses of both models by getting fair scores for both similarity and diversity unlike the other two models. This work also showed that the proposed metrics perform the task of measuring the quality of extracted sets effectively.

Both the RIS framework and the proposed equations are reusable and can be adapted to other datasets in any domain, as long as adequate analysis to determine the appropriate choice of unsupervised algorithm to be used in each step of the framework, their hyper-parameters and proper weights are assigned to each metric in the final equation.

Finally, the results of this study indicate that the proposed RIS framework can be useful for the task of extracting representative but diverse IA's with respect to to a given AOI

7.2 Limitations

Like most scientific studies, this research has some limitations, this is what I discuss in this section.

The minimum size of an extracted set that the proposed RIS algorithm for extraction of representative but diverse samples can handle is equal to the number of clusters in the dataset. This is not a problem for datasets with small clusters (5 or less), which were used for the analysis, but for large datasets with many clusters, the extracted set can contain too many instances that may not be necessary for a user depending on this task or domain. A solution for this can be to extract samples from k clusters closest to the point where the AOI lies, thereby restricting the algorithm to extract instances less than the number of clusters in the dataset.

Furthermore, PCA was used extensively in this research, this is not ideal for datasets with categorical features. Although, I used a categorical encoding step to navigate this problem and also tried FAMD before settling on PCA as discussed in section 4.3.1, it is not optimal to use encoding approaches for PCA. It is important to use other techniques of dimensionality reduction to handle mixed datasets.

Although the flexible weighting of the each metric in the final score is an advantage which offers adaptation to different use cases, datasets and domains, it can also be a disadvantage because this research did not propose a unified optimal weights for the metrics. To sacrifice unification in favor of flexibility means that the formula does not offer a a standard with which can be used to evaluate every potential nuances involved in working with a different domain or dataset.

Finally, for a medical research, a qualitative evaluation is important regardless of how good the quantitative evaluation is, it is pivotal that medical experts examine the results of the work. Unfortunately, this was impossible for reasons stated in section 6.1.4. Evaluating this approach using a larger dataset of IA's with sufficient instances that permit for the building localized models for extracting representative but diverse IA's is necessary to facilitate a qualitative evaluation.

7.3 Future Work

The RIS model introduced in this work is focused on extracting a subset of representative but diverse samples from a larger dataset. This approach presented here was tested on tabular data, however, the robustness of this idea could be improved by testing other types of datasets, such as text and image data. This would make for a more robust evaluation of the framework if expanded to also cover various datasets.

In addition, all the metrics used for evaluating an extracted set were all distance based, it may be beneficial to evaluate the quality of an extracted set by non-distance based metrics. Despite the fact that I exploited methodologies used for instance selection as discussed in chapter 3, a new set of metrics for evaluation also eliminates any form of bias because this proposed RIS framework and the evaluation metrics were developed by me.

The size and the composition of the dataset can also be improved. A dataset with a considerable percentage of outliers would be beneficial to adequately stress test the anomaly metric $Anom_k(S)$. Furthermore, the development of a dataset that matches groups on instances based on the notion of representative but diverse would be useful as a baseline to evaluate the performance of this model and other models that attempt to solve this problem.

A larger IA dataset with more samples for each aneurysm location and type would be adequate to build a localized model for IA's that can be effectively evaluated both qualitatively and quantitatively.

Furthermore, exploring other approaches to this task, such as developing an optimization function which we try to optimize by an iterative selection of samples until the size of the extracted set is reached can serve as a useful and an adequate alternative to RIS.

Overall, testing the approach on different types of data, development of non-distance based metrics, use of a larger dataset with more samples, and exploring other approaches would provide a more comprehensive understanding and analysis of this task.

A

List of Figures

2.1 The circle of willis and surrounding arteries(FLANAGAN et al. (2015)) 6

2.2 Endovascular and surgical treatments for IA. (A) Endovascular coiling of the aneurysm sac. (B) Surgical clipping of the aneurysm neck. (C) Endovascular treatment using coils and a stent. (D) Endovascular treatment using flow diverter. (PERRONE et al. (2015)) 8

3.1 Prototype selection strategy. (GARCÍA et al. (2015)) 23

3.2 Training set selection strategy. (GARCÍA et al. (2015)) 23

3.3 Experiment results from KIM et al. (2017). In the rightmost column, the Recall, Precision, and F1 outcomes of LDC instance selection are displayed, together with the total counts of true positives (TP) and the number and percentage of true positive increases (in comparison to supervised learning). In the Recall, Precision, and F1 columns, the numbers in parentheses reflect the difference between the supervised classifier and the LDC technique. Asterisks (*) indicate results that are significantly different from supervised learning at the 95% confidence level. 25

3.4 Experiment results from PAN et al. (2005). (a) shows the scores of the coverage metric while (b) shows the score for clustering accuracy metric. R represents the number of instances in the extracted set 26

4.1 Low-level overview of the proposed RIS framework 30

4.2 Image of multiple aneurysms from a patient 32

4.3 K-means clustering of data augmented with 500 data points . 34

4.4 DBSCAN clustering of data augmented with 500 data points . 34

4.5 DBSCAN clustering of data after FAMD dimensionality reduction 38

4.6 Steps for instance extraction with strategies explored for each step. Strategies highlighted in green were used after experiments. 40

4.7 Outlier detection scatter-plots showing assignment of outliers by autoencoders (top), isolation forest (second), LOF (third), DBSCAN (bottom). X and Y axis of the scatter-plots are the first 2 principal components used only for visualisation while the assignments was done on the 3 principal components datasets 42

4.8 Selecting number of clusters using plots of silhouette coefficient against number of clusters (left) and SSE against number of clusters (right) 43

4.9 Cluster assignments: DBSCAN assignments (top) with 2 clusters and 4 points as outliers (0: 39, 1: 27, -1: 4), OPTICS assignments (middle) with 4 clusters and 43 points as outliers (-1: 43, 3: 8, 0: 8, 1: 7, 2: 4), Kmeans assignments (bottom) with 3 clusters (1: 30, 0: 20, 2: 20) 44

4.10 K-means clustering of data with 3 components of PCA visualised using the first 2 PC's as the X and Y axis respectively . . 45

4.11 K-means clustering of full dataset without PCA visualised using the first 2 PC's as the X and Y axis respectively 45

6.1 Line plot of the performance of each model on the selected AOI instances on the main dataset for $N = 3$ for similarity metric. 60

6.2 Line plot of the performance of each model on the selected AOI instances on the main dataset for $N = 3$ for redundancy metric. 61

6.3 Line plot of the performance of each model on the selected AOI instances on the main dataset for $N = 3$ for anomaly metric. 62

6.4 Line plot of the performance of each model on the selected AOI instances on the main dataset for $N = 3$ for final score. . . 63

6.5 Line plot of the performance of each model on the selected AOI instances on the main dataset for $N = 5$ for similarity metric. 63

6.6 Line plot of the performance of each model on the selected AOI instances on the main dataset for $N = 5$ for redundancy metric. 64

6.7 Line plot of the performance of each model on the selected AOI instances on the main dataset for $N = 5$ for anomaly metric. 64

6.8 Line plot of the performance of each model on the selected AOI instances on the main dataset for $N = 5$ for final score. . . 65

6.9 Bar plot of the change in scores of similarity metric for each instance on the main dataset for all models. 67

6.10 Bar plot of the change in scores of redundancy metric for each instance on the main dataset for all models. 67

6.11 Bar plot of the change in scores of anomaly metric for each instance on the main dataset for all models. 68

6.12 Bar plot of the change in scores of final for each instance score on the main dataset for all models. 69

6.13 Line plot of the performance of each model on the selected AOI instances on the supplementary dataset for $N = 3$ for similarity metric. 71

6.14 Line plot of the performance of each model on the selected AOI instances on the supplementary dataset for $N = 3$ for redundancy metric. 71

6.15 Line plot of the performance of each model on the selected AOI instances on the supplementary dataset for $N = 3$ for anomaly metric. 72

6.16 Line plot of the performance of each model on the selected AOI instances on the supplementary dataset for $N = 3$ for final score. 73

6.17 Line plot of the performance of each model on the selected AOI instances on the supplementary dataset for $N = 5$ for similarity metric. 74

6.18 Line plot of the performance of each model on the selected AOI instances on the supplementary dataset for $N = 5$ for redundancy metric. 74

6.19 Line plot of the performance of each model on the selected AOI instances on the supplementary dataset for $N = 5$ for anomaly metric. 75

6.20 Line plot of the performance of each model on the supplementary dataset for $N = 5$ for final score. 75

6.21 Bar plot of the change in scores of similarity metric for each instance on the supplementary dataset for all models. 76

6.22 Bar plot of the change in scores of redundancy metric for each instance on the supplementary dataset for all models. . 77

6.23 Bar plot of the change in scores of anomaly metric for each instance on the supplementary dataset for all models. 78

6.24 Bar plot of the change in scores of final score for each instance on the supplementary dataset for all models. 79

6.25 Accuracy of models trained with different percentages of representative sets extracted from the training set of the iris dataset 81

6.26 Accuracy of models trained with different percentages of representative sets extracted from the training set of the supplementary dataset 82

6.27 Distribution of LRD score for the dataset 85

B

List of Tables

1.1 Absolute Differences between AOI and Database of IA 3

4.1 IS Vs RIS 30

4.2 Software used for Development and Experiment 31

4.3 Co-operating Medical Facilities for Main Dataset 31

4.4 IA feature type and the number present in dataset 31

4.5 Description of the Features of the Main Dataset After Pre-processing 35

4.6 Explained Variance of Principal Components 37

4.7 Description of the Features of the Supplementary Dataset After Pre-processing 38

5.1 Evaluation Results $N = 3$ 52

5.2 Evaluation Results $N = 5$ 53

5.3 Evaluation Results Supplementary Dataset $N = 3$ 54

5.4 Evaluation Results Supplementary Dataset $N = 5$ 55

5.5 Evaluation results of proposed IS adaptation on iris dataset. The table shows the accuracy on the test data of the models trained using different percentages of representative set extracted from the training set. 57

5.6 Evaluation results of proposed IS adaptation on supplementary dataset. The table shows the accuracy on the test data of the models trained using different percentages of representative set extracted from the training set. 57

6.1 Experiment Summary on Main Dataset 60

6.2 Experiment Summary on Main Dataset 70

6.3 LRD Scores and DBSCAN Outliers 84

C

Bibliography

[ABBOUD et al. 2017] T. Abboud, J. Rustom, M. Bester, P. Czorlich, E. Vittorazzi, H. O. Pinnschmidt, M. Westphal und J. Regelsberger. **Morphology of ruptured and unruptured intracranial aneurysms**. World neurosurgery, Vol. 99:610–617, 2017.

[AGHA und FOWLER 2015] R. A. Agha und A. J. Fowler. **The role and validity of surgical simulation**. International surgery, Vol. 100(2):350–357, 2015.

[AHN et al. 2017] J.-M. Ahn, J.-S. Oh, S.-M. Yoon, J.-H. Shim, H.-J. Oh und H.-G. Bae. **Procedure-related complications during endovascular treatment of intracranial saccular aneurysms**. Journal of cerebrovascular and endovascular neurosurgery, Vol. 19(3):162–170, 2017.

[AHUJA et al. 2019] R. Ahuja, A. Solanki und A. Nayyar. **Movie recommender system using k-means clustering and k-nearest neighbor**. In: 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), pp. 263–268. 2019, IEEE.

[AJMAL et al. 2023] S. Ajmal, R. A. R. Ashfaq und K. Saleem. **Uncertainty Based Optimal Sample Selection for Big Data**. IEEE Access, Vol. 11:6284–6292, 2023.

[ALLGAIER et al. 2022] M. Allgaier, B. Neyazi, I. Erol Sandalcioglu, B. Preim und S. Saalfeld. **Immersive VR training system for clipping intracranial aneurysms**. Current Directions in Biomedical Engineering, Vol. 8(1):9–12, 2022.

[ALWALID et al. 2022] O. Alwalid, X. Long, M. Xie und P. Han. **Artificial Intelligence Applications in Intracranial Aneurysm: Achievements,**

Challenges and Opportunities. Academic Radiology, Vol. 29:S201–S214, 2022. Special Issue on Neuroradiology.

[AN et al. 2022] X. An, J. He, Y. Di, M. Wang, B. Luo, Y. Huang und D. Ming. **Intracranial aneurysm rupture risk estimation with multi-dimensional feature fusion**. Frontiers in Neuroscience, Vol. 16:813056, 2022.

[ANKERST et al. 1999] M. Ankerst, M. M. Breunig, H.-P. Kriegel und J. Sander. **OPTICS: Ordering points to identify the clustering structure**. ACM Sigmod record, Vol. 28(2):49–60, 1999.

[BACKES et al. 2015] D. Backes, M. D. Vergouwen, A. T. Tiel Groenestege, A. S. E. Bor, B. K. Velthuis, J. P. Greving, A. Algra, M. J. Wermer, M. A. van Walderveen, K. G. terBrugge et al. **PHASES score for prediction of intracranial aneurysm growth**. Stroke, Vol. 46(5):1221–1226, 2015.

[BELAVADI et al. 2021] R. Belavadi, S. V. R. Gudigopuram, C. C. Raguthu, H. Gajjela, I. Kela, C. L. Kakarala, M. Hassan und I. Sange. **Surgical Clipping Versus Endovascular Coiling in the Management of Intracranial Aneurysms**. Cureus, Vol. 13(12), 2021.

[BOULOUIS et al. 2017] G. Boulouis, C. Rodriguez-Régent, E. Rasolonjatovo, W. B. Hassen, D. Trystram, M. Edjlali-Goujon, J.-F. Meder, C. Oppenheim und O. Naggara. **Unruptured intracranial aneurysms: An updated review of current concepts for risk factors, detection and management**. Revue neurologique, Vol. 173(9):542–551, 2017.

[BREUNIG et al. 2000] M. M. Breunig, H.-P. Kriegel, R. T. Ng und J. Sander. **LOF: identifying density-based local outliers**. In: Proceedings of the 2000 ACM SIGMOD international conference on Management of data, 2000, pp. 93–104.

[BRISMAN et al. 2005] J. L. Brisman, Y. Niimi, J. K. Song und A. Berenstein. **Aneurysmal rupture during coiling: low incidence and good outcomes at a single large volume center**. Neurosurgery, Vol. 57(6):1103–1109, 2005.

[BROWN und BRODERICK 2014] R. D. Brown und J. P. Broderick. **Unruptured intracranial aneurysms: epidemiology, natural history, man-**

agement options, and familial screening. The Lancet Neurology, Vol. 13(4):393–404, 2014.

[BURKART und HUBER 2021] N. Burkart und M. F. Huber. **A survey on the explainability of supervised machine learning**. Journal of Artificial Intelligence Research, Vol. 70:245–317, 2021.

[CANO et al. 2003] J. Cano, F. Herrera und M. Lozano. **Using evolutionary algorithms as instance selection for data reduction in KDD: an experimental study**. IEEE Transactions on Evolutionary Computation, Vol. 7(6):561–575, 2003.

[CEBRAL et al. 2011] J. R. Cebral, F. Mut, J. Weir und C. Putman. **Quantitative characterization of the hemodynamic environment in ruptured and unruptured brain aneurysms**. American Journal of Neuroradiology, Vol. 32(1):145–151, 2011.

[CHEN et al. 2018] G. Chen, C. Wang, M. Zhang, Q. Wei und B. Ma. **How "small" reflects "large"?—Representative information measurement and extraction**. Information Sciences, Vol. 460:519–540, 2018.

[CZARNOWSKI und JĘDRZEJOWICZ 2018] I. Czarnowski und P. Jędrzejowicz. **Cluster-Based Instance Selection for the Imbalanced Data Classification**. In: N. T. Nguyen, E. Pimenidis, Z. Khan und B. Trawiński, Eds., Computational Collective Intelligence, pp. 191–200. 2018, Springer International Publishing, Cham.

[DEHARIYA et al. 2010] V. K. Dehariya, S. K. Shrivastava und R. Jain. **Clustering of image data set using k-means and fuzzy k-means algorithms**. In: 2010 International conference on computational intelligence and communication networks, pp. 386–391. 2010, IEEE.

[DETMER et al. 2019] F. J. Detmer, S. Hadad, B. J. Chung, F. Mut, M. Slawski, N. Juchler, V. Kurtcuoglu, S. Hirsch, P. Bijlenga, Y. Uchiyama et al. **Extending statistical learning for aneurysm rupture assessment to Finnish and Japanese populations using morphology, hemodynamics, and patient characteristics**. Neurosurgical focus, Vol. 47(1):E16, 2019.

[DHAR et al. 2008] S. Dhar, M. Tremmel, J. Mocco, M. Kim, J. Yamamoto, A. H. Siddiqui, L. N. Hopkins und H. Meng. **Morphology parameters**

for intracranial aneurysm rupture risk assessment. Neurosurgery, Vol. 63(2):185, 2008.

[DUBES und JAIN 1980] R. Dubes und A. K. Jain. **Clustering methodologies in exploratory data analysis**. Advances in computers, Vol. 19:113–228, 1980.

[ETMINAN et al. 2019] N. Etminan, H.-S. Chang, K. Hackenberg, N. K. De Rooij, M. D. Vergouwen, G. J. Rinkel und A. Algra. **Worldwide incidence of aneurysmal subarachnoid hemorrhage according to region, time period, blood pressure, and smoking prevalence in the population: a systematic review and meta-analysis**. JAMA neurology, Vol. 76(5):588–597, 2019.

[FLANAGAN et al. 2015] M. F. Flanagan et al. **The role of the craniocervical junction in craniospinal hydrodynamics and neurodegenerative conditions**. Neurology Research International, Vol. 2015, 2015.

[FRÉNEAU et al. 2022] M. Fréneau, C. Baron-Menguy, A.-C. Vion und G. Loirand. **Why are women predisposed to intracranial aneurysm?** Frontiers in cardiovascular medicine, Vol. 9:815668, 2022.

[GARCÍA et al. 2015] S. García, J. Luengo und F. Herrera. **Data preprocessing in data mining**. Springer, 2015.

[GÖLITZ et al. 2014] P. Gölitz, T. Struffert, O. Ganslandt, S. Lang, F. Knossalla und A. Doerfler. **Contrast-enhanced angiographic computed tomography for detection of aneurysm remnants after clipping: a comparison with digital subtraction angiography in 112 clipped aneurysms**. Neurosurgery, Vol. 74(6):606–614, 2014.

[GOODFELLOW et al. 2016] I. Goodfellow, Y. Bengio und A. Courville. **Deep learning**. MIT press, 2016.

[GORDON 2000] J. A. GORDON. **The human patient simulator™: acceptance and efficacy as a teaching tool for students**. Academic Medicine, Vol. 75(5):522, 2000.

[HAHNE et al. 2008] F. Hahne, W. Huber, R. Gentleman, S. Falcon, R. Gentleman und V. Carey. **Unsupervised machine learning**. Bioconductor case studies, pp. 137–157, 2008.

[HAWKINS 1980] D. M. Hawkins. **Identification of outliers**, Vol. 11. Springer, 1980.

[HOCHBAUM und PATHRIA 1998] D. S. Hochbaum und A. Pathria. **Analysis of the greedy approach in problems of maximum k-coverage**. Naval Research Logistics (NRL), Vol. 45(6):615–627, 1998.

[HUANG et al. 2018] M.-W. Huang, W.-C. Lin und C.-F. Tsai. **Outlier removal in model-based missing value imputation for medical datasets**. Journal of healthcare engineering, Vol. 2018, 2018.

[HUANG et al. 2021] M.-W. Huang, C.-F. Tsai und W.-C. Lin. **Instance selection in medical datasets: a divide-and-conquer framework**. Computers & Electrical Engineering, Vol. 90:106957, 2021.

[IHN et al. 2018] Y. K. Ihn, S. H. Shin, S. K. Baik und I. S. Choi. **Complications of endovascular treatment for intracranial aneurysms: management and prevention**. Interventional Neuroradiology, Vol. 24(3):237–245, 2018.

[IM et al. 2009] S.-H. Im, M. Han, O.-K. Kwon, B. Kwon, S. Kim, J. Kim und C. Oh. **Endovascular coil embolization of 435 small asymptomatic unruptured intracranial aneurysms: procedural morbidity and patient outcome**. American Journal of Neuroradiology, Vol. 30(1):79–84, 2009.

[INVESTIGATORS 2012] U. J. Investigators. **The natural course of unruptured cerebral aneurysms in a Japanese cohort**. New England Journal of Medicine, Vol. 366(26):2474–2482, 2012.

[ISSENBERG und SCALESE 2008] S. B. Issenberg und R. J. Scalese. **Simulation in health care education**. Perspectives in biology and medicine, Vol. 51(1):31–46, 2008.

[JENNETT et al. 1981] B. Jennett, J. Snoek, M. Bond und N. Brooks. **Disability after severe head injury: observations on the use of the Glasgow Outcome Scale.** Journal of Neurology, Neurosurgery & Psychiatry, Vol. 44(4):285–293, 1981.

[JIN und HAN 2010] X. Jin und J. Han. K-Means Clustering, pp. 563–564. 2010. Springer US, Boston, MA.

[JOLLIFFE und CADIMA 2016] I. T. Jolliffe und J. Cadima. **Principal component analysis: a review and recent developments**. Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences, Vol. 374(2065):20150202, 2016.

[JUVELA et al. 2013] S. Juvela, K. Poussa, H. Lehto und M. Porras. **Natural history of unruptured intracranial aneurysms: a long-term follow-up study**. Stroke, Vol. 44(9):2414–2421, 2013.

[KAELBLING et al. 1996] L. P. Kaelbling, M. L. Littman und A. W. Moore. **Reinforcement learning: A survey**. Journal of artificial intelligence research, Vol. 4:237–285, 1996.

[KEEDY 2006] A. Keedy. **An overview of intracranial aneurysms**. McGill Journal of Medicine: MJM, Vol. 9(2):141, 2006.

[KIM 2006] K.-j. Kim. **Artificial neural networks with evolutionary instance selection for financial forecasting**. Expert Systems with Applications, Vol. 30(3):519–526, 2006.

[KIM et al. 2017] Y. Kim, E. Riloff und S. M. Meystre. **Exploiting unlabeled texts with clustering-based instance selection for medical relation classification**. In: AMIA annual symposium proceedings, Vol. 2017, p. 1060. 2017, American Medical Informatics Association.

[KOCKRO et al. 2007] R. A. Kockro, A. Stadie, E. Schwandt, R. Reisch, C. Charalampaki, I. Ng, T. T. Yeo, P. Hwang, L. Serra und A. Perneczky. **A collaborative virtual reality environment for neurosurgical planning and training**. Operative Neurosurgery, Vol. 61(suppl_5):ONSE379–ONSE391, 2007.

[LEE et al. 2021] K. Lee, D. Ippolito, A. Nystrom, C. Zhang, D. Eck, C. Callison-Burch und N. Carlini. **Deduplicating training data makes language models better**. arXiv preprint arXiv:2107.06499, 2021.

[LI et al. 2022] K. Li, X. Gao, X. Jia, B. Xue, S. Fu, Z. Liu, X. Huang und Z. Huang. **Detection of local and clustered outliers based on the density–distance decision graph**. Engineering Applications of Artificial Intelligence, Vol. 110:104719, 2022.

[LIN et al. 2017] W.-C. Lin, C.-F. Tsai, Y.-H. Hu und J.-S. Jhang. **Clustering-based undersampling in class-imbalanced data**. Information Sciences, Vol. 409:17–26, 2017.

[LIN et al. 2015] W.-C. Lin, C.-F. Tsai, S.-W. Ke, C.-W. Hung und W. Eberle. **Learning to detect representative data for large scale instance selection**. Journal of Systems and Software, Vol. 106:1–8, 2015.

[LINDGREN et al. 2018] A. Lindgren, M. D. Vergouwen, I. van der Schaaf, A. Algra, M. Wermer, M. J. Clarke und G. J. Rinkel. **Endovascular coiling versus neurosurgical clipping for people with aneurysmal subarachnoid haemorrhage**. Cochrane Database of Systematic Reviews, (8), 2018.

[LIU et al. 2012] F. T. Liu, K. M. Ting und Z.-H. Zhou. **Isolation-based anomaly detection**. ACM Transactions on Knowledge Discovery from Data (TKDD), Vol. 6(1):1–39, 2012.

[LIU et al. 2022] J. Liu, X. Zou, Y. Zhao, Z. Jin, J. Tu, X. Ning, J. Li, X. Yang und J. Wang. **Prevalence and Risk Factors for Unruptured Intracranial Aneurysms in the Population at High Risk for Aneurysm in the Rural Areas of Tianjin**. Frontiers in Neurology, Vol. 13, 2022.

[MA und WEI 2012] B. Ma und Q. Wei. **Measuring the coverage and redundancy of information search services on e-commerce platforms**. Electronic Commerce Research and Applications, Vol. 11(6):560–569, 2012.

[MA et al. 2017] B. Ma, Q. Wei, G. Chen, J. Zhang und X. Guo. **Content and Structure Coverage: Extracting a Diverse Information Subset**. INFORMS J. Comput., Vol. 29:660–675, 2017.

[MACQUEEN 1967] J. MacQueen. **Classification and analysis of multivariate observations**. In: 5th Berkeley Symp. Math. Statist. Probability, pp. 281–297. 1967, University of California Los Angeles LA USA.

[MADHULATHA 2012] T. S. Madhulatha. **An overview on clustering methods**. arXiv preprint arXiv:1205.1117, 2012.

[MALHAT et al. 2020] M. Malhat, M. El Menshawy, H. Mousa und A. El Sisi. **A new approach for instance selection: Algorithms, evaluation, and comparisons**. Expert Systems with Applications, Vol. 149:113297, 2020.

[MAUPU et al. 2022] C. Maupu, H. Lebas und Y. Boulaftali. **Imaging modalities for intracranial aneurysm: more than meets the eye**. Frontiers in Cardiovascular Medicine, Vol. 9:793072, 2022.

[MCGUIRE et al. 2021] L. S. McGuire, A. Fuentes und A. Alaraj. **Three-dimensional modeling in training, simulation, and surgical planning in open vascular and endovascular neurosurgery: a systematic review of the literature**. World Neurosurgery, Vol. 154:53–63, 2021.

[MCLAUGHLIN und BOJANOWSKI 2004] N. Mclaughlin und M. W. Bojanowski. **Early surgery-related complications after aneurysm clip placement: an analysis of causes and patient outcomes**. Journal of neurosurgery, Vol. 101(4):600–606, 2004.

[MITCHELL et al. 2007] T. M. Mitchell et al. **Machine learning**, Vol. 1. McGraw-hill New York, 2007.

[MORITA et al. 2005] A. Morita, S. Fujiwara, K. Hashi, H. Ohtsu und T. Kirino. **Risk of rupture associated with intact cerebral aneurysms in the Japanese population: a systematic review of the literature from Japan**. Journal of neurosurgery, Vol. 102(4):601–606, 2005.

[NAGGARA et al. 2011] O. Naggara, J. Raymond, F. Guilbert und D. Altman. **The problem of subgroup analyses: an example from a trial on ruptured intracranial aneurysms**. American journal of neuroradiology, Vol. 32(4):633–636, 2011.

[NAGGARA et al. 2012] O. N. Naggara, A. Lecler, C. Oppenheim, J.-F. Meder und J. Raymond. **Endovascular treatment of intracranial unruptured aneurysms: a systematic review of the literature on safety with emphasis on subgroup analyses**. Radiology, Vol. 263(3):828–835, 2012.

[NIEMANN et al. 2018] U. Niemann, P. Berg, A. Niemann, O. Beuing, B. Preim, M. Spiliopoulou und S. Saalfeld. **Rupture status classification of intracranial aneurysms using morphological parameters**. In: 2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS), pp. 48–53. 2018, IEEE.

[NIEUWKAMP et al. 2009] D. J. Nieuwkamp, L. E. Setz, A. Algra, F. H. Linn, N. K. de Rooij und G. J. Rinkel. **Changes in case fatality of aneurysmal**

subarachnoid haemorrhage over time, according to age, sex, and region: a meta-analysis. The Lancet Neurology, Vol. 8(7):635–642, 2009.

[OISHI et al. 2012] H. Oishi, M. Yamamoto, T. Shimizu, K. Yoshida und H. Arai. **Endovascular therapy of 500 small asymptomatic unruptured intracranial aneurysms**. American journal of neuroradiology, Vol. 33(5):958–964, 2012.

[OLVERA-LÓPEZ et al. 2010] J. A. Olvera-López, J. A. Carrasco-Ochoa, J. F. Martínez-Trinidad und J. Kittler. **A review of instance selection methods**. Artificial Intelligence Review, Vol. 34:133–143, 2010.

[PAN et al. 2005] F. Pan, W. Wang, A. Tung und J. Yang. **Finding representative set from massive data**. In: Fifth IEEE International Conference on Data Mining (ICDM'05), 2005, pp. 8 pp.–.

[PARK et al. 2005] H.-K. Park, M. Horowitz, C. Jungreis, J. Genevro, C. Koebbe, E. Levy und A. Kassam. **Periprocedural morbidity and mortality associated with endovascular treatment of intracranial aneurysms**. American Journal of Neuroradiology, Vol. 26(3):506–514, 2005.

[PERRONE et al. 2015] R. D. Perrone, A. M. Malek und T. Watnick. **Vascular complications in autosomal dominant polycystic kidney disease**. Nature Reviews Nephrology, Vol. 11(10):589–598, 2015.

[PIEROT et al. 2008] L. Pierot, L. Spelle und F. Vitry. **Immediate clinical outcome of patients harboring unruptured intracranial aneurysms treated by endovascular approach: results of the ATENA study**. Stroke, Vol. 39(9):2497–2504, 2008.

[PU et al. 2020] G. Pu, L. Wang, J. Shen und F. Dong. **A hybrid unsupervised clustering-based anomaly detection method**. Tsinghua Science and Technology, Vol. 26(2):146–153, 2020.

[RAHMAH und SITANGGANG 2016] N. Rahmah und I. S. Sitanggang. **Determination of optimal epsilon (eps) value on dbscan algorithm to clustering data on peatland hotspots in sumatra**. In: IOP conference series: earth and environmental science, Vol. 31, p. 012012. 2016, IoP Publishing.

[REINARTZ 2002] T. Reinartz. **A unifying view on instance selection**. Data Mining and Knowledge Discovery, Vol. 6:191–210, 2002.

[ROKED und REDDY 2020] F. Roked und U. Reddy. **Management of subarachnoid haemorrhage**. Anaesthesia & Intensive Care Medicine, Vol. 21(6):305–311, 2020.

[RYAN et al. 2016] J. R. Ryan, K. K. Almefty, P. Nakaji und D. H. Frakes. **Cerebral aneurysm clipping surgery simulation using patient-specific 3D printing and silicone casting**. World neurosurgery, Vol. 88:175–181, 2016.

[RYTTLEFORS et al. 2008] M. Ryttlefors, P. Enblad, R. S. Kerr und A. J. Molyneux. **International subarachnoid aneurysm trial of neurosurgical clipping versus endovascular coiling: subgroup analysis of 278 elderly patients**. Stroke, Vol. 39(10):2720–2726, 2008.

[SAALFELD et al. 2018] S. Saalfeld, P. Berg, A. Niemann, M. Luz, B. Preim und O. Beuing. **Semiautomatic neck curve reconstruction for intracranial aneurysm rupture risk assessment based on morphological parameters**. International journal of computer assisted radiology and surgery, Vol. 13:1781–1793, 2018.

[SANDER et al. 1998] J. Sander, M. Ester, H.-P. Kriegel und X. Xu. **Density-based clustering in spatial databases: The algorithm gdbscan and its applications**. Data mining and knowledge discovery, Vol. 2:169–194, 1998.

[SCHUBERT et al. 2017] E. Schubert, J. Sander, M. Ester, H. P. Kriegel und X. Xu. **DBSCAN revisited, revisited: why and how you should (still) use DBSCAN**. ACM Transactions on Database Systems (TODS), Vol. 42(3):1–21, 2017.

[SEGARAN 2007] T. Segaran. **Collective intelligence-building smart web 2.0 applications**. Newton: O'Reilly, 2007.

[SEIL et al. 2022] R. Seil, C. Hoeltgen, H. Thomazeau, H. Anetzberger und R. Becker. **Surgical simulation training should become a mandatory part of orthopaedic education**. Journal of Experimental Orthopaedics, Vol. 9(1):1–5, 2022.

[SHARMA und KUMAR 2016] S. K. Sharma und S. Kumar. **Comparative analysis of Manhattan and Euclidean distance metrics using A* algorithm**. J. Res. Eng. Appl. Sci, Vol. 1(4):196–198, 2016.

[SONG et al. 2017] Y. Song, J. Liang, J. Lu und X. Zhao. **An efficient instance selection algorithm for k nearest neighbor regression**. Neurocomputing, Vol. 251:26–34, 2017.

[SONOBE et al. 2010] M. Sonobe, T. Yamazaki, M. Yonekura und H. Kikuchi. **Small unruptured intracranial aneurysm verification study: SUAVe study, Japan**. Stroke, Vol. 41(9):1969–1977, 2010.

[SPITZ et al. 2020] L. Spitz, U. Niemann, O. Beuing, B. Neyazi, I. E. Sandalcioglu, B. Preim und S. Saalfeld. **Combining visual analytics and case-based reasoning for rupture risk assessment of intracranial aneurysms**. International Journal of Computer Assisted Radiology and Surgery, Vol. 15:1525–1535, 2020.

[SPITZ et al. 2021] L. Spitz, V. M. Swiatek, B. Neyazi, I. E. Sandalcioglu, B. Preim und S. Saalfeld. **An interactive tool for identifying patient subgroups based on arbitrary characteristics for medical research**. Current Directions in Biomedical Engineering, Vol. 7(1):43–46, 2021.

[STIENEN et al. 2018] M. N. Stienen, M. Germans, J.-K. Burkhardt, M. C. Neidert, C. Fung, D. Bervini, D. Zumofen, M. Roethlisberger, S. Marbacher, R. Maduri et al. **Predictors of in-hospital death after aneurysmal subarachnoid hemorrhage: analysis of a nationwide database (Swiss SOS [Swiss Study on Aneurysmal Subarachnoid Hemorrhage])**. Stroke, Vol. 49(2):333–340, 2018.

[TAN et al. 2018] P.-N. Tan, M. Steinbach, A. Karpatne und V. Kumar. **Introduction to Data Mining (2nd Edition)**. Pearson, 2nd Edn., 2018.

[TANG et al. 2022] X. Tang, L. Zhou, L. Wen, Q. Wu, X. Leng, J. Xiang und X. Zhang. **Morphological and hemodynamic characteristics associated with the rupture of multiple intracranial aneurysms**. Frontiers in Neurology, Vol. 12:2564, 2022.

[TANIOKA et al. 2020] S. Tanioka, F. Ishida, A. Yamamoto, S. Shimizu, H. Sakaida, M. Toyoda, N. Kashiwagi und H. Suzuki. **Machine learning classification of cerebral aneurysm rupture status with morphologic**

variables and hemodynamic parameters. Radiology: Artificial Intelligence, Vol. 2(1):e190077, 2020.

[THAKER et al. 2012] N. G. Thaker, J. D. Turner, W. S. Cobb, I. Hussain, N. Janjua, W. He, C. D. Gandhi und C. J. Prestigiacomo. **Computed tomographic angiography versus digital subtraction angiography for the postoperative detection of residual aneurysms: a single-institution series and meta-analysis**. Journal of neurointerventional surgery, Vol. 4(3):219–225, 2012.

[THOMPSON et al. 2015] B. G. Thompson, R. D. Brown Jr, S. Amin-Hanjani, J. P. Broderick, K. M. Cockroft, E. S. Connolly Jr, G. R. Duckwiler, C. C. Harris, V. J. Howard, S. C. Johnston et al. **Guidelines for the management of patients with unruptured intracranial aneurysms: a guideline for healthcare professionals from the American Heart Association/American Stroke Association**. Stroke, Vol. 46(8):2368–2400, 2015.

[TOTH und CEREJO 2018] G. Toth und R. Cerejo. **Intracranial aneurysms: Review of current science and management**. Vascular Medicine, Vol. 23(3):276–288, 2018.

[TSAI et al. 2013] C.-F. Tsai, W. Eberle und C.-Y. Chu. **Genetic algorithms in feature and instance selection**. Knowledge-Based Systems, Vol. 39:240–247, 2013.

[UPTON 1992] G. J. Upton. **Fisher's exact test**. Journal of the Royal Statistical Society: Series A (Statistics in Society), Vol. 155(3):395–402, 1992.

[VAN ENGELEN und HOOS 2020] J. E. Van Engelen und H. H. Hoos. **A survey on semi-supervised learning**. Machine learning, Vol. 109(2):373–440, 2020.

[VERNOOIJ et al. 2007] M. W. Vernooij, M. A. Ikram, H. L. Tanghe, A. J. Vincent, A. Hofman, G. P. Krestin, W. J. Niessen, M. M. Breteler und A. van der Lugt. **Incidental findings on brain MRI in the general population**. New England Journal of Medicine, Vol. 357(18):1821–1828, 2007.

[WANG et al. 2021] M.-D. Wang, Q.-H. Fu, M.-J. Song, W.-B. Ma, J.-H. Zhang und Z.-X. Wang. **Novel Subgroups in Subarachnoid Hemorrhage and Their Association With Outcomes—A Systematic Review and Meta-Regression**. Frontiers in Aging Neuroscience, Vol. 12:573454, 2021.

[WEIR et al. 2002] B. Weir, L. Disney und T. Karrison. **Sizes of ruptured and unruptured aneurysms in relation to their sites and the ages of patients**. Journal of neurosurgery, Vol. 96(1):64–70, 2002.

[WERMER et al. 2007] M. J. Wermer, I. C. van der Schaaf, A. Algra und G. J. Rinkel. **Risk of rupture of unruptured intracranial aneurysms in relation to patient and aneurysm characteristics: an updated meta-analysis**. Stroke, Vol. 38(4):1404–1410, 2007.

[WIEBERS 2003] D. O. Wiebers. **Unruptured intracranial aneurysms: natural history, clinical outcome, and risks of surgical and endovascular treatment**. The Lancet, Vol. 362(9378):103–110, 2003.

[WILSON und MARTINEZ 2000] D. R. Wilson und T. R. Martinez. **Reduction techniques for instance-based learning algorithms**. Machine learning, Vol. 38(3):257–286, 2000.

[WU und LIN 2005] J. Wu und Z. Lin. **Research on customer segmentation model by clustering**. In: Proceedings of the 7th international conference on Electronic commerce, 2005, pp. 316–318.

[XIANG et al. 2011] J. Xiang, S. K. Natarajan, M. Tremmel, D. Ma, J. Mocco, L. N. Hopkins, A. H. Siddiqui, E. I. Levy und H. Meng. **Hemodynamic–morphologic discriminants for intracranial aneurysm rupture**. Stroke, Vol. 42(1):144–152, 2011.

[XU et al. 2019] Z. Xu, Y.-N. Rui, J. P. Hagan und D. H. Kim. **Intracranial aneurysms: pathology, genetics, and molecular mechanisms**. Neuromolecular medicine, Vol. 21(4):325–343, 2019.

[ZHANG et al. 2022] H. Zhang, L. Li, H. Zhang, J. Liu, D. Song, Y. Zhao, S. Guan, A. Maimaitili, Y. Wang, W. Feng et al. **Small and medium-sized aneurysm outcomes following intracranial aneurysm treatment using the pipeline embolization device: a subgroup analysis of the PLUS registry**. Frontiers in Neurology, Vol. 13, 2022.

[ZHAO et al. 2018] J. Zhao, H. Lin, R. Summers, M. Yang, B. G. Cousins und J. Tsui. **Current treatment strategies for intracranial aneurysms: an overview**. Angiology, Vol. 69(1):17–30, 2018.

[ZHUANG et al. 2008] J. Zhuang, S. C. H. Hoi und A. Sun. **On Profiling Blogs with Representative Entries**. In: Proceedings of the Second Workshop on Analytics for Noisy Unstructured Text Data, AND '08, p. 55–62. 2008, Association for Computing Machinery, New York, NY, USA.

Declaration of Academic Integrity

I hereby declare that I have written the present work myself and did not use any sources or tools other than the ones indicated.

Datum: 02-11-23 .....

(Signature)