CHAPTER 1

INTRODUCTION

This chapter includes the following subtopics, namely: (1) Rationale; (2) Theoretical Framework; (3) Conceptual Framework/Paradigm; (4) Statement of the problem; (5) Hypothesis (Optional); (6) Assumption (Optional); (7) Scope and Delimitation; and (8) Importance of the study.

1.1 Rationale

Imbalanced data refers to a condition where one class has significantly fewer instances than other classes [1]. Within the context of binary classification, the class with the smallest number of instances is referred to as the minority class, while the class with the largest number of instances is designated as the majority class [1]. In real-world scenarios, many datasets have differences in the number of instances between classes [2], such as credit card fraud detection [3], hypertension classification [4], customer reviews analysis [5], mobile malware detection [6], air quality prediction [7], and anomaly detection [8]. In imbalanced data conditions, the lack of balance between classes poses a major challenge [1], as it can significantly impact the performance of classifiers. This issue becomes more critical when the class imbalance is extreme, making it harder for models to accurately identify instances of the minority class [1]. The Imbalance Ratio (IR) is a key metric used to measure class imbalance in a dataset, particularly in binary classification problems. A dataset is considered balanced when the IR equals 1 [9]. Conversely, an IR greater than 1.5 indicates that the dataset is imbalanced [10], [11], with larger IR values reflecting a more severe imbalance between the majority and minority classes [9]. This condition can affect model performance because the majority of classification algorithms do not consider class imbalance and tend to treat all instances as equally important [1]. As a result, these algorithms often treat all instances uniformly, leading to models that are biased toward the majority class [10].

To overcome this problem, various resampling methods have been proposed. Generally, these methods are divided into two categories, namely data-level approaches and algorithm-level approaches [1]. Currently, most of the related literature examines data-level methods [12] because the resampling process is performed independently of model training, allowing the resampled dataset to be used directly with traditional machine learning algorithms [13].

At the data-level, the most commonly used techniques are oversampling, undersampling, and hybrid sampling [13]. In the oversampling method, a certain number of instances in the minority class are duplicated or synthetic samples are generated to balance the data distribution with the majority class [14]. In contrast, the undersampling method reduces

the number of instances in the majority class by removing some samples, thereby aligning the data distribution more closely with the minority class [15]. In other side, hybrid combines both oversampling and undersampling techniques [16].

Several studies have been conducted to apply data-level methods to address the challenges of imbalanced datasets. Among these, several studies have specifically examined the application of oversampling [17], [18], undersampling [19], [20], and hybrid techniques [21], [22]. These three approaches represent effective solutions for addressing the issue of imbalanced data. Among these methods, undersampling techniques have demonstrated a notable advantage in tackling this challenge [23], [24]. However, it is important to note that one of the primary limitations of undersampling is the potential removal of critical instances from the dataset [25].

To address the challenges associated with undersampling, a robust and adaptive global search strategy is required [26]. One promising method is Evolutionary Computation (EC), which is well known for its effectiveness, especially because of its global search capability that supports wide exploration of the solution space [26]. This strength allows EC to effectively identify specific instances to be removed from the majority class, while reducing the risk of losing important information [27].

Several studies have explored the application of EC in sampling methods [2], [28], [29]. However, existing literature indicates that no metaheuristic optimizer can guarantee a global optimum across all problem domains [30], particularly when dealing with imbalanced data. This limitation continues to drive research toward the development of novel algorithms that offer improved scalability, stability and reliability [30].

To address the limitations mentioned above, undersampling can be combined with recent metaheuristic algorithms, such as the Komodo Mlipir Algorithm (KMA) [30]. KMA is a developed algorithm designed to address the limitations of previous metaheuristic approaches in terms of scalability, stability, and exploration capability within complex and high-dimensional search spaces. This approach adopts the feeding and reproduction patterns of the Komodo dragons native to East Nusa Tenggara, Indonesia, which are modeled through three role-based groups differentiated by quality: large males, females, and small males [30]. The key advantage of KMA lies in its introduction of two novel movement strategies: high-exploitation low-exploration (HILE), and low-exploitation high-exploration (LIHE). These strategies enable the algorithm to perform intensive local search while simultaneously exploring the broader solution space, thereby reducing the risk of entrapment in local optima. Moreover, KMA has been proven to deliver stable performance and effectively handle optimization problems across various dimensionalities [30].

This study proposes a method called Komodo Mlipir Algorithm-based Undersampling (KMAUS) to enhance the performance of classification models, particularly in handling class imbalance issues in the utilized dataset. KMAUS integrates the principles of the Komodo Mlipir Algorithm [30] with undersampling techniques to reduce instances from

the majority class by removing them without losing important information.

1.2 Statement of the Problem

The main issue with imbalanced data is that classification models tend to be biased toward the majority class, as most classification algorithms do not consider class imbalance and treat all instances as equally important [1]. To address this problem, various resampling methods have been proposed, including undersampling. Undersampling has proven to be effective in handling imbalanced data issue [23], [24]. However, a major limitation of this method is the potential removal of critical instances from the dataset [25]. To address the limitation of undersampling, a robust and adaptive global search strategy is needed [26], such as Evolutionary Computation (EC). The Komodo Mlipir Algorithm, one of the algorithms based on EC, has been shown to guarantee global optimum solutions across various problems [30]. Therefore, combining undersampling with the Komodo Mlipir Algorithm offers a promising solution to overcome the challenges associated with undersampling techniques.

To address these challenges, this study introduces an improved undersampling approach based on the Komodo Mlipir Algorithm (KMA). Inspired by the feeding and reproduction behaviors of Komodo dragons, KMA divides candidate solutions into three groups: large males, females, and small males, based on solution quality [30]. Large males perform high-exploitation low-exploration (HILE) movements that focus on intensifying the search around promising regions. In contrast, small males perform *mlipir* or stealthy movements characterized by low-exploitation high-exploration (LIHE), aiming to explore new areas of the solution space. This adaptive strategy supports a more balanced sampling process that dynamically searches the solution space without discarding potentially informative samples.

This study proposes a method called Komodo Mlipir Algorithm-based Undersampling (KMAUS) to enhance the performance of classification models, particularly in handling class imbalance issues in the utilized dataset. KMAUS is designed to reduce the majority class by retaining relevant instances and discarding excessive or less informative instances.

1. How is the performance of handling imbalanced data problems using the Komodo Mlipir Algorithm-based Undersampling (KMAUS) method?

2. How do the parameters of the Komodo Mlipir Algorithm-based Undersampling (KMAUS) method affect the model performance in imbalanced data classification?

3. How does the Komodo Mlipir Algorithm-based Undersampling (KMAUS) method perform in comparison to baseline and benchmark methods across 100 imbalanced datasets from the KEEL repository?

1.3 Objective and Hypotheses

1.3.1 Objectives

1. Developing a method for handling imbalanced data using the Komodo Mlipir Algorithm-Based Undersampling (KMAUS) approach.
2. Evaluating the impact of the Komodo Mlipir Algorithm-based Undersampling (KMAUS) parameters on the performance of imbalanced data classification.
3. Comparing the performance of KMAUS to baseline and benchmark methods on 100 imbalanced datasets.

1.3.2 Hypotheses

This study hypothesizes that the Komodo Mlipir Algorithm-based Undersampling (KMAUS) approach can improve classification performance on imbalanced datasets. The hypothesis is founded in several premises. First, undersampling methods such as RUS, TL, ENN, and NM have limitations, as they may potentially remove important data instances from the majority class [25]. Therefore, a robust and adaptive global search strategy is needed to more effectively select which majority class instances should be removed [26], [27]. Second, KMAUS combines an undersampling technique with the Komodo Mlipir Algorithm, which employs two distinct movement patterns based on data quality. High-quality big males focus the search on promising regions of the solution space, while low-quality small males explore broader areas [30]. In addition, KMAUS includes a population adaptation mechanism that adjusts the number of individuals during the solution search process [30]. Based on these premises, this study evaluates the effectiveness of KMAUS by comparing it with undersampling methods as baseline method, and undersampling that use several metaheuristic algorithms as benchmark methods.

1.4 Assumption

In this study, several assumptions were made. First, it is assumed that the binary classification datasets from the KEEL repository contain sufficient feature representations for effective learning and classification under imbalanced conditions. Second, it is assumed that the 100 binary-class datasets taken from the KEEL repository represent real-world imbalanced data scenarios, so that the experimental results can provide a valid basis for evaluating the effectiveness and capability of the proposed method compared to baseline and benchmark methods.

1.5 Scope and Delimitation

This study focuses on improving binary classification performance on imbalanced datasets by employing the Komodo Mlipir Algorithm-based Undersampling (KMAUS) method. Specifically, it addresses the issue of class imbalance by applying KMAUS to reduce the number of majority class samples using an undersampling approach. The datasets utilized in this study consist of 100 binary-class datasets from the KEEL dataset repository, which is widely recognized as a benchmark in studies on imbalanced classification. This study is limited to binary classification tasks with datasets obtained from the KEEL repository and does not involve more complex data types such as multiclass, time-series, or image data. This study concentrates on undersampling techniques and does not incorporate oversampling or hybrid approaches.

1.6 Significance of the Study

The main contribution of this study is the introduction of the Komodo Mlipir Algorithm-based Undersampling (KMAUS) method, which was developed to solve the problem of data imbalance in binary classification tasks. This method helps handle class imbalance effectively and works well on datasets with different levels of imbalance, from mild to severe.