Soft Set Weighted Classification for Wilson Disease Detection Using Genetic Algorithm*

Emilia Bezrutczyk, Weronika Stefańska

Faculty of Applied Mathematics, Silesian University of Technology, Kaszubska 23, 44100 Gliwice, Poland

Abstract

Wilson's disease is a rare genetic disorder caused by mutations that impair copper metabolism, potentially leading to severe neurological and hepatic damage if not diagnosed and treated promptly. Early diagnosis is therefore critical. In this study, we utilize a medical dataset comprising both individuals diagnosed with Wilson's disease and those at risk. Prior to classification, the dataset was preprocessed by handling missing values, normalizing numerical features, and splitting the data into training and testing subsets in a 70:30 ratio.We apply a soft set-based classification algorithm, incorporating two distinct soft decision tables to explore the most effective configuration for diagnostic accuracy. The best-performing configuration achieved an accuracy of 99.47%. Furthermore, a genetic algorithm-based feature weighting mechanism was employed to identify the most influential attributes, further enhancing the classifier's performance. After 10 generations, the final weighted classifier reached an accuracy of 99.97%. These results demonstrate the potential of our approach for interpretable, accurate, and efficient early diagnosis of Wilson's disease.

Keywords

Wilson disease, soft sets, weighted classification, disease detection, genetic algorithm, heuristic algorithm

1. Introduction

Wilson's disease is a rare genetic disorder of copper metabolism that leads to copper accumulation in several organs [1]. Since the disease primarily affects the liver and brain, it causes various neurological, hepatic, and psychological symptoms. These symptoms can be stabilized and controlled if the disease is diagnosed early and treated properly. If left untreated, Wilson's disease can be fatal. The disorder is caused by mutations in the ATP7B gene, which is responsible for copper transport. Affected individuals usually carry compound heterozygous mutations, meaning they inherited two different mutations, or homozygous mutations, where the same mutation is present in both copies of the gene. People who inherit only one mutated copy of the ATP7B gene are carriers; they usually do not show symptoms but can pass the mutation to their offspring. While genetic testing provides the most definitive diagnosis, hepatic involvement is nearly always present, making liver-specific biomarkers (such as serum albumin, liver enzymes, and bilirubin levels) and liver biopsy essential tools in clinical diagnosis. Additionally, copper levels in blood and 24-hour urine samples are important diagnostic indicators. However, interpreting serum copper levels can be challenging, as low ceruloplasmin concentrations may falsely lower total copper estimates.[2] One of the most distinctive clinical signs—besides the genetic mutation—is the presence of Kayser-Fleischer rings, which result from copper deposits in the cornea and are highly suggestive of neurological involvement. [3] Diagnosis of Wilson's disease can be challenging due to the variability of symptoms, which often depend on the patient's age. For example, younger individuals tend to present with more advanced hepatic involvement, while patients aged 15–35 more frequently exhibit neurological symptoms. Moreover, there is no single definitive diagnostic test, and the clinical presentation often overlaps with other hepatic or neurological disorders, leading to delayed or missed diagnoses.

Given these challenges, diagnostic support using soft classifier algorithms holds significant value. Machine learning techniques can detect subtle patterns in clinical and biochemical data that may not be immediately apparent, thereby enhancing early detection and supporting decision-making in complex

*IVUS 2025: Information Society and University Studies 2025, May 15, Kaunas, Lithuania

eb310420@student.polsl.pl (E. Bezrutczyk); wk310769@student.polsl.pl (W. Stefańska)

cases such as Wilson's disease. Based on medical systems analysis [4] and its different components like nature-inspired/heuristic algorithms applications [5, 6, 7, 8], we proposed a hybridization of the soft set [9] approach with a genetic algorithm (as a representative of the heuristic algorithms). More precisely, we present two separate classification tables, to determine the most effective diagnostic setup. In the following sections, we evaluate the clinical relevance of the extracted diagnostic features using a weighted soft set-based classification approach. With the use of a genetic algorithm applied over 10 generations to optimize feature weights and further improve classification performance. This evolutionary approach simulates natural selection to iteratively search for the most effective weight vector. Each individual in the population represents a candidate weight vector.

2. Methodology

2.1. Soft decision tables

To construct soft decision tables, we calculated representative values for each feature in each class, such as the mean and median, along with corresponding margins. We used two approaches. In the first approach, the soft decision table was built using the mean values of each feature as the center points and the standard deviations as the margins. In the second approach, we used the median as the center and the interquartile range (IQR) as the margin.

The interquartile range (IQR) measures the spread of the middle 50% of the data. For a given feature x_i in class c, it is defined as:

$$\text{IQR}_i^{(c)} = Q_{3,i}^{(c)} - Q_{1,i}^{(c)} \tag{1}$$

where $Q_{1,i}^{(c)}$ and $Q_{3,i}^{(c)}$ are the first (25th percentile) and third quartiles (75th percentile), respectively.

2.2. Soft Set Classifier

The proposed soft set classification method assigns membership scores to each class based on how well the test object's feature values fit within the statistical margins of each class, see Algorithm 1. For each class, the function initializes a membership score counter and iterates through every feature, comparing its value to the corresponding feature's mean and margin for that class. For each feature i, the center of the interval μ_i represents either the mean or the median of that feature in a given class, depending on the chosen soft decision table. The margin δ_i corresponds to either the standard deviation or the interquartile range (IQR), defining the width of the interval used to assess the feature match.

$$\text{Interval}_i = [\mu_i - \delta_i,\ \mu_i + \delta_i]$$

If the test feature value lies within this interval, it is considered a match. After evaluating all features, the total weighted match count represents the membership score of the test object for that class. This process is repeated for each class, and the function returns a dictionary containing membership scores for all classes.

Further on we implemented a function classify_with_method with uses the soft set classification approach, see Algorithm 2. It calculates statistics using soft decision tables (calculate_stats_func) from the training dataset. For each instance in the test set, it extracts feature values and computes membership scores. The predicted class is selected as the one with the highest membership score. Finally, the function calculates classification accuracy and returns both the classification results and overall accuracy.

Algorithm 1: Soft Set Membership Calculation

Input: $test\_obj$, $class\_stats$, $weights$
Output: $memberships$

$memberships =$ empty dictionary;
foreach $(class\_label, (means, margins))$ *in* $class\_stats$ **do**
  $match\_count = 0$;
  $total\_features =$ length of $means$;
  for $i = 0$ **to** $total\_features - 1$ **do**
    $feature\_val = test\_obj[i]$;
    $mean\_val = means[i]$;
    $margin = margins[i]$;
    if $mean\_val - margin \leq feature\_val \leq mean\_val + margin$ **then**
      $match\_count = match\_count + weights[i]$;
    end
  end
  $memberships[class\_label] = match\_count$;
end
return $memberships$;

Algorithm 2: Soft Set Classification Algorithm

Input: $train\_df$, $test\_df$, $calculate\_stats\_func$, $weights$
Output: $classification\_table$, $accuracy$

$class\_stats = calculate\_stats\_func(train\_df)$;
$correct = 0$;
$total =$ number of rows in $test\_df$;
$classification\_table =$ empty list;
$feature\_columns =$ columns in $test\_df$ excluding label column;
foreach *row in* $test\_df$ **do**
  $true\_class =$ value of label column in row;
  $test\_data =$ feature values in row;
  $memberships = soft\_set\_membership(test\_data, class\_stats, weights)$;
  $predicted\_class =$ class with highest membership value;
  if $predicted\_class == true\_class$ **then**
    $correct = correct + 1$;
  end
  Append $\{true\_class, predicted\_class, memberships\}$ to $classification\_table$;
end
$accuracy = \frac{correct}{total}$;
return $classification\_table$, $accuracy$;

2.3. Genetic algorithm for calculating weights

To enhance classification performance, we employed a genetic algorithm (GA) inspired by the approach described in Simulation of the characteristics of the ball movement on a beam by the use of genetic algorithm [10]. The GA evolves a population of candidate weight vectors over multiple generations to optimize feature weights for the soft set classifier.

Initially, the algorithm generates a random population of weight vectors normalized to sum to one. In each generation, it evaluates the classification accuracy of every individual weight vector using the soft set classifier on the training and testing datasets. The average accuracy of the population is computed, and the best-performing individual is recorded.

Selection of individuals for reproduction is performed using roulette wheel selection, which proba-

bilistically favors higher accuracy. New offspring are generated by blending pairs of parents through a linear crossover operation, followed by mutation, which introduces small random perturbations to maintain diversity.

This iterative process continues for a fixed number of generations, with the population ideally converging towards weight configurations that have higher classification accuracy. Throughout, all weight vectors remain normalized to ensure their sums equal one.

Algorithm 3: Genetic Algorithm for Soft Set Weight Optimization

Input: $trainSet$, $testSet$, $generations$, pop_size, $mutation\_rate$
Output: $best\_weights$, $best\_accuracy$

$n\_features =$ number of features in $trainSet$ (excluding label);
Initialize $population$ with pop_size random normalized weight vectors;
$best\_weights =$ None;
$best\_accuracy = 0$;
for $t = 1$ **to** $generations$ **do**
    $scored\_population =$ empty list;
    foreach $weights$ *in* $population$ **do**
        $accuracy =$ Evaluate accuracy using soft set classifier with $weights$;
        Append $(weights, accuracy)$ to $scored\_population$;
    end
    Sort $scored\_population$ by accuracy descending;
    if *best accuracy in* $scored\_population > best\_accuracy$ **then**
        $best\_weights =$ best weights;
        $best\_accuracy =$ best accuracy;
    end
    $survivors =$ Roulette wheel selection from $scored\_population$;
    $new\_population =$ empty list;
    while $|new\_population| < pop\_size$ **do**
        Select two parents $parent1$, $parent2$ from $survivors$;
        $\eta =$ random value in $[0, 1]$;
        $child1 = parent1 + \eta \cdot (parent2 - parent1)$;
        $child2 = parent2 + \eta \cdot (parent1 - parent2)$;
        foreach $child$ *in* $\{child1, child2\}$ **do**
            Clip $child$ to non-negative values;
            Normalize $child$ so that weights sum to 1;
            if *random()* $< mutation\_rate$ **then**
                Apply small random mutation to $child$;
                Clip and renormalize $child$;
            end
            Append $child$ to $new\_population$;
        end
    end
    $population = new\_population$;
end
return $best\_weights$, $best\_accuracy$;

3. Experiments

In our project, we used the Wilson Disease Synthetic Dataset for Data Analysis [1], which was made publicly available on the Kaggle platform by Guldanika Osmonova. The dataset consists of 60,000 records containing patients' information, including demographic variables (age, gender, region, and

socioeconomic status), biochemical indicators (e.g., ceruloplasmin levels, copper levels in blood and urine, liver enzymes), genetic markers, and ophthalmologic signs (e.g., Kayser–Fleischer rings). It also includes data on neurological and psychiatric symptoms.The final column of the dataset indicates whether an individual is affected by Wilson's disease. This binary classification label is encoded numerically, with 0 representing healthy individuals and 1 representing patients diagnosed with Wilson's disease. The structure of the dataset description was inspired by Malaria Detection Using Advanced Deep Learning Architecture [11].

Prior to soft set classification, the data underwent preprocessing. Due to the presence of missing values in the dataset, these were imputed using the median. Subsequently, the data was normalized using min-max normalization to scale the features to the range [0, 1], which helps improve the accuracy of the classification algorithm. Finally, the dataset was split into training and testing sets in a 70:30 ratio. The soft set classifier was then tested with identical weights on each feature on two different soft tables, with the goal of achieving the highest classification accuracy:

- Mean-based Soft table: 97.64%
- Median-based Soft table: 99.47%

To further enhance the classification performance, a genetic algorithm was implemented to iteratively optimize feature weights. To ensure the robustness of the results, the algorithm was executed multiple times on the same dataset, and the three best-performing runs were selected for evaluation. This approach evaluates the contribution of each feature toward classification accuracy over multiple generations, allowing the model to fine-tune its parameters effectively. After 10 generations, the optimized model achieved improved best accuracies as follows:

- Mean-based Soft table: 98.37%
- Median-based Soft table: 99.97%

To visualize relationships among features, we analyzed the correlations between variables in the dataset. Figure 1 presents the correlation matrix. As observed, Copper in Blood Serum, Copper in Urine, Kayser–Fleischer Rings, and ATP7B Gene Mutation are strong indicators for diagnosing Wilson's disease. Conversely, Free Copper in Blood Serum, Albumin, ALP, INR, Psychiatric Symptoms, and BMI show low correlation with the disorder. Analysis of correlations between certain features provides insight into what to expect when classifying with weighted algorithms.

To further evaluate the significance of each feature, a genetic algorithm was employed to optimize feature weights and to visualize their contribution to the classification process. As shown in Figure 2, features such as Age, Kayser-Fleischer Rings, Free Copper in Blood Serum, Neurological Symptoms Score, and Region consistently receive high weights across different soft tables, indicating their strong relevance in the diagnosis of Wilson's disease.

Conversely, features like Sex, BMI, Alcohol Use, and Total Bilirubin exhibit relatively low weights, suggesting that they have a limited impact on the classification performance. Interestingly, features such as ATP7B Gene Mutation, Family History, and liver enzymes (ALT and AST) demonstrate moderate to high importance, indicating their potential diagnostic value and justifying further clinical attention.

Comparing the results of the correlation heatmap and the feature weights plot reveals a strong alignment between copper-related biomarkers, genetic indicators, and liver markers in diagnosing Wilson's disease. Several features exhibit high importance in both analyses. Interestingly, some features that showed little or no correlation with the diagnosis in the heatmap were still assigned high importance by the genetic algorithm weights, indicating their subtle but significant role in classification. However, when combining the genetic algorithm weights with the correlation matrix, it becomes evident that BMI and psychiatric symptoms have low relevance for the diagnosis.

4. Conclusion

The implemented soft classifier demonstrates strong accuracy even before optimizing feature weights with the genetic algorithm. By leveraging heuristics, the algorithm rapidly converges to near-optimal

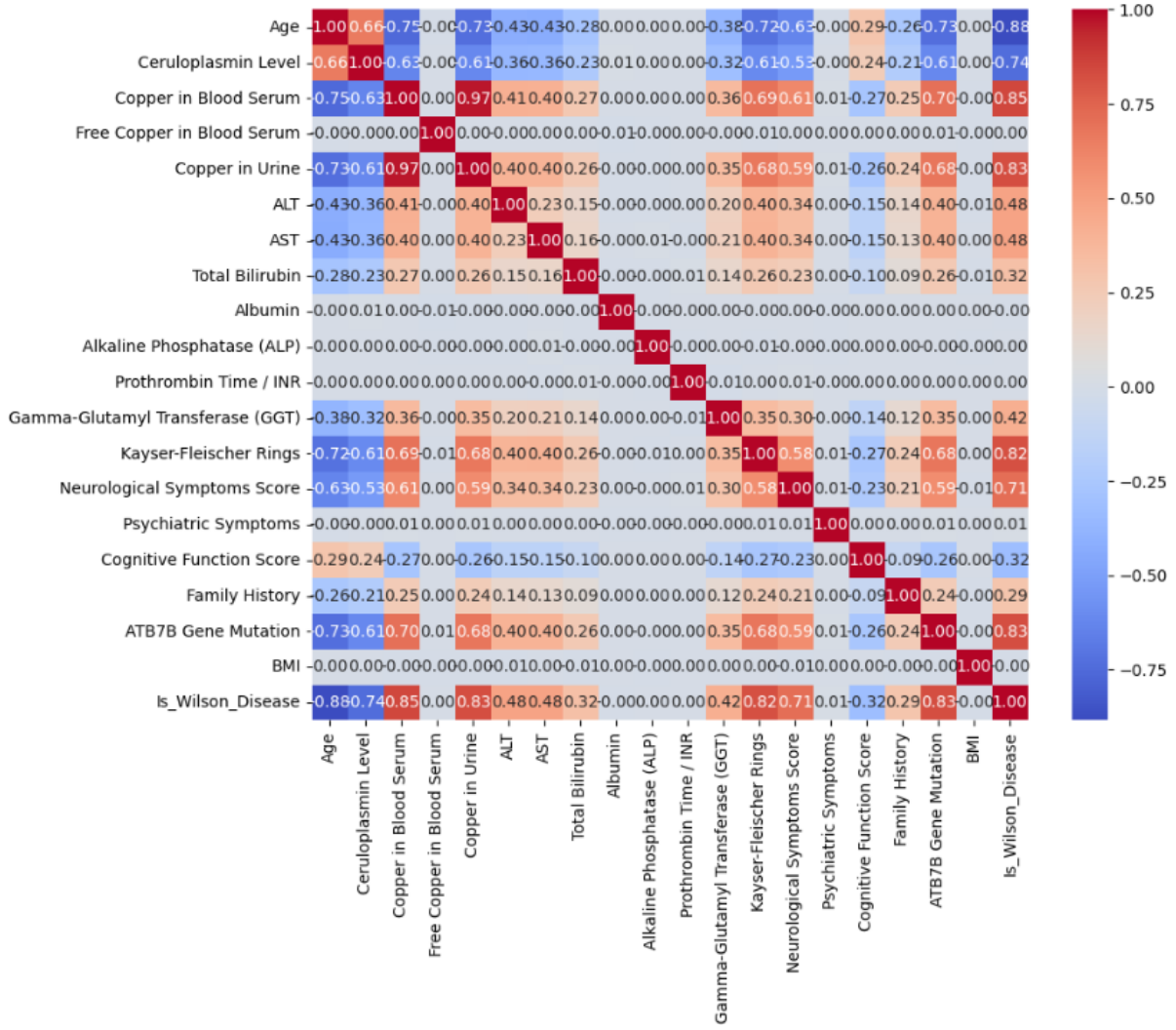

Figure 1: Correlation matrix for Wilson's disease dataset

weights within the early generations. This swift convergence to a local maximum likely occurs because the classifier already performs well on normalized data without explicit weighting. Beyond optimizing classification accuracy, the genetic algorithm also provides valuable insights into feature relevance. While certain features—such as Kayser-Fleischer rings and liver markers—play a dominant role in the classification, many others contribute less significantly. The weight analysis reveals underlying nonlinear relationships within the data. Consequently, this approach not only enhances the classifier's accuracy but also deepens our understanding of the diagnostic challenges associated with Wilson's disease. Overall, the classifier attains the upper limit of accuracy achievable using this soft set classification method.

The genetic algorithm experiments were conducted on approximately 30% of the full dataset due to computational limitations. Expanding the analysis to the entire dataset or using larger sample sizes could provide further validation of the robustness and generalizability of the proposed weighted soft set classifier. Future work may also consider developing hybrid models that integrate soft set theory with other machine learning techniques, potentially enhancing the model's ability to capture complex feature interactions effectively.

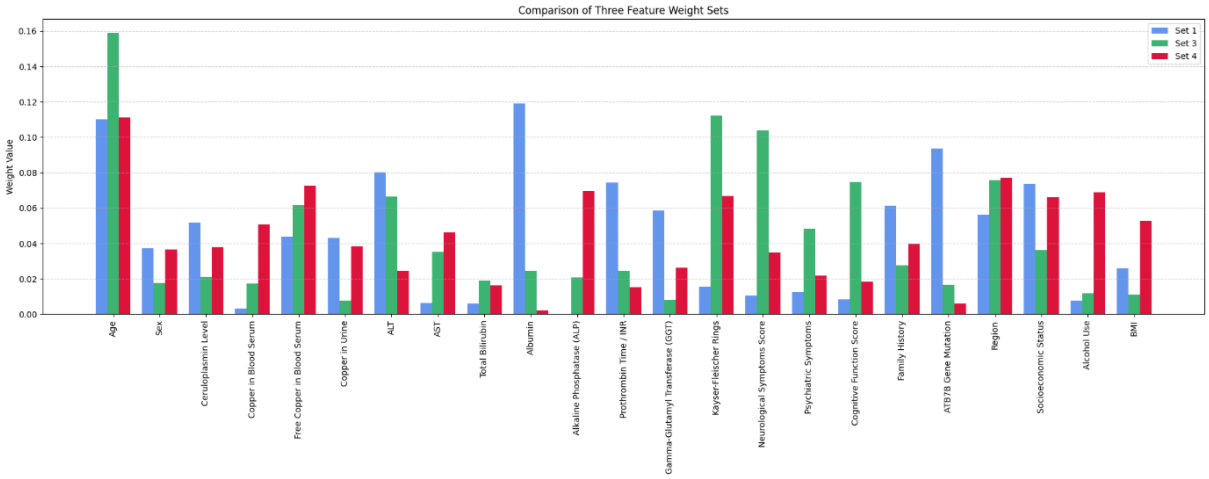

Figure 2: Feature weights optimized by the genetic algorithms

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] G. Osmonova, Wilson disease synthetic dataset for data analysis, https://www.kaggle.com/datasets/guldanikaosmonova/wilson-disease-dataset, 2024.

[2] A. Członkowska, T. Litwin, P. Dusek, P. Ferenci, S. Lutsenko, V. Medici, J. K. Rybakowski, K. H. Weiss, M. L. Schilsky, Wilson disease, Nature Reviews Disease Primers 4 (2018) 21. doi:10.1038/s41572-018-0018-3.

[3] H. S. Aboalam, M. K. Hassan, N. El-domiaty, N. F. Ibrahim, A. M. Ali, W. Hassan, E. G. Abu El Wafa, A. Elsaghier, H. F. Hetta, M. Elbadry, M. El-Kassas, Challenges and recent advances in diagnosing wilson disease, Journal of Clinical and Experimental Hepatology 15 (2025) 102531. doi:https://doi.org/10.1016/j.jceh.2025.102531.

[4] D. Połap, A. Jaszcz, Decentralized medical image classification system using dual-input cnn enhanced by spatial attention and heuristic support, Expert Systems with Applications 253 (2024) 124343.

[5] K. Prokop, D. Połap, Heuristic-based image stitching algorithm with automation of parameters for smart solutions, Expert Systems with Applications 241 (2024) 122792.

[6] W. Wang, W. Chen, L. Yan, Y. Yang, H. Zhao, Heuristic genetic algorithm parameter optimizer: Making lossless compression algorithms efficient and flexible, Expert Systems with Applications (2025) 126693.

[7] R. Brociek, M. Pleszczyński, A. Zielonka, A. Wajda, S. Coco, G. Lo Sciuto, C. Napoli, Application of heuristic algorithms in the tomography problem for pre-mining anomaly detection in coal seams, Sensors 22 (2022) 7297.

[8] A. Sikora, A. Zielonka, M. Woźniak, Heuristic optimization of 18-pulse rectifier system, in: 2021 IEEE Congress on Evolutionary Computation (CEC), IEEE, 2021, pp. 673–680.

[9] J. C. R. Alcantud, A. Z. Khameneh, G. Santos-García, M. Akram, A systematic literature review of soft set theory, Neural Computing and Applications 36 (2024) 8951–8975.

[10] M. Woźniak, M. Gabryel, R. K. Nowicki, Simulation of the characteristics of the ball movement on a beam by the use of genetic algorithm (2012).

[11] W. Siłka, M. Wieczorek, J. Siłka, M. Woźniak, Malaria detection using advanced deep learning architecture, Sensors 23 (2023). doi:10.3390/s23031501.