Journal of Theoretical and Applied Information Technology

15th September 2023. Vol.101. No 17

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ENHANCING RARE DISEASE DIAGNOSIS: A WEIGHTED COSINE SIMILARITY APPROACH FOR IMPROVED K-NEAREST NEIGHBOR ALGORITHM

SOMIYA ABOKADR[1], AZREEN AZMAN[2], HAZLINA HAMDAN[3], NURUL AMELINA[4]

[1]Department of Intelligent Systems, Faculty of Computer Science & Information Technology, University Putra Malaysia, Seri Kembangan 43400, Malaysia and Faculty of Science, Al Zintan University Libya
[2]Deputy Dean (Development, Industry and Community Linkages), Faculty of Computer Science & Information Technology, University Putra Malaysia, Seri Kembangan 43400 Malaysia
[3]Department of Computer Science, Faculty of Computer Science and Information Technology, University Putra Malaysia
[4]Department of Multimedia Faculty of Computer Science and Information Technology University Putra Malaysia43400 UPM Serdang, Selangor Darul Ehsan
E-mail: soma.almoktar@gmail.com, 2azreenazman@upm.edu.my, 3hazlina@upm.edu.my, 4nurulamelina@upm.edu.my

Corresponding authors : SOMIYA ABOKADR, AZREEN AZMAN

ABSTRACT

Diagnosing rare diseases is challenging because they affect only a restricted group of individuals, usually identified as one out of every 2,000 people within the European Union and no more than one out of 1,250 individuals in the United States. This makes it difficult for doctors to recognize the symptoms of these diseases. This paper focuses on the challenges of diagnosing rare diseases due to their low prevalence rates and difficulties in recognizing their symptoms.

Machine learning techniques often face difficulties in classifying patients with rare diseases because of their small sample sizes, leading to biased results. They proposed a weighted cosine similarity approach as a distance measure for the k-nearest neighbours algorithm instead of the conventional cosine similarity to address this issue. The use of genetic optimization to select the best weights for the weighted cosine similarity.

The Rare Metabolic Diseases Database was used as a case study, and the results demonstrated that reducing the classification bias between majority and minority classes improves all classification performance measures. However, as the number of classes and imbalance ratio increase, the approach's effectiveness decreases, eventually reaching zero. Future work will focus on reformulating the g-mean to smooth its values and avoid assigning a zero score when all class instances are misclassified.

Keywords: *K-Nearest Neighbor, Cosine Similarity, Imbalance Data, Genetic Algorithm, Imbalance Ratio.*

1. INTRODUCTION

Diagnosing rare diseases is a complex and challenging process that involves identifying symptoms and conducting various tests to confirm the presence of the disease [1]. Due to their rarity, healthcare professionals often struggle with correctly diagnosing these diseases, and the shared symptoms with more common conditions can further complicate the diagnostic process. Sophisticated diagnostic tools and collaboration between medical professionals, including specialists and genetic counselors, are often necessary for a definitive diagnosis [2]. A disease is a deviation from good health or abnormal functioning, typically caused by various factors, including infection, genetic defects, or environmental stress. It is a pathological state affecting an organism's part, organ, or system. Patients may identify as having either a disease or a disorder[3].

A rare or uncommon type of disease is referred to as a rare disease, which affects only a limited number of individuals in the general population. For a specific disease to be categorized as rare, it must not affect more than a limited number of people from the entire population[4].

In addition to disease prevalence, several factors typical of uncommon diseases are considered while diagnosing these diseases [5]. For instance, there is no effective cure for most of these disorders, which are chronic, progressive, life-threatening, induce degeneration of body tissues, and result in impairment [5],[6]. Therefore, in 80% of cases, these diseases have hereditary causes, 50% to 70% of patients are children, and 30% pass away before they turn five [5], [6]. There are currently between 5,000 and 7,000 known rare diseases, and more are continually being discovered. In addition, a majority of recognized illnesses can be categorized into several vital domains, such as autoimmune conditions, metabolic disorders, neuromuscular ailments, blood disorders, cardiovascular and respiratory issues, skin disorders, and rare tumors are among the ailments that have been identified [6].

The World Health Organization (WHO) defines a disease as uncommon if it occurs in less than 6.5 to 10 out of every 10,000 individuals. Applying this criterion, Less than 5 out of every 10,000 people in the EU, or 1 in 2,000, are affected by such diseases. Conversely, less than 200,000 people in the USA have been identified as having uncommon diseases, compared to less than 50,000 in Japan and 2,000 in Australia [5],[7]. Hence, Most nations have used the EU definition as the basis for their rare national disease strategies[5].

Diagnosis targets vary from one rare disease to all rare diseases, and the percentage of patients with the same disease also varies, which leads to an imbalanced dataset[8]. According to a survey by Global Genes, approximately 40% of general practitioners and 24% of specialists lack time to concentrate on these diagnoses [3]. The majority of clinicians also have little awareness of these disorders. Despite medical advancements, patients with rare diseases often face misdiagnosis or lack of diagnosis due to limited scientific knowledge and clinical experience. One major challenge in dealing with rare diseases is accurately diagnosing patients. Based on research carried out by the European Union, a quarter of individuals with uncommon medical conditions had to endure a wait of 5 to 30 years before receiving an accurate diagnosis[9].

Although next-generation sequencing technology has effectively pinpointed the genetic cause of uncommon Mendelian diseases, the current diagnosis rate is still not considered satisfactory due to the variability, vagueness, and inaccuracies in describing disease symptoms. Moreover, the research does not fully utilize clinical genetics experts' knowledge. Most information utilized to support the diagnosis was found in phenotypic concepts, pictures, or fluid laboratory testing. The research aims to delimit the topic of rare disease diagnosis, explicitly addressing the challenges of accurate identification, the use of AI methodologies, and the importance of addressing imbalanced class distributions. The propose a method combining the k-nearest neighbors' algorithm, cosine similarity weighting, and a genetic algorithm to improve classification performance in rare diseases.

Rare disease diagnosis is considered a classification problem, which can be solved using artificial intelligence (AI). In this contest, two fundamentally distinct classifications can be made of AI methodologies: Data-driven AI versus knowledge-based AI. The knowledge-based approaches openly describe knowledge as symbols representing rules, ontologies, historical data, or other knowledge structures. To function at an acceptable level, data-driven AI needs large amounts of data and powerful processing [10]. In this context, studies that used knowledge-based AI accounted for 57% of the studies. Two-thirds of the data-driven AI utilized machine learning methods, and the remaining one-third used superficial similarities [8].

Real-world datasets often exhibit an imbalanced nature, with some classes being overrepresented while others are underrepresented. This imbalance can result in a bias towards the majority classes in classification to the detriment of the minority classes. Therefore, addressing imbalanced class distributions in datasets is essential to avoid such biases. Even the minority class is considered more interesting when the minority classes are insufficiently represented. The issue of imbalanced class distributions is widely recognized and acknowledged in everyday applications such as medical diagnosis, spam detection, image processing, and fraud detection in the banking sector.

In the case of a medical diagnosis involving rare diseases, it is crucial to identify such conditions within the general population accurately. Also, this condition occurs within the dataset with the rare disease only where the percent of patients with the same disease. As a result, any mistakes made in diagnosis will harm the patient's health and can cause added stress and complications for the patient. Furthermore, it can also have implications for the treatment plan and medication regimen. Therefore, a classification model should be capable of achieving a higher level of accuracy in identifying the minority classes within datasets[11]. Even though the imbalance issue exists in rare disease datasets, no

existing study using phenotype concepts material reported the result of reducing imbalance impact on rare disease diagnosis. There are two suggested methods for resolving this problem. The initial method consists of amplifying the number of records for underrepresented categories, and the second method involves adjusting the classification algorithm[12].

This paper focus on diagnosing rare diseases and addressing the challenges associated with accurately identifying these conditions.

- The aim is to develop a classification model using artificial intelligence (AI) techniques, specifically the k-nearest neighbors algorithm enhanced with a genetic algorithm, to improve the accuracy of rare disease diagnosis.
- The paper delves into the characteristics of rare diseases, their prevalence, and the factors considered during diagnosis. We highlight these diseases' chronic, progressive, and life-threatening nature and their hereditary causes and impact on different age groups, particularly children.
- Identify rare disease diagnosis as a classification problem and propose AI methodologies, specifically data-driven and knowledge-based approaches, as potential solutions.
- Developing a balanced margin between majority and minority classes in rare disease classification. We propose using the k-nearest neighbors' algorithm with a genetic algorithm to optimize the cosine similarity weighting method and maximize the G-mean value, which evaluates unbiased performance across all classes.

2. RELATED WORK

Diagnosis support for rare diseases is based on various materials, mainly phenotypic concepts, images, or fluids. The Human Phenotype Ontology (https://hpo.jax.org) (HPO) is considered the source for providing a set of commonly used terms and phrases to describe the physical and observable characteristics associated with human illnesses [8]. Machine learning techniques applied to Clinical databases aim to identify connections and trends within medical and pathological data related to patients and their health conditions. The goal is to gain insight into the development and characteristics of particular illnesses, enabling earlier detection and diagnosis. Machine learning is considered a broadly accepted approach for rare disease diagnosis 42 studies out of 61 used the Machine learning approach for rare disease diagnosis [8].

K-nearest neighbors (KNN) is the most popular algorithm used for rare disease diagnosis based on phenotype concepts material. Jia et al. [13] Four diagnostic models were created utilizing a method of measuring physical similarities (phenotypic similarity) and a machine learning technique to diagnose rare diseases.

The Rare Metabolic Disease Database (RAMEDIS) open medical records validated each diagnostic model. In this case, the classifier's performance varied in different cases. However, after implementing the Bayesian averaging approach in the four models, merging the classifiers' predictions, and ranking the potential rare diseases according to the score. The outcome of each model produces a list of the top 10 likely diseases for diagnosis. The findings revealed that all models demonstrated exceptional diagnostic accuracy, with precision levels of up to 98% and the highest recall rate of 95%. In comparison, the models utilizing machine learning techniques exhibited superior performance.

The website (http://www.unimd.org/RDAD/) allows unrestricted access to the system. Schaaf et al. [14] apply similarity analysis to 10 university hospitals in Germany using similarity methods to offer physicians hints about a possible patient diagnosis. Feichen Shen et al. [15] Collected data from Mayo Clinic and utilized similar methods with the KNN algorithm to assist in rate disease diagnosis. Shen et al. [16] also collected data from Mayo Clinic and utilized four similarity methods with two neighborhood algorithms to hypothesize the phenotype data can be leveraged to accelerate disease diagnosis. Sheikhzadeh et al. [17]used multivariate logistic regression on a sample of patients to assess the inclusion of clinical variables for inclusion in the prediction model for rare disease diagnosis. Li et al. [18] proposed new similarity methods to improve the effectiveness of similar approaches that rely on physical characteristics shared among individuals with a given disease, a new technique called Emission-Reception Information Content (ERIC) has been developed. The new methods include terms calculated based on the relationship between phenotype and genotype.

Healthcare professionals contributed factual clinical data, including Human Phenotype Ontology (HPO) traits and causative genetic variants. These data were then utilized to evaluate the efficacy of the

ERIC similarity methods. In other consideration, Burange & Chatur [19] proposed a deep artificial neural network model (ANN) using multi-layers for rare disease diagnosis. RAKE (Rapid Automatic Keywords Extraction) algorithm extracts keywords from specified symptoms. These keywords are encoded to become the input for the ANN input layer. Garcelon et al. [20] did also a study to identify rare disease people using medical report phenotype.

The researchers utilized Electronic Health Records as a secondary data source to expand their understanding of uncommon illnesses. Through their investigation, they implemented a technique to identify the characteristics linked with specific diseases. They employed frequency and TF-IDF measures to examine the connection between clinical features and rare diseases in a practical manner. This approach was employed in six scenarios to determine phenotypes related to the conditions known as Rett syndrome, Lowe syndrome, Silver Russell syndrome, Bardet-Biedl syndrome, DOCK8 deficiency, and Activated PI3-kinase Delta Syndrome (APDS) are mentioned or incorporated[21].

A. Problem Statement

Despite advancements in next-generation sequencing technology and AI methodologies, the current diagnosis rate for rare diseases is still not considered satisfactory. The variability, vagueness, and inaccuracies in describing disease symptoms pose significant obstacles to accurate identification. Moreover, the issue of imbalanced class distributions within real-world datasets further compounds the problem. This imbalance leads to biases in classification, favoring majority classes and hindering the accurate identification of minority classes. This bias can harm patients, including misdiagnosis, delayed treatment, and potential health complications.

The paper aims to address this problem by proposing a novel method that combines the k-nearest neighbors' algorithm, cosine similarity weighting, and a genetic algorithm. This approach seeks to enhance classification performance in the context of rare diseases and mitigate the impact of imbalanced class distributions.

3. METHODOLOGY

This section describes the dataset used to apply the experiment and how the features are extracted, followed by a preliminary for the KNN algorithm, genetic algorithm, and cosine similarity function. Finally, the proposed classification method is illustrated.

3.1 Dataset Description

RAMEDIS dataset is a public biomedical data dataset, as depicted in Figure 1. The dataset contains

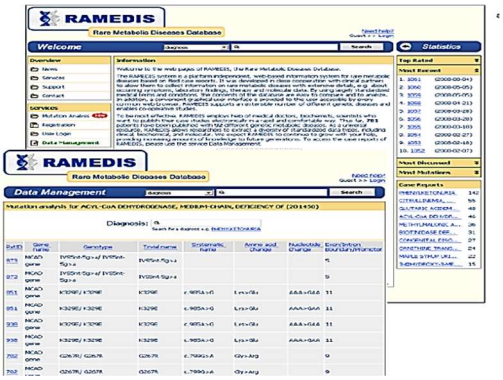


Figure 1: A Screenshot of the RAMEDIS data repository.

Five hundred eighty patients and 80 diseases. Figure 2 shows the structure of the dataset. Table 1 lists the statistics about the dataset in September 2009.

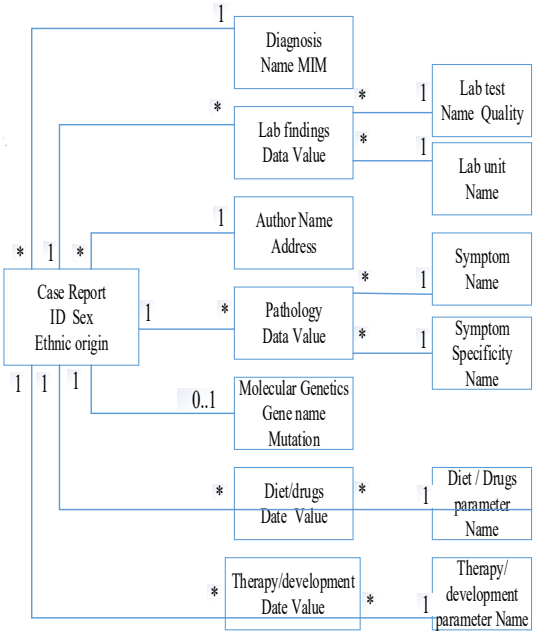


Figure 2: Schema representation of data in the RAMEDIS repository

RAMEDIS is an internet-based information system for rare diseases not tied to any platform. To prepare the data for analysis, a preliminary stage involves purging the database of duplicate entries and performing other cleaning procedures.

3.2 Feature Extraction

Feature extraction involves converting raw data into numerical features that can be processed using machine learning techniques while retaining the information present in the original dataset.

Using equation 1, features were extracted for each patient by applying TF-IDF to medical terms found on the list of records in symptom fields and historical description fields words which have the corresponding item in HPO ontology item (https://hpo.jax.org/app/download/ontology).

$$TF - IDR(t,d) = TF(t,d) * IDF(t) \quad (1)$$

The TF-IDF equation used in the study considered N, the total number of patients, t, the specific symptom being analyzed, d, the current patient being evaluated, TF, which represents the frequency of the symptom of the IDF is a value calculated from the number of patients with a particular symptom present in their medical records.

3.3 K-Nearest Neighbors Algorithm

KNN is a simple supervised machine learning algorithm that can be applied to classification problems. In KNN, each object is classified based on the voting of its k nearest neighbors. KNN relays on a distance metric to get nearest neighbors, i.e., Euclidian distance and cosine similarity …etc.

3.4 Cosine Similarity Function

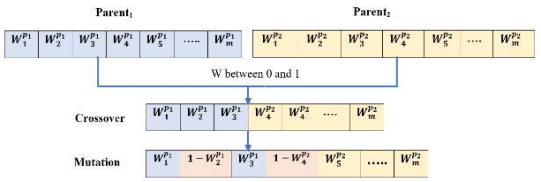
Cosine similarity measures the similarity between two vectors to determine if two vectors point in the same direction and are used in document analysis. It has two forms, the classic form, which is given by Equation 2, and the weighted form, which is given by Equation 3:

$$sim(X,Y) = \frac{\sum_{t \in T}(X_t)*(Y_t)}{\sqrt{\sum_{t \in T}(X_t)^2}*\sqrt{\sum_{t \in T}(Y_t)^2}} \quad (2)$$

$$sim(X,Y,W) = \frac{\sum_{t \in T}(X_t*W_t)*(Y_t*W_t)}{\sqrt{\sum_{t \in T}(X_t*W_t)^2}*\sqrt{\sum_{t \in T}(Y_t*W_t)^2}} \quad (3)$$

In the given context, Xt and Yt refer to dimension t for vectors X and Y, respectively, while Wt represents the weight assigned to symptom t.

3.5 Genetic Algorithm



A genetic algorithm is a search algorithm that uses meta-heuristic methods to find an optimized solution. It inspires by biological operations like genes crossover and mutation. Figure 3 shows the main steps of the standard genetics algorithm. It starts with generating the initial solutions population, which is evaluated for the best objective function and follows.

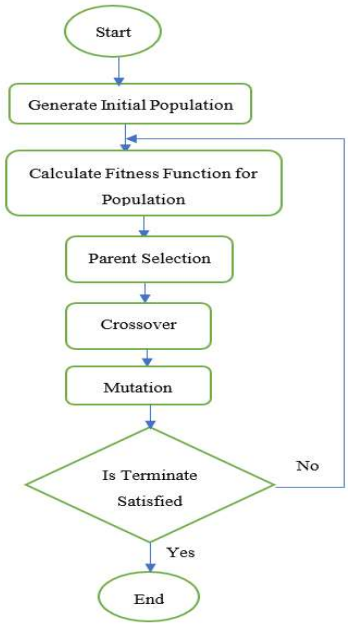


Figure 3: Standard Genetic Algorithm Flow Chart

Select candidate parent solutions to generate new sibling solutions for the next generation. Generic operations use crossover and mutation to produce sibling solutions from parent solutions. The genetic algorithm can solve many optimization problems, i.e., salesman problems, hyper-parameter optimization …etc. The Genetic algorithm can use different selections for parents, and the methodology employed selection, crossover, and mutation techniques, which can be observed in Figure 4 and Figure 5, displaying the specific operations utilized.

In this work, each solution consists of weights equal to the number of unique symptoms in the database. Each weight is a random value between 0 and 1. The selection operation uses roulette-wheel selection. All solutions are lined up in a queue with a length equal to their g-mean value to get a selection change proportional to their g-mean value.

Figure 4: Crossover and Mutation method

The selection is based on a random selection point over the queue line. The crossover operation is done by selecting two parents and generating a new sibling based on an arbitrary point. The new sibling consists of part of parent one before this point and part of parent two after this point. A mutation is done by selecting a random set of random points to set its values to complement the cell value.

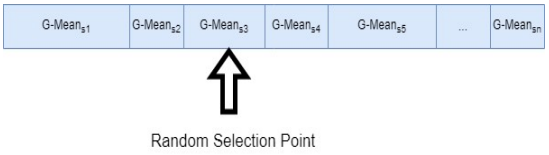


Figure5 : Roulette-wheel selection

3.6 Performance Measure

The usual metrics for evaluating binary classification performance are accuracy, precision, recall, and f1-score. These measures are adapted for multi-class classification to report the average values for each class using arithmetic [22],[23], and geometric means. The concluded set of harmonic measures for multiclass classification averages the measure for each class without considering the class size. The following measure can compute as follow:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

$$Pr\,e\,cision_k = \frac{TP_k}{TP_k+FP_k} \quad (5)$$

$$R\,e\,call_k = \frac{TP_k}{TP_k+FN_k} \quad (6)$$

$$MacroAverage\,Pr\,e\,cision_k = \frac{\sum_{k=1}^{K} Pr\,ecisio_{\ k}}{K} \quad (7)$$

$$MacroAverage\,Re\,c\,all = \frac{\sum_{k=1}^{K} Re\,ca_{\ k}}{K} \quad (8)$$

$$MacroF_1 - Score = 2 * \left(\frac{MAP*MAR}{MAP^{-1}+MA^{-1}}\right) \quad (9)$$

Where, MAP: Macro Average Precision
MAR : Macro Average Recall

$$G - mean = \sqrt[k]{\prod_{k=1}^{K} \frac{TP_k}{TP_k+FN_k}} \quad (10)$$

The given parameters are defined as follows:
K represents the total number of classes, while k represents the evaluated class.
TP refers to the count of positive samples accurately recognized as positive. Simultaneously, FN denotes the count of positive samples incorrectly classified as unfavorable, while TN denotes the count of negative samples accurately classified as harmful. FP indicates the count of negative samples erroneously classified as positive. These parameters are used for each of the K classes being analyzed.

3.7 Cross-validation

Cross-validation generates reliable results, and the technique of 5-fold cross-validation is applied, whereby the dataset is split into five sections. In each iteration, one section is designated as the test set. The other folds are used for training. Each class instance should distribute equally on folds. This condition restricted the number of instances to at least five for each class. The computation is repeated five times to get the average of the classification performance. Figure 6 shows how the 5-fold cross-validation is performed.

3.8 Classification Procedure

Classification algorithm one consists of three phases. In the first phase, the features are extracted from the dataset regarding TF-IDF for each patient. Search for optimized cosine similarity weights using a genetic algorithm in the second phase. The objective function for the genetic is the G-mean reported by the KNN algorithm that uses weighted cosine similarity. In the final phase, the best weights that generate max G-mean using the KNN algorithm are reported with all classification performance measures in section 3.5.

In algorithm 1, the input consists of the dataset, genetic algorithm parameters, and KNN algorithm number of neighbors. The output reports the classification performance for the best population individual based on G-Mean for classification using the KNN algorithm with optimized weighted cosine similarity. In steps 1-3, unusual symptoms for all patients are extracted and stored as a variable. In steps 4-6, each patient feature is calculated for each period in *terms* of the variable using Eq. (1). In step 7, an initial population for the genetic algorithm is generated as the matrix of two dimensions with the number of rows equals to *N,* and length equals to the number of *terms*. Each row represents an expected solution for cosine similarity weights, and each cell is initialized with a random value from 0 to 1.

In Steps 8-19, the genetic algorithm runs from *T* of iteration to find the optimized weights that report the best G-mean. In each iteration, all population individuals, which represent weights for

cosine similarity, classification performance is evaluated based on KNN and cosine similarity using corresponding individual weights. The evaluation helps to select two parents using Roulette-wheel selection based on G-mean to apply crossover and mutation operation to generate new individuals based on probabilities C_*prop* and *M_prop*. At the end of each iteration, the best individual is saved based on the best G-mean. In step 20, the best solution is reported with all classification performance measures in section 3.6.

Algorithm 1: Enhance the k-nearest neighbors algorithm

Input:
DS: Rare Disease Dataset
T: Number of Iteration
N: Population Number
C_prop: Crossover Probability
M_prop: Mutation Probability
K: number of nearest neighbors
Output:
Optimized Cosine Similarity Weights
Classification Performance

Phase 1: Features Extraction
Terms ← {}
For each *patient* **in** *DS*
    Terms ← *Terms* ∪ *Terms* (*Patient*)
For each *patient* in *DS*
    For each *term* **in** the *Terms*
        Calculate TF-IDF (*patient*, *term*)
Phase 2: Optimize Cosine Similarity Weights using Genetics Algorithm
Initial population *P* with number of *N* solutions and dimension equals to length set *Terms*
t = 0
While t < T
    Evaluation of each solution in *P* using cross-validation
    For each *solution* **in** *P*
        Select two parents, p_1, *and* p_2 solutions, using a roulette-wheel algorithm based on G-mean for each solution.
        If random < c_prop
            Apply crossover between p_1 *and* p_2 *to generate a new solution.*
            If random < m_prop
                Apply mutation to the new generation new solution.
            Replace the *solution* in population *P* with the new solution.
    Save the best solution based on G-Mean.
    t ← t – 1
Phase 3: Report Results
Report the Best solution and classification performance.

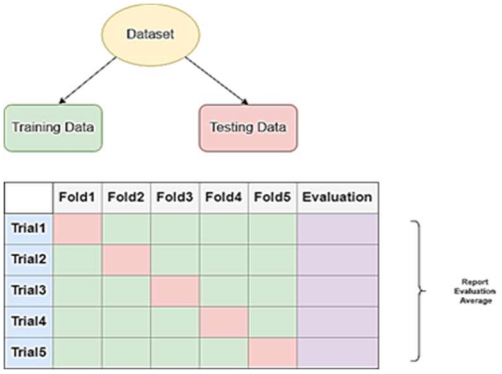


Figure 6: Cross-validation method

4. EXPERIMENTS

This section outlines the experimental setup for the study, specifically examining how the existing method performs as the number of classes increases, as well as evaluating the efficacy of the suggested approach in enhancing the current technique.

4.1 Experimental Setup

In this experiment, the RAMEDIS proposed a number of diseases equal to 80. The diseases in descending sorted in order according to the number of patients in each disease. Diseases were excluded with several patients less than five patients to match the cross-validation condition, which resulted in 29 diseases only. Then, apply algorithm 1 for the first three up to 29 diseases using the setting in *Table 2*.

Table 1: Experimental Settings

Parameter	Value
Number of Population	50
Crossover Probability	90%
Mutation Probability	10%
Number of Iterations	30
Number of Nearest Neighbors	5

Table 2 : Experimental Settings

Dataset Objects	Number of Entries
valid authors	77
valid case reports	589
molecular genetics	372
unique symptoms	580
lab findings	40,804
therapy, development	6,622
diet, drugs	2,210
pictures	91
references	73
genes	26,779
Enzymes	4,609

pathways	169
compounds	13,972
diagnoses	17,508
symptoms	580
lab findings in different specimen	1,360

5. RESULT AND DISCUSSION

In this section, they perform the proposed method shown in the extensive simulation results. The results were evaluated regarding imbalance ratio, accuracy, precision, recall, f1-score, and G-mean at various numbers of classes. It is assumed that the number of diseases equals 80 and the number of classes is 29. The multi-class classification starts from 3 classes and ends with 29 classes where the number of patients equals 5, where at least one patient exits on each fold in the cross-validation process. The genetic optimization proposed to measure the classification performance of the majority and minority classes based on a weighted cosine similarity as a distance measure for the k-nearest neighbor's algorithm.

The imbalanced ratio will influence the performance of the number of classes, as recorded can be too increased to be delivered correctly. Thus, as illustrated in Fig.7, we examine the ratio to ensure its effect on the number of classes. The performance of imbalance ratio parameters will versus increase the number of classes mainly due to a decrease in the G-mean characteristics. For example, in the case of the number of classes less than 13, the imbalance ratio performance is between (0-5). In addition, when the number of classes increased to the maximum value of 29, the imbalance ratio performance increased to approximately 10. This is due to the fact the disease classes are sorted in descending according to the number of patients with each disease.

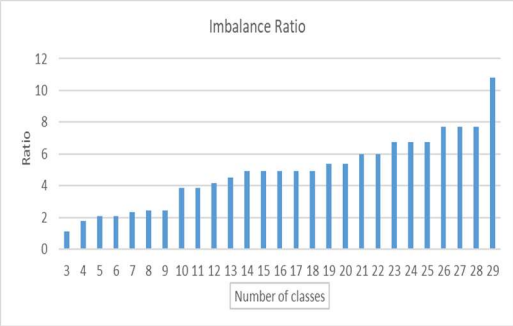


Figure 7 : Imbalance Ratio For Each Number of classes

Figures 8-12 present the classification performance metrics for varying classes, demonstrating that the number of classes significantly impacts the proposed method's effectiveness.

This becomes particularly clear for the imbalanced data set, which resulted in bias between the majority and minority classes.

In this case, changing classes effect improves the percentage of good correspondence by 3 to 12 classes. This dataset contains the best weight to get the maximum value for G-mean, which consider unbiased for all classes. As anticipated, a small number of classes does not provide enough detail to accurately assess the classification performance, while including too many classes may result in excessive computations for all performance measurements. The decrease in classification accuracy, precision, recall, and G-mean as the number of classes increases can be attributed to using smaller ranges for each class. This results in more fuzzy and unstable regions, with data points that may be assigned to neighboring ranges.

To support this claim, we evaluated the classification accuracy, precision, recall, and G-mean of the relationship between the number of classes and the corresponding model points, allowing us to examine the data for each point in the model through a k-nearest neighbors algorithm to enhanced weights with the genetic algorithm based on cosine similarity.

As previously stated, the increase in the number of classes resulted in a decline in classification accuracy, ultimately reducing overall classification performance.

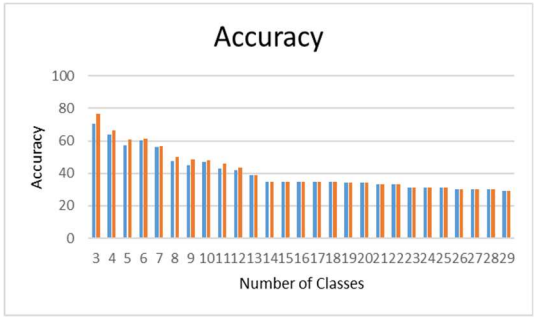


Figure 8 : Comparison Of Classification Accuracy

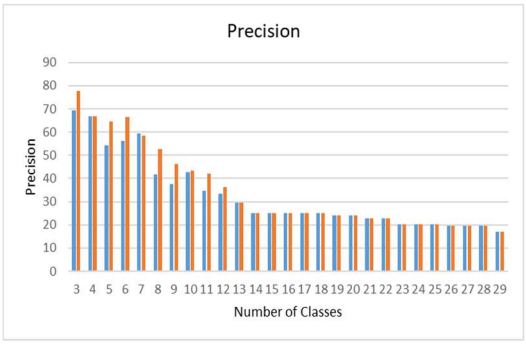


Figure 9 : Comparison Of Classification Macro-Precision.

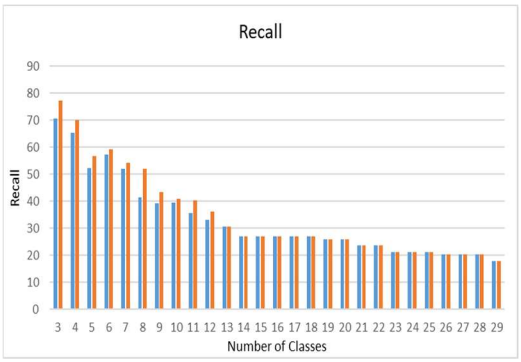


Figure 1 : Comparison Of Classification Macro-Recall

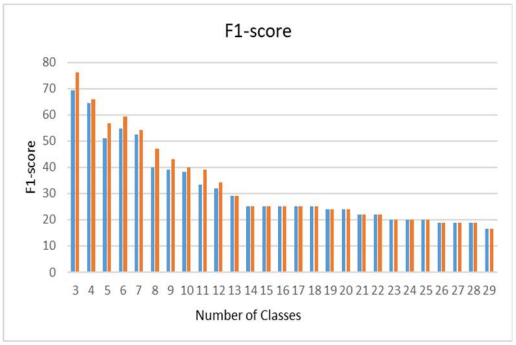


Figure 11 : Comparison Of Classification Macro-F1-Score.

Table 3 : G-Mean result

No. of Classes	Benchmark	Work of paper
3	69.72	76.13
4	63.14	68.97
5	19.78	51.15
6	11.86	54.04
7	30.12	41.26
8	0	34.24
9	0	17.51
10	0	9.02
11	0	9.117
12	0	9.68

The results show that using the genetic algorithm to optimize the weights used in the cosine similarity function enhances all performance metrics. The average for the classification performance is based on macro for accuracy, macro-precision, macro-recall, f1-score, and G-mean. The accuracy is increased to 6.66 %, precision up to 8.51 %, recall up to 6.61%, f1-score up to 6.84, and G-mean up to 52.87%. The results show that the significant impact is on G-mean as it's used as an objective for generic algorithms, and the enhancement of the G-mean leads to the enhancement of all classification performance metrics. The G-mean metric reaches zero when the number of classes and imbalance ratio increases.

6. CONCLUSION

Unbalanced class distribution is a significant challenge in machine learning classification. Often, the resulting classification can exhibit a bias toward majority classes at the expense of minority classes. Two methods have been put forward to overcome this problem: increasing the number of records for minority classes and modifying the classification algorithm. In this study, we utilize the genetic algorithm to select an optimized set of weights for the cosine similarity distance measure used in the k-nearest neighbors algorithm. This enhanced approach is applied to the RAMEDIS dataset as a case study. The experiment enhances the accuracy, precision, recall, f1-score, and G-mean. This enhancement is decreased and reaches zero when the number of classes and imbalance ratio increase. Future work can handle this issue by reformulating the G-mean by smoothing its values to avoid giving zero when all instances of any classes are misclassified.

REFERENCES:

[1] A. Briolay, L. Bessueille, and D. Magne, "TNAP: a new multitask enzyme in energy metabolism," *International Journal of Molecular Sciences,* vol. 22, no. 19, p. 10470, 2021.

[2] H. Cleckley, "Future Implications of the Psychopathy Construct for Criminology and

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Criminal Justice Policy and Practice," *No Remorse: Psychopathy and Criminal Justice,* p. 277, 2018.

[3] J. Rode, "Rare diseases: understanding this public health priority," *Rare Dis,* vol. 14, p. 3, 2005.

[4] H. MacLeod, K. Oakes, D. Geisler, K. Connelly, and K. Siek, "Rare world: Towards technology for rare diseases," in *Proceedings of the 33rd Annual ACM Conference on human factors in computing systems*, 2015, pp. 1145-1154.

[5] H. MacLeod, S. Yang, K. Oakes, K. Connelly, and S. Natarajan, "Identifying rare diseases from behavioural data: a machine learning approach," in *2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, 2016: IEEE, pp. 130-139.

[6] M. Santoro *et al.*, "Rare disease registries classification and characterization: a data mining approach," *Public health genomics,* vol. 18, no. 2, pp. 113-122, 2015.

[7] S. Saeb *et al.*, "Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study," *Journal of medical Internet research,* vol. 17, no. 7, p. e4273, 2015.

[8] J. Schaaf, M. Sedlmayr, J. Schaefer, and H. Storf, "Diagnosis of Rare Diseases: a scoping review of clinical decision support systems," *Orphanet journal of rare diseases,* vol. 15, no. 1, pp. 1-14, 2020.

[9] R. Dragusin *et al.*, "FindZebra: a search engine for rare diseases," *International journal of medical informatics,* vol. 82, no. 6, pp. 528-538, 2013.

[10] S. Montani and M. Striani, "Artificial intelligence in clinical decision support: a focused literature survey," *Yearbook of medical informatics,* vol. 28, no. 01, pp. 120-127, 2019.

[11] H. Jegierski and S. Saganowski, "An "outside the box" solution for imbalanced data classification," *IEEE Access,* vol. 8, pp. 125191-125209, 2020.

[12] A. Ali, S. M. Shamsuddin, and A. L. Ralescu, "Classification with class imbalance problem," *Int. J. Advance Soft Compu. Appl,* vol. 5, no. 3, 2013.

[13] M. A. U. H. Tahir, S. Asghar, A. Manzoor, and M. A. Noor, "A classification model for class imbalance dataset using genetic programming," *IEEE Access,* vol. 7, pp. 71013-71037, 2019.

[14] Y. Mi, "Imbalanced classification based on active learning SMOTE," *Research Journal of Applied Science Engineering and Technology,* vol. 5, pp. 944-949, 2013.

[15] F. Shen, S. Liu, Y. Wang, A. Wen, L. Wang, and H. Liu, "Utilization of electronic medical records and biomedical literature to support the diagnosis of rare diseases using data fusion and collaborative filtering approaches," *JMIR medical informatics,* vol. 6, no. 4, p. e11301, 2018.

[16] F. Shen, S. Liu, Y. Wang, L. Wang, N. Afzal, and H. Liu, "Leveraging collaborative filtering to accelerate rare disease diagnosis," in *AMIA Annual Symposium Proceedings*, 2017, vol. 2017: American Medical Informatics Association, p. 1554.

[17] S. Sheikhzadeh *et al.*, "A simple clinical model to estimate the probability of Marfan syndrome," *QJM: An International Journal of Medicine,* vol. 105, no. 6, pp. 527-535, 2012.

[18] Q. Li, K. Zhao, C. D. Bustamante, X. Ma, and W. H. Wong, "Xrare: a machine learning method jointly modeling phenotypes and genetic evidence for rare disease diagnosis," *Genetics in Medicine,* vol. 21, no. 9, pp. 2126-2134, 2019.

[19] T. P. Burange and P. Chatur, "Analysis of Symptoms Wise Disease Inference System Using Data Mining Technique," in *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2018: IEEE, pp. 1160-1165.

[20] N. Garcelon *et al.*, "Finding patients using similarity measures in a rare diseases-oriented clinical data warehouse: Dr. Warehouse and the needle in the needle stack," *Journal of biomedical informatics,* vol. 73, pp. 51-61, 2017.

[21] L. Cheng *et al.*, "Computational methods for identifying similar diseases," *Mol Ther-Nucl Acids,* vol. 18, pp. 590-604, 2019.

[22] M. Grandini, E. Bagli, and G. Visani, "Metrics for multi-class classification: an overview," *arXiv preprint arXiv:2008.05756,* 2020.

[23] E. R. Fernandes and A. C. de Carvalho, "Evolutionary inversion of class distribution in overlapping areas for multi-class imbalanced learning," *Information Sciences,* vol. 494, pp. 141-154, 2019.