Ingeniare. Revista chilena de ingeniería, (2025) 33:38

Enhanced Dempster-Shafer classifier with metaheuristics and feature selection for predicting temporomandibular osteoarthritis progression

Clasificador Dempster-Shafer potenciado con metaheurísticas y selección de características para la predicción de la progresión de osteoartritis temporomandibular

Emilio Flores[1, 4*] https://orcid.org/0000-0002-6714-3597
Loreto Ramos-Rojas[2] https://orcid.org/0009-0001-8101-8944
Alejandro Veloz[3] https://orcid.org/0000-0001-9394-6660
Rodrigo Olivares[4] https://orcid.org/0000-0003-0582-954X

Recibido 21 de octubre de 2025, aceptado 29 de diciembre de 2025
Received: October 21, 2025 Accepted: December 29, 2025

ABSTRACT

Accurately predicting temporomandibular joint osteoarthritis (TMJ-OA) progression represents a significant clinical challenge. An enhanced version of the Dempster-Shafer classifier, optimized using the Reptile Search Algorithm (RSA), a metaheuristic optimization technique is introduced to minimize the Binary Cross-Entropy (BCE) loss function and frame TMJ-OA progression as a binary classification problem. A feature selection strategy using the Permutation Feature Importance (PFI) technique is implemented, guided by relevance scores determined by the Dempster-Shafer model. This methodology effectively reduced the dimensionality from 141 to 10 relevant features, derived from clinical and imaging data of 66 patients. The selected features were used to train eight classical machine learning models: Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machine, K-Nearest Neighbors, Gaussian Naive Bayes (GNB), Decision Tree, and AdaBoost. The optimized classifier achieved competitive performance (accuracy = 0.9095 ± 0.0744; ROC AUC = 0.9467 ± 0.0705; F1-score = 0.8590 ± 0.1280), comparable to Random Forest and GNB, while providing superior probabilistic calibration. The integration of metaheuristic optimization and model-specific feature selection enhances both the accuracy and interpretability of the system, offering an efficient and explainable alternative for predicting TMJ-OA progression.

Keywords: Temporomandibular joint osteoarthritis, Dempster-Shafer classifier, feature selection.

RESUMEN

La predicción temprana de la progresión de la osteoartritis temporomandibular (TMJ-OA) representa un desafío clínico significativo. En este trabajo se propone una versión mejorada del clasificador Dempster-Shafer, potenciado mediante el algoritmo metaheurístico Reptile Search Algorithm (RSA) para optimizar la función de Binary Cross-Entropy (BCE) y abordar la identificación de la progresión de TMJ-OA como un problema de clasificación binaria. Se implementa una estrategia de selección de características basada en la técnica de Permutation Feature Importance (PFI), guiada por la importancia atribuida por

[1] Universidad de las Américas. Facultad de Ingeniería y Negocios. Santiago, Chile. E-mail: eflores@udla.cl
[2] Universidad de los Andes. Facultad de Odontología. Santiago, Chile. E-mail: lcramos@miuandes.cl
[3] Universidad de Valparaíso. Escuela de Ingeniería Biomédica. Valparaíso, Chile. E-mail: alejandro.veloz@uv.cl
[4] Universidad de Valparaíso. Escuela de Ingeniería Informática. Valparaíso, Chile. E-mail: rodrigo.olivares@uv.cl
* Autor de correspondencia: eflores@udla.cl

DOI: 10.64966/ingeniare.v33.38.

el propio modelo Dempster-Shafer. Esta metodología permitió reducir eficazmente la dimensionalidad de 141 a 10 características relevantes, derivadas de datos clínicos e imagenológicos de un conjunto de 66 pacientes. Las características seleccionadas se emplearon para entrenar ocho modelos clásicos de aprendizaje automático: Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machine, K-Nearest Neighbors, Gaussian Naive Bayes (GNB), Decision Tree, y AdaBoost. El clasificador optimizado mostró un rendimiento competitivo (accuracy = 0,9095 ± 0,0744; ROC AUC = 0,9467 ± 0,0705; F1-score = 0,8590 ± 0,1280), comparable al de Random Forest y GNB, pero con una calibración probabilística superior. La integración de metaheurísticas y una selección de características específica del modelo mejora la precisión y la interpretabilidad del sistema, proponiendo una alternativa eficiente y explicable para la predicción de la progresión de TMJ-OA.

Palabras clave: Osteoartritis temporomandibular, clasificador Dempster-Shafer, selección de características.

INTRODUCTION

The quality of life for individuals is closely linked to the proper functioning of the stomatognathic system, in which the TMJ plays a fundamental role. This joint participates in essential functions such as mastication, speech, and deglutition, being crucial for an individual's physical and social well-being. Among the pathologies that can compromise its integrity, temporomandibular joint osteoarthritis (TMJ-OA) stands out due to its prevalence and functional impact. Additionally, early and timely diagnosis is of utmost importance for correct planning and treatment.

TMJ-OA is a degenerative joint disease characterized by the progressive deterioration of intra-articular tissues, including cartilage and the synovial membrane, accompanied by structural changes in the osseous components, such as the mandibular condyle and articular eminence [1]. It lacks a single, clearly established etiology and is therefore considered a multifactorial condition. The etiology of TMJ-OA remains unclear, and the condition is therefore considered multifactorial. Risk factors associated with its onset and progression include age, female sex, a history of direct or indirect trauma to the joint, and joint overloading resulting from parafunctional habits. Additional recognized factors include alterations in synovial lubrication, generalized joint hyperlaxity, genetic predisposition, hormonal imbalances, and anatomical and biomechanical abnormalities inherent to the TMJ are recognized. These factors can act independently or synergistically, fostering an environment conducive to the progressive deterioration of joint structures [2].

TMJ-OA has been the subject of considerable investigation in the scientific literature, particularly regarding its early detection, progression analysis, and support for clinical decision-making through artificial intelligence (AI) tools [3], [4]. In this context, [5] proposes TMJOAI, a web-based tool for the early prediction of TMJ-OA diagnosis. This tool employs multiple machine learning algorithms for feature selection, based on p-values and the Area Under the Receiver Operating Characteristic (ROC) curve (AUC). Among the evaluated algorithms, XGBoost and LightGBM achieved the highest performance results, with AUC values ranging from 0.83 to 0.88. Similarly, [6] reports promising results for the early detection of TMJ-OA by combining XGBoost and LightGBM models, achieving an accuracy of 0.82, an AUC of 0.87, and an F1-score of 0.81. Furthermore, [7] presents a methodology for predicting disease progression over a 2 to 3 years of clinical follow-up period, incorporating explainability through SHAP (SHapley Additive exPlanations). This technique allows for the interpretation of the relevance of features selected by the models, obtaining an accuracy of 0.87, an AUC of 0.72, and an F1-score of 0.82.

Furthermore, diagnostic imaging combined with deep neural networks has been extensively explored in the field of automated TMJ-OA diagnosis, primarily utilizing radiological images and computed tomography scans [8]. In [9] investigates the application of deep learning to two-dimensional panoramic images of the mandibular condyle, employing pre-trained models such as ResNet-152 and EfficientNet-B7, which achieved accuracy levels of 0.87 and 0.88, respectively. Likewise, reference [10] utilizes a version of the ResNet model developed in Keras to classify orthopantomography images into three categories: normal, indeterminate, and TMJ-OA, achieving an accuracy of 0.78. While these studies show promising results, they primarily focus on predictive performance, often lacking the granular interpretability required for clinical justification provided by evidential reasoning.

This research aims to predict the progression of TMJ-OA through a binary classification approach, utilizing the dataset available in [7]. An improvement to the classifier based on Dempster-Shafer theory described in [11], by incorporating RSA [12] as an optimization mechanism, is proposed to achieve this objective. RSA was selected for its remarkable balance between exploration and exploitation capabilities, a critical attribute for efficiently navigating the intricate, high-dimensional belief-mass assignment space inherent to Dempster-Shafer theory. This balance enables RSA to pinpoint optimal solutions with a notably faster convergence rate compared to other metaheuristics. In this study, RSA's core function is to meticulously determine the optimal values for these belief masses, which are the fundamental parameters that directly govern the Dempster-Shafer classifier's inference and decision-making process. The proposed improvements also include integrating rules with equidistant granularity and user-defined divisions, along with the use of a BCE-based loss function. As part of this enhancement process, a dimensionality-reduction methodology is introduced based on the PFI technique. PFI was chosen for its model-agnostic nature and its ability to provide interpretable insights into feature relevance, directly aligning with the explainability goals of the Dempster-Shafer framework. This allows for the selection of the most relevant features, guided by the behavior of the Dempster-Shafer classifier itself, ensuring that the feature set contributes meaningfully and transparently to the diagnostic prediction.

The main contributions of this article are as follows:

- An improved version of the Dempster-Shafer classifier, optimized using a metaheuristic algorithm (RSA) to enhance model fitting based on the BCE loss function, is proposed.
- A feature selection strategy based on the PFI technique, which allows for the selection of relevant attributes guided by the model's own behavior, contributing to an explainable dimensionality reduction, is integrated.
- The system's explanatory capability and diagnostic accuracy, offering an efficient, robust, and easily interpretable alternative for clinical support in the early diagnosis of TMJ-OA, are highlighted.

MATERIALS AND METHODS

This study presents experiments conducted using an open-access dataset to predict TMJ-OA progression [7]. The prediction was carried out using a Dempster-Shafer classifier, enhanced in this work by integrating the RSA metaheuristic. RSA serves as the core optimization mechanism, diligently searching the complex belief mass assignment space to determine the optimal belief masses that govern the classifier's inference process, specifically by minimizing the BCE loss function.

A two-phase methodology was implemented to ensure robust dimensionality reduction and strictly prevent data leakage. First, a stability analysis was performed using PFI integrated within a 10-fold cross-validation loop. In this phase, feature selection was performed exclusively on the training data for each fold, creating a voting system to identify a stable subset of clinically relevant features. Second, the identified stable features were used to train and assess the enhanced Dempster-Shafer classifier and compare it with eight classic machine learning models: Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machine, K-Nearest Neighbors, GNB, Decision Tree, and AdaBoost. All models were assessed in an independent cross-validation process using only the selected stable features. The methodology employed for the proposed classifier is illustrated in Figure 1.

Dataset

The dataset used in this study comprises clinical information and image-derived features of the condyle and articular fossa [7]. The clinical data includes individual patient characteristics such as gender and age, as well as results from a Likert-scale questionnaire based on the Diagnostic Criteria for Temporomandibular Disorders (DC/TMD) protocol, which focuses on TMJ-related issues. These considerations include symptoms like

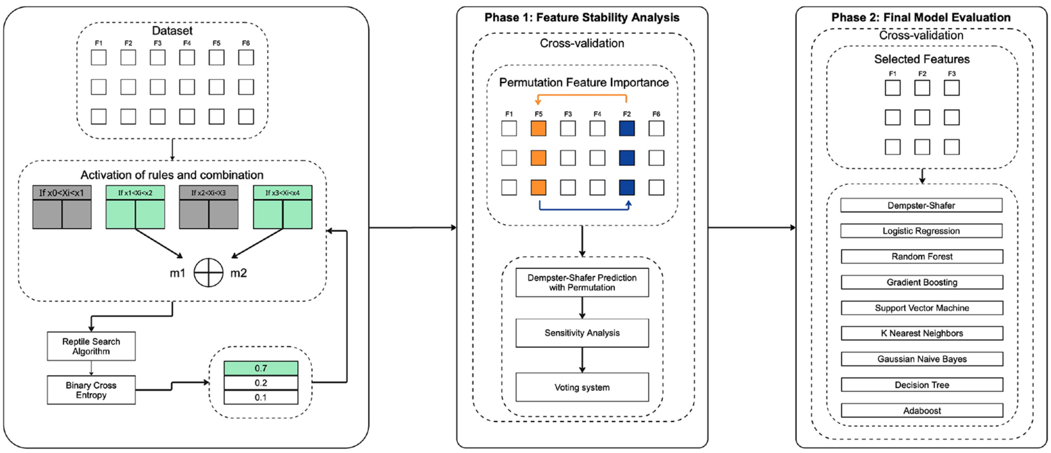

Figure 1. Methodological diagram for the development of the proposed classifier.

headaches, clicking sounds during mastication, as well as other functional and pain parameters (e.g., pain intensity, interference with daily activities, among others). In total, 106 patients participated; however, only 66 patients were ultimately validated through the meticulous fusion of available datasets. This validation process was crucial to ensuring data consistency, completeness, and quality across the merged sources, thereby enabling the selection of a robust, reliable subset for the study. The sample consists of 55 females and 11 males, with an average age of 40.21 ± 12.49 years. It is important to note the sensitive nature of data collected via Likert-scale questionnaires, given their inherent subjectivity and susceptibility to biases. These can include central tendency bias (avoiding extreme responses), acquiescence bias (tendency to agree), and social desirability bias (responding in a socially favorable way), which may potentially distort the true underlying information or introduce errors in the statistical analysis [13]. After data fusion and exclusion of target variables and baseline health information columns from each dataset, a total of 148 raw features were initially obtained. However, this set underwent a strict cleaning process described in the following section. The target variable is binary, indicating the progression of the patient's health status, where 0 corresponds to healthy and 1 corresponds to TMJ-OA. The distribution comprises 43 healthy patients and 23 patients with TMJ-OA. Figure 2 illustrates the distribution of patient health status during follow-up by gender.

Preprocessing

Data preprocessing focused on handling missing values through eliminating features rather than imputing them. Specifically, any column from the Likert-scale questionnaire that did not contain 100% complete responses (i.e., contained NaN values) was removed. This strict criterion was adopted because imputing values in subjective psychometric scales introduces a high risk of information distortion and bias [13]. As a result, 141 features were retained for initial model construction, preserving 95.27% of the original data structure.

Conversely, regarding data transformation, the proposed Dempster-Shafer classifier enhanced by RSA does not require feature scaling or standardization. This is because the classifier uses rule-based activation thresholds. It does not internally weight the magnitude of dataset values during training or inference.

Dempster-Shafer classifier

The Dempster-Shafer Theory of Evidence (DST) [14], also known as the Dempster-Shafer Theory, generalizes Bayesian theory by using "mass" assignments to multiple outcomes to measure the degree of uncertainty in the process.

The interpretable Dempster-Shafer classifier [11] is a DST-based model that uses rules and mass assignments to infer a result based on probabilities and uncertainty. In the Dempster-Shafer classifier [11], mass assignment is performed using stochastic

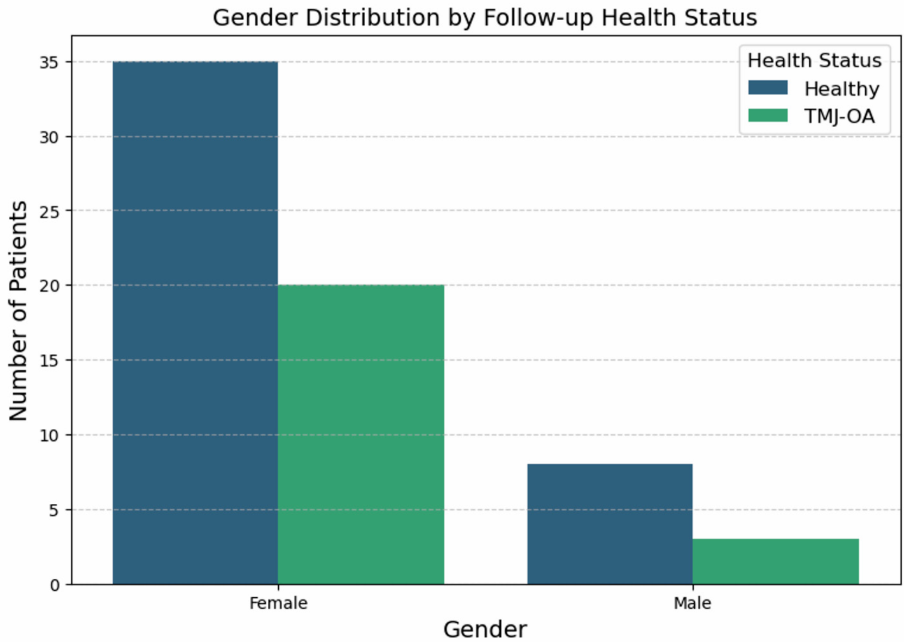

Figure 2. Distribution of health status at follow-up by gender.

gradient descent (SGD) and a mean squared error (MSE) loss function. This section briefly details how the classifier works.

Let X be the set of all states of a system, called the frame of discernment. The mass assignment must satisfy the conditions presented in equation (1).

$$m:2^{x} \rightarrow [0,1],\ m(\varnothing)=0,\ \Sigma_{A\subseteq X} m(A)=1 \tag{1}$$

Where A is a subset of X and Ø is an empty set. The term $m(A)$ represents the probability of obtaining exactly the result of A and not of a proper subset of A.

The Belief metric, or credence, is presented as the total evidence supporting the result, and is given by the expression in equation (2).

$$Bel_{m}(A)=\Sigma_{B\subseteq A}(B) \tag{2}$$

The plausibility metric, or plausibility, is defined as the total amount of evidence that could support a given set. It is defined as shown in equation (3) below.

$$Pl_{m}(A)=\Sigma_{B\cap A=A=\varnothing} m(B) \tag{3}$$

The combination of masses can be performed in various ways; in this case, Dempster's Rule of Combination (DR) is presented in equation (4).

$$m_1(A)\oplus m_2(B)=\frac{1}{1-K}\Sigma_{B\cap C=A\neq\varnothing} m_1(B)m_2(C) \tag{4}$$

Here, K represents a constant indicating the degree of conflict between $m1$ and $m2$. It also normalizes the values of Dempster's combination, as given by equation (5).

$$K=\Sigma_{B\cap C=\varnothing} m_1(B)m_2(C) \tag{5}$$

Rule generation is accomplished using statistical granular separators, such as the mean, quartiles, and variances, as illustrated in Figure 3.

Figure 3 illustrates an example where μ corresponds to the mean and σ to the standard deviation. Some example intervals might be the following:

If $x_i < \mu - \sigma$, then
If $\mu - \sigma \leq x_i < \mu + \sigma$, then
If $x_i \geq \mu + \sigma$, then

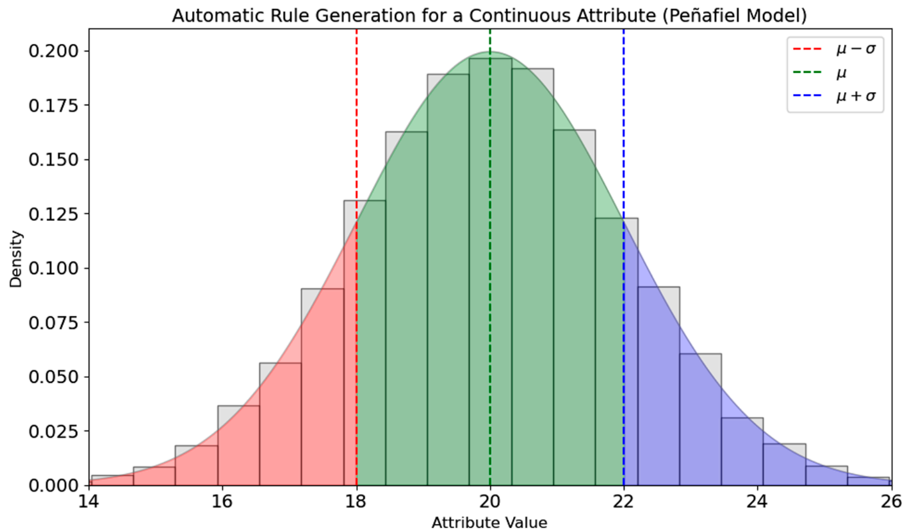

Figure 3. Example of interval generation for the Dempster-Shafer classifier.

Dempster-Shafer classifier enhanced with metaheuristic algorithms

This work, building upon the framework of the interpretable Dempster-Shafer classifier [11], proposes a substantial modification to the optimization process of belief masses. This is achieved by employing metaheuristic algorithms and the BCE loss function for learning the rules.

In the proposed Dempster-Shafer classifier, a set of rules is used to account for the distribution of each class. Equidistant divisions (M) are applied within these distributions, meaning the user can determine how many divisions (M+1) the algorithm performs to activate each rule. A representation of the equidistant distribution division for rule generation is illustrated in Figure 4.

Some example intervals could be the following:

If $13.5 \leq x_i < 16.5$ then
If $16.5 \leq x_i < 20$ then
If $20 \leq x_i < 23.25$ then
If $23.25 \leq x_i < 26.5$ then

Here, the symbol x_i corresponds to the feature evaluated within the interval and rule activation.

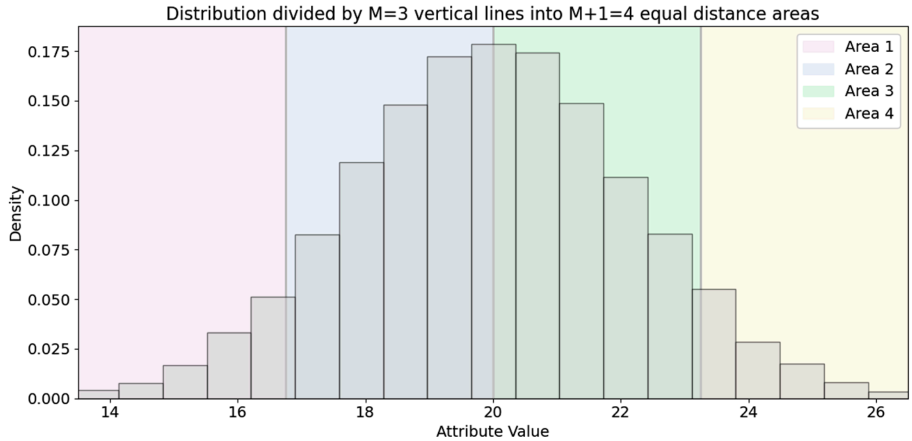

Figure 4. Example of interval generation for the proposed Dempster-Shafer classifier.

An example rule is as follows: "If x_i is between 13.5 and 16.5, then m (*healthy*) = 0.75 , m (*TMJ – OA*) = 0.05, and m (Ø) = 0.20." This implies that each interval will have an associated rule, and for each rule, three masses, corresponding to the healthy class, the TMJ-OA class, and uncertainty, respectively.

Another modification made to the Dempster-Shafer classifier involves using the BCE loss function, as presented in equation (6).

$$BCE = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log\left(p_i\right) + \left(1 - y_i\right)\log\left(1 - p_i\right)\right] \quad (6)$$

Here, N is the number of observations, y_i is the true label for the i-th observation, and p_i is the predicted probability for the i-th observation belonging to class 1.

The RSA [12], a powerful metaheuristic, is employed in this study for its demonstrated efficacy across diverse applications [15]-[18]. Within our framework, the RSA algorithm's primary role is to optimize the combined belief masses derived from the Dempster-Shafer theory. Specifically, the raw values generated by the RSA are passed through an L1 normalization layer to ensure they strictly satisfy the mass summation constraint presented in Equation (1). This optimization is driven by the imperative to minimize the BCE loss function, which quantifies the discrepancy between the classifier's probabilistic outputs and the true labels, effectively guiding the model towards more accurate predictions. The optimization process is intricately linked to the rule activation within the Dempster-Shafer framework, where each rule contributes to the overall belief assignment for a given input.

It is crucial to acknowledge that, due to the inherent stochastic nature of metaheuristic algorithms like RSA, the initial starting points (initialization) can significantly influence the optimization trajectory. Consequently, different runs of the algorithm may yield diverse optimized belief mass values. This variability, in turn, can lead to slight differences in the resulting accuracy, highlighting the importance of robust evaluation and potentially multiple runs for reliable performance assessment.

Feature Selection via Permutation

PFI was employed to enhance the interpretability of the Dempster-Shafer classifier and strictly prevent data leakage. Unlike standard approaches that apply feature selection on the entire dataset, a stability analysis framework integrated within a stratified 10-fold cross-validation was implemented.

In each fold of the cross-validation, the dataset was split into training (90%) and validation (10%) sets. Crucially, the PFI calculation and feature ranking were performed exclusively using the training data of the current fold. For each iteration, a Dempster-Shafer classifier was trained, and the baseline accuracy was established on the training subset. The importance of a feature was quantified by randomly shuffling its values within the training set in order to break its correlation with the target variable. The subsequent drop in accuracy served as the metric for that feature's contribution.

Within each fold, the top 15 features causing the most significant performance drop were identified. This process was repeated 10 times (once per fold), creating a voting system. From this aggregation, only features present in at least 40% of the folds (i.e., 4 out of 10) were selected for the final model, resulting in a stable set of 10 features.

Table 1 presents the meticulously configured parameters chosen for the Dempster-Shafer classifier to perform feature selection, ensuring an optimized and robust process. Future research can further refine this procedure to strengthen the steps and enhance the interpretability of the Dempster-Shafer classifier model.

Training and validation

Following the feature stability analysis, the final evaluation methodology proceeds with a second, independent Stratified 10-Fold Cross-Validation

Table 1. Parameters for feature selection in the Dempster-Shafer classifier.

Classifier	Parameters
Dempster-Shafer	M = 3, optimizer = RSA, n_iterations = 1000, n_agents = 30, random_state = 40

process. In this phase, the proposed Dempster-Shafer classifier and the eight benchmark models are trained and validated using exclusively the subset of features that satisfied the stability criterion (i.e., those selected in at least 40% of the folds during the screening phase). For the benchmark models, standard default hyperparameters from the Scikit-learn library were used, unless otherwise specified in Table 2.

This rigorous approach ensures a fair and consistent comparison by training all algorithms on identical data splits across all 10 partitions. Crucially, by restricting the input to the pre-validated stable subset, this methodology decouples the feature selection process from this final performance evaluation, effectively mitigating data leakage and ensuring the validity of the model's generalization assessment. Table 2 details the specific parameters configured for each algorithm.

Performance evaluation metrics

A set of standard performance metrics for binary classification problems was used to assess the proposed Dempster-Shafer classifier and the baseline models' performance in predicting the progression of temporomandibular joint osteoarthritis. These metrics were calculated from the cross-validation results and are presented as the mean along with the standard deviation. The metrics are computed from *TP* (true positives), *TN* (true negatives), *FP* (false positives), *FN* (false negatives), and *N* (total number of instances).

Accuracy: This represents the proportion of correct predictions out of the total number of instances. Its formulation is presented in equation (7).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

Precision: Measures the proportion of positive predictions that were actually correct. Its formulation is presented in equation (8).

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

Recall (Sensitivity or Completeness): This measures the proportion of actual positive instances that were correctly identified by the model. Its formulation is presented in equation (9).

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

F1-Score: This is the harmonic mean of Precision and Recall. It provides a balanced metric of a classifier's performance, being particularly useful in datasets with imbalanced classes, as it penalizes errors in both precision and recall. Its formulation is presented in equation (10).

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{10}$$

ROC AUC: This metric measures the model's discriminative capacity to distinguish between positive and negative classes, regardless of the chosen classification threshold. An AUC value closer to 1.0 indicates a higher discriminative ability.

Log Loss (Logarithmic Loss): This quantifies the accuracy of a model's probability predictions. It

Table 2. Parameters of classification algorithms.

Classifier	Parameters
Dempster-Shafer	M = 3, optimizer = RSA, n_iterations = 1000, n_agents = 30, random_state = 40
Logistic Regression	max_iter = 1000, random_state = 40
Random Forest	n_estimators = 100, random_state = 40
Gradient Boosting	n_estimators = 100, random_state = 40
SVC	probability= True, random_state = 40
KNN	–
GNB	–
Decision Tree	random_state = 40
AdaBoost	n_estimators = 100, random_state = 40

heavily penalizes highly confident but incorrect predictions. A lower value indicates better-calibrated predicted probabilities. Log Loss is calculated using the BCE loss function from equation (6).

Cohen's Kappa (κ): This measures the degree of agreement between the classifier's predictions and the true labels, adjusting for the possibility of random agreement. It is more robust than Accuracy for imbalanced datasets. Its formulation is presented in equation (11)

$$K = \frac{P_o - P_e}{1 - P_e} \tag{11}$$

Where Po is the observed accuracy, and Pe is the probability of expected agreement by chance, calculated as presented in equation (12).

$$P_e \frac{(TP+FP)(TP+FN)+(FN+TN)(FP+TN)}{N^2} \tag{12}$$

Matthews Correlation Coefficient (MCC): This metric assesses the quality of binary classifications by considering true positives, true negatives, false positives, and false negatives. It generates a coefficient ranging from -1 (completely opposite prediction) to +1 (perfect prediction), with 0 indicating a random prediction. MCC is considered a balanced and robust metric for imbalanced datasets. Its formulation is presented in equation (13).

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{13}$$

Balanced Accuracy: This calculates the average of the Recall obtained for each class (positive and negative classes). It is a useful metric when classes are imbalanced, as it gives equal weight to each class's predictive capability. Its formulation is presented in equation (14).

$$Balanced\ Accuracy = \frac{Recall_{positive} + Recall_{negative}}{2} \tag{14}$$

EXPERIMENTAL RESULTS

This section presents the empirical results from evaluating the proposed Dempster-Shafer classifier and the benchmark models for predicting TMJ-OA progression. All models were trained and evaluated exclusively using the stable subset of features identified through the proposed stability-based PFI methodology to ensure a rigorous assessment. The performance evaluation was conducted using a stratified 10-fold cross-validation scheme to maintain class distribution consistency across all partitions.

Feature selection

The stability analysis, integrated within the stratified 10-fold cross-validation, evaluated the recurrence of relevant features selected by the PFI method. Figure 5 illustrates the selection frequency for the top features across the 10 folds. The vertical axis lists the features, while the horizontal axis represents the number of folds in which each feature was identified as one of the top contributors to the model's performance.

Based on the established stability criterion, the dimensionality of the dataset was effectively reduced from 141 initial attributes to a robust core of 10 features:

- ProblemOrPreventChewing: Difficulty or prevention of chewing.
- C_entropy_FU: Entropy of the condylar image texture at follow-up (Radiomic feature).
- JawGratingGrinding: Grating or grinding sounds in the jaw.
- PainLocation: Specific location of facial/jaw pain.
- ProblemOrPreventEatingHardFoods: Difficulty eating hard foods.
- ProblemOrPreventTalking: Difficulty or limitation when talking.
- GrindTeethClenchWhileSleeping: Teeth grinding or clenching during sleep.
- PainRight: Presence of pain on the right side.
- ProblemOrPreventSexualActivity: Limitations in sexual activity due to TMJ pain.
- FeelingWeak: Perception of general weakness.

Notably, while the stability analysis prioritized clinical variables derived from the DC/TMD protocol, the inclusion of C_entropy_FU is a significant finding. This indicates that the entropy (texture complexity) of the condyle in follow-up images carries a unique predictive signal that complements the patient-reported symptoms, which was captured by the Dempster-Shafer model's evidential reasoning process.

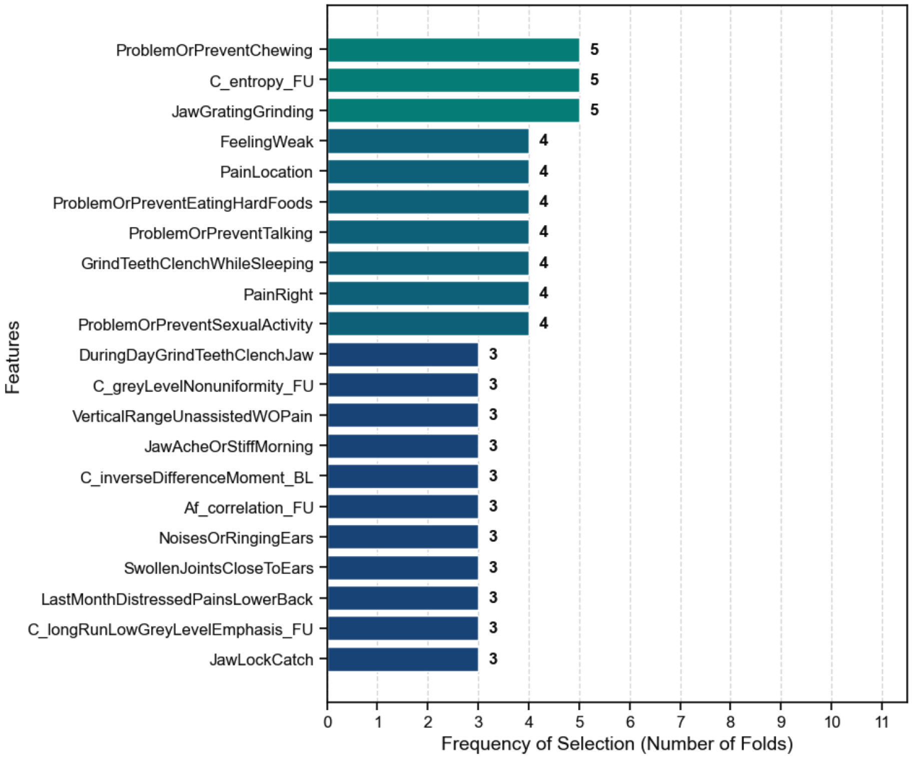

Figure 5. Feature stability analysis based on selection frequency across 10-fold cross-validation.

A sensitivity analysis was conducted by varying the stability threshold from 1 to 10 votes to validate the optimality of the selected subset. Figure 6 illustrates the trade-off between model parsimony and performance.

A clear trend is observed: as the stability threshold increases, the number of retained features drops drastically (from >80 at threshold 1 to 3 at threshold 5 and to 10 features at threshold 4, which represents the optimal trade-off point selected for this study), while the classification accuracy exhibits a consistent upward trend. Specifically, the model utilizing the minimal set of 10 highly stable features achieved the highest mean accuracy (~0.90) with reduced variance. This confirms that excluding unstable features effectively filters out noise, resulting in a more generalizable and parsimonious model.

The discriminatory power of the ten selected features is visually inspected in Figure 7. The histograms reveal distinct distributional patterns between "Healthy" and "TMJ-OA" patients. For instance, variables such as ProblemOrPreventEatingHardFoods and ProblemOrPreventChewing show a clear separation, where the presence of the symptom is strongly associated with the TMJ-OA class.

Stability analysis of the metaheuristic optimization

Given the stochastic nature of the RSA, evaluating the consistency of the optimization process is essential to ensure reliability. 30 independent execution runs were conducted using the selected stable feature subset on fixed data, varying only the random seeds to strictly assess the stability of the proposed Dempster-Shafer classifier. Figure 8 displays the boxplot distribution of the Accuracy and ROC AUC metrics obtained across these iterations.

The quantitative analysis reveals a mean Accuracy of 0.8095 ± 0.0621, with values ranging from 0.7143 to 0.9286. More notably, the ROC AUC metric demonstrates remarkable stability, achieving a mean of 0.9000 with a low standard deviation of ± 0.0417 (Range: 0.8222 – 0.9778). The compact interquartile range observed in the ROC AUC

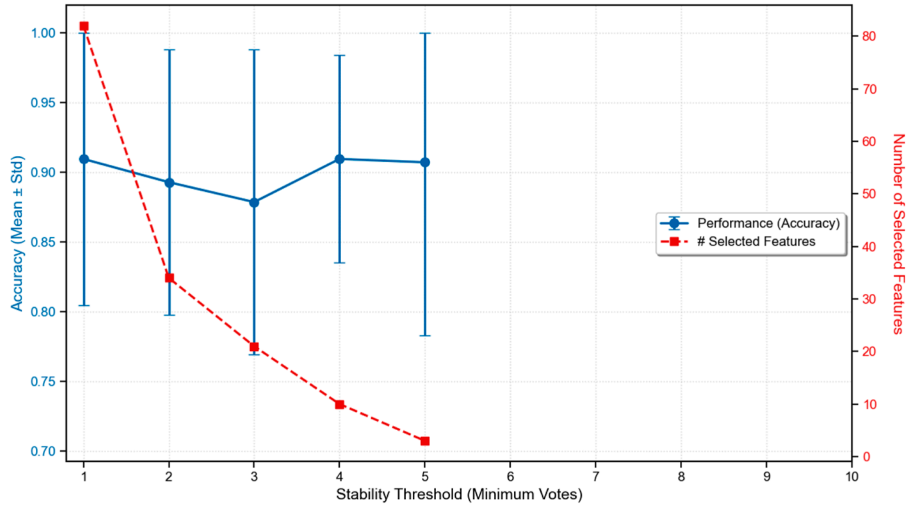

Figure 6. Sensitivity analysis showing the trade-off between Accuracy (blue line) and number of features (red dashed line) at different stability thresholds.

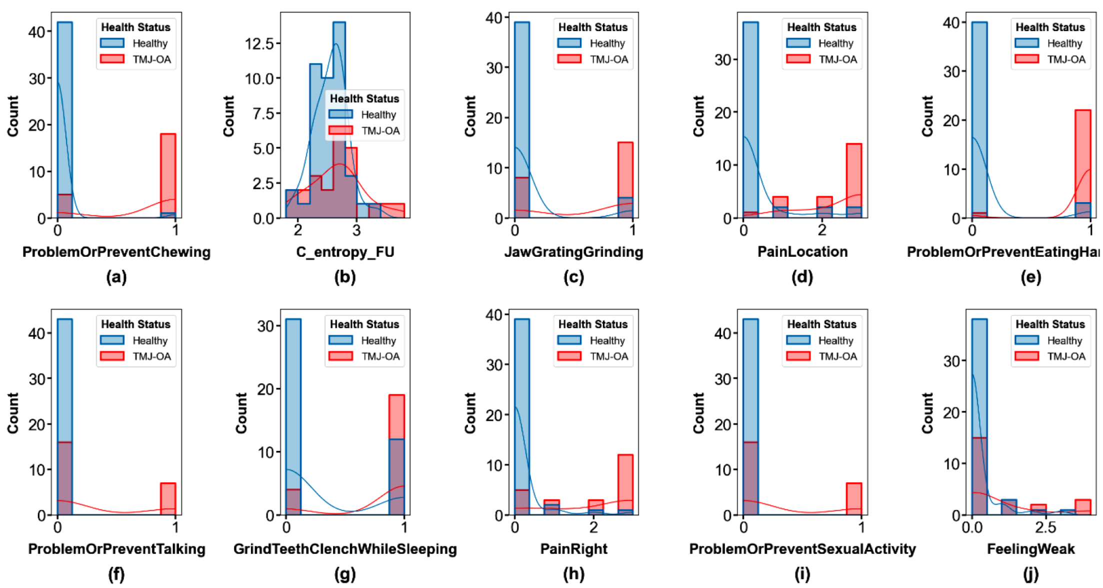

Figure 7. Distribution of the ten stable features for Healthy (blue) and TMJ-OA (red) groups: a) ProblemOrPreventChewing, b) C_entropy_FU, c) JawGratingGrinding, d) PainLocation, e) ProblemOrPreventEatingHardFoods, f) ProblemOrPreventTalking, g) GrindTeethClenchWhileSleeping, h) PainRight, i) ProblemOrPreventSexualActivity, and(j) FeelingWeak.

distribution confirms that the RSA consistently converges to near-optimal belief mass assignments, demonstrating that the model's discriminative capability is robust against random initialization.

Comparison with machine learning algorithms

Table 3 presents the consolidated performance metrics (mean ± standard deviation) for each of the nine classifiers evaluated using the stratified

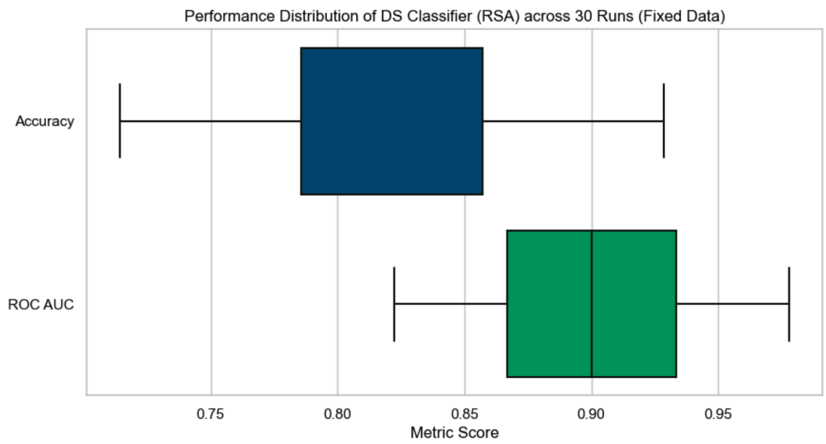

Figure 8. Performance distribution of the DS classifier optimized by RSA across 30 independent runs.

Table 3. Classifier Results with 10-Fold Cross-Validation.

Classifier	Accuracy	Precision	Recall	F1-score	ROC AUC	Log Loss	Cohen's Kappa	MCC	Balanced accuracy
Dempster-Shafer	0.9095 ± 0.0744	0.9083 ± 0.1417	0.8667 ± 0.2082	0.8590 ± 0.1280	0.9467 ± 0.0705	0.4825 ± 0.5477	0.7938 ± 0.1740	0.8196 ± 0.1513	0.8983 ± 0.0952
AdaBoost	0.8810 ± 0.1086	0.9000 ± 0.1528	0.8000 ± 0.2560	0.8100 ± 0.1739	0.9137 ± 0.0943	0.5702 ± 0.0581	0.7302 ± 0.2397	0.7603 ± 0.2156	0.8650 ± 0.1283
GNB	0.9381 ± 0.0762	0.9667 ± 0.1000	0.8500 ± 0.2291	0.8800 ± 0.1514	0.9650 ± 0.0776	1.5292 ± 2.2102	0.8427 ± 0.1955	0.8641 ± 0.1685	0.9150 ± 0.1119
Random Forest	0.9381 ± 0.0762	0.9333 ± 0.1333	0.9000 ± 0.2000	0.8933 ± 0.1373	0.9525 ± 0.0810	0.7627 ± 1.7573	0.8522 ± 0.1840	0.8715 ± 0.1595	0.9275 ± 0.0990
Logistic Reg.	0.9071 ± 0.1067	0.9167 ± 0.1708	0.8167 ± 0.2291	0.8433 ± 0.1752	0.9650 ± 0.0776	0.2511 ± 0.2339	0.7801 ± 0.2485	0.7989 ± 0.2375	0.8858 ± 0.1316
Gradient Boosting	0.9095 ± 0.0981	0.9000 ± 0.1528	0.8667 ± 0.2082	0.8600 ± 0.1474	0.9342 ± 0.0904	0.6554 ± 0.7372	0.7939 ± 0.2173	0.8132 ± 0.2026	0.8983 ± 0.1150
Decision Tree	0.8810 ± 0.1413	0.8833 ± 0.1833	0.8167 ± 0.2833	0.8167 ± 0.2156	0.8733 ± 0.1578	4.2909 ± 5.0916	0.7384 ± 0.3069	0.7633 ± 0.2895	0.8733 ± 0.1578
SVC	0.9381 ± 0.0762	0.9333 ± 0.1333	0.9000 ± 0.2000	0.8933 ± 0.1373	0.9625 ± 0.0800	0.2132 ± 0.1785	0.8522 ± 0.1840	0.8715 ± 0.1595	0.9275 ± 0.0990
KNN	0.9214 ± 0.1086	0.9167 ± 0.1708	0.8500 ± 0.2291	0.8633 ± 0.1804	0.9575 ± 0.0791	0.7148 ± 1.7703	0.8105 ± 0.2548	0.8258 ± 0.2434	0.9025 ± 0.1344

10-fold cross-validation scheme on the selected stable feature subset. Key metrics included are Accuracy, Precision, Recall, F1-Score, ROC AUC, Log Loss, Cohen's Kappa, Matthews Correlation Coefficient (MCC), and Balanced Accuracy.

Although algorithms such as Random Forest and Gaussian Naive Bayes (GNB) achieved higher nominal accuracy (0.9381), this metric requires careful interpretation. interpreted with caution. A critical analysis of the Log Loss metric reveals that GNB exhibits poor probabilistic calibration (Log Loss of 1.5292), indicating that the model is often "confidently wrong" in its predictions. In contrast, the Dempster-Shafer classifier maintains a significantly lower and stable Log Loss (0.4825), suggesting superior reliability in its probability estimates.

Furthermore, by retaining 10 features (a mix of key symptoms and texture biomarkers), the proposed model preserves a semantic layer. This allows clinicians to understand why a prediction is made based on specific evidence, such as the difficulty in chewing combined with changes in image entropy. This explainability, combined with a Balanced Accuracy of 0.8983, makes the DS classifier a more practical tool for clinical settings than opaque models with marginally higher but less reliable metrics.

Finally, Figure 9 visually summarizes the performance distribution of the top algorithms across the four most critical metrics (Accuracy, ROC AUC, F1-score, and Log Loss). The bar charts highlight the consistency of the Dempster-Shafer classifier (red bars), which shows contained dispersion and balanced results across all indicators, unlike competitors that

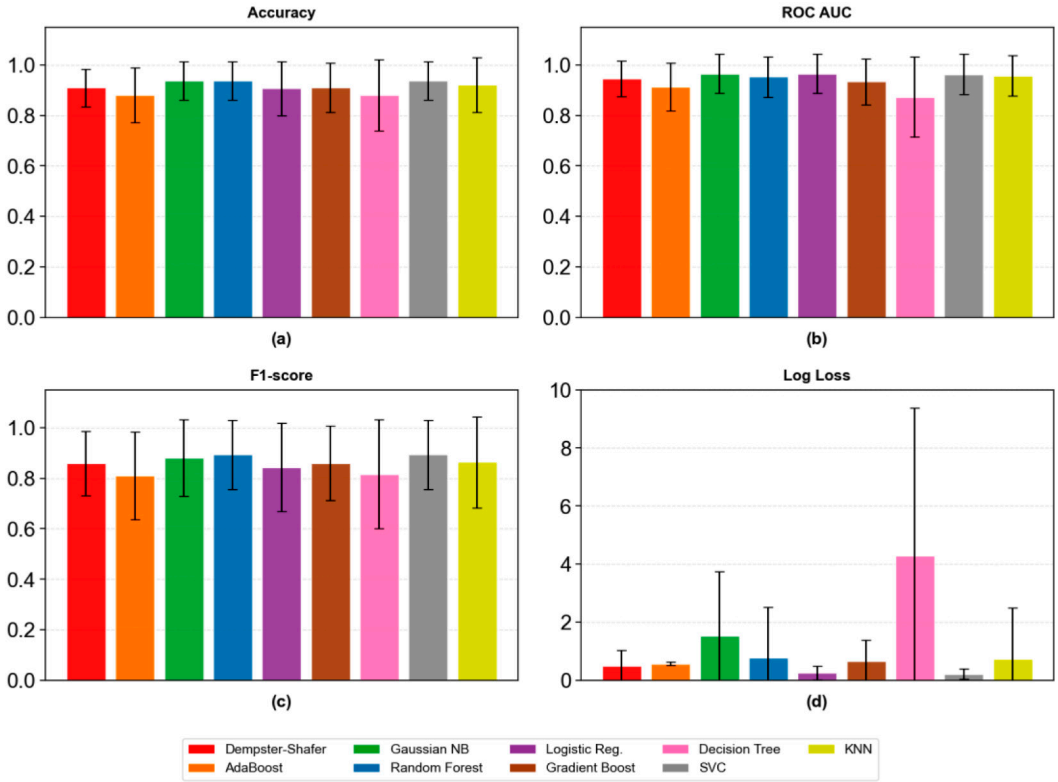

Figure 9. Comparative performance analysis of the classifiers across the four most critical metrics: a) Accuracy, b) ROC AUC, c) F1-score, and d) Log loss.

exhibit high variance or extreme trade-offs between precision and recall.

DISCUSSION

The results obtained in this study validate the efficacy of the proposed methodology, which integrates an enhanced Dempster-Shafer classifier (optimized via RSA) with a rigorous feature selection scheme. The model achieved robust performance metrics, registering a mean Accuracy of 0.9095 ± 0.0744 and a ROC AUC of 0.9467 ± 0.0705. Furthermore, the model demonstrated a solid balance between sensitivity and specificity, reflected in a Balanced Accuracy of 0.8983 ± 0.0952 and F1-score of 0.8590 ± 0.1280. These figures position the proposed method competitively against complex ensemble algorithms, outperforming Gradient Boosting and AdaBoost in terms of reliability and balanced performance.

A central finding of this research in the Dempster-Shafer classifier's capability to guide its own dimensionality reduction. The system effectively identified the intrinsic relevance of attributes using evidence theory by integrating the PFI technique into a stratified cross-validation cycle. The sensitivity analysis played a crucial role in this process, empirically determining the optimal stability threshold. As observed in the results, raising the selection criterion to a minimum of 4 votes effectively filtered out noise, reducing the feature space from 141 original variables to a robust core of 10 features.

This drastic reduction not only simplified the model but also revealed an interesting phenomenon: the transversal improvement of performance. The subset of features identified exclusively by the dynamics of the Dempster-Shafer classifier enabled classic models, such as Random Forest and Logistic Regression, to achieve high-level metrics. This suggests that the Dempster-Shafer classifier, thanks to its explicit handling of uncertainty and ignorance (mass assigned to Ø), possesses superior sensitivity for identifying truly informative biomarkers that other models might overlook. This finding opens a new avenue of research, proposing that the Dempster-Shafer classifier serves not only as a prediction tool but also as a highly effective wrapper-type dimensionality-reducing technique for complex medical problems, capable of enhancing the performance of other machine learning algorithms.

When comparing performance, it was observed that algorithms such as GNB and Random Forest achieved a marginally higher nominal accuracy (0.9381). However, this data must be interpreted with caution, with the trade-off between raw accuracy and clinical reliability in mind. The analysis of the Log Loss metric revealed that GNB exhibits poor probabilistic calibration (Log Loss of 1.5292), indicating a tendency to make incorrect predictions with high confidence. In contrast, the Dempster-Shafer classifier maintained a significantly lower Log Loss (0.4825 ± 0.5477), demonstrating a much safer, more calibrated uncertainty estimation, which is critical in healthcare applications.

The balance achieved between model parsimony and practical utility must be valued. Although retaining a larger number of variables might marginally increase numerical precision in some "black-box" models (such as deep learning approaches [9], [10]), the current proposal prioritizes interpretability. By retaining 10 features, composed predominantly of patient-reported symptoms but critically including a radiomic marker (e.g., ProblemOrPreventChewing, C_entropy_FU), the model preserves an essential semantic layer. This allows the clinician not only to obtain a binary prediction but also to understand the multimodal evidence supporting it, facilitating informed and personalized decision-making in the early diagnosis of TMJ-OA.

Finally, it is necessary to address the scale of this study. While the results are robust, the sample size ($N = 66$) limits broad generalization. Future studies should consider larger multicenter datasets.

CONCLUSIONS

The empirical results presented in this study demonstrate the efficacy of the proposed hybrid framework, which integrates the RSA-enhanced Dempster-Shafer classifier with a stability-based feature selection strategy. This rigorous methodological approach successfully reduced the high-dimensional clinical dataset from 141 attributes to a robust core of just 10 features, prioritizing a powerful combination of patient-reported symptoms and radiomic texture patterns. This significant reduction not only streamlined the model but also enhanced its interpretability without compromising predictive performance.

The Dempster-Shafer classifier achieved a robust accuracy of 0.9095 ± 0.0744 and a ROC AUC of 0.9467 ± 0.0705, proving to be highly competitive against established ensemble methods like Random Forest and AdaBoost. Notably, while some benchmark models achieved marginally higher nominal accuracy, the proposed classifier demonstrated superior probabilistic calibration, evidenced by a significantly lower Log Loss (0.4825 ± 0.5477). This balance between accuracy and reliability reinforces the model's suitability for clinical applications, where trustworthy uncertainty estimation is as critical as the classification label itself.

Furthermore, this study highlights the potential of the Dempster-Shafer classifier not only as a predictor but as a powerful wrapper for feature selection. The "stable feature subset" identified by the model's dynamics proved to be universally relevant, boosting the performance of other classifiers and confirming that the method captures genuine clinical and structural signals rather than model-specific noise.

Future research directions will focus on refining the evidential reasoning framework. Specifically, the plan is to investigate alternative combination rules, such as Yager's rule or Dubois and Prade's rule, to more effectively manage conflict redistribution in cases of high evidential disagreement. Additionally, it is aimed to explore novel dimensionality reduction techniques that directly exploit the classifier's internal mass assignments (Belief and Plausibility functions) to rank feature relevance, potentially offering a more computationally efficient alternative to iterative permutation methods. Expanding this framework to multiclass progression stages and integrating it into telemedicine platforms and Electronic Health Records (EHR) remains a priority. Furthermore, conducting pilot studies to validate the tool in real-world clinical settings is essential to democratize access to specialized TMJ-OA diagnostics.

DATA AND CODE AVAILABILITY

The source code developed for this study, including the metaheuristic optimization and the evidential reasoning framework, can be accessed at [19].

ACKNOWLEDGMENTS

This work was funded by the Escuela de Ingeniería Informática, Universidad de Valparaíso, Chile, through resolution REXE Nº 2552/2023. E. Flores was supported by the ANID BECAS/DOCTORADO NACIONAL 21242003. R. Olivares has been supported by FONDECYT de Iniciación 11231016 from ANID, Chile.

REFERENCES

[1] C.C. Peck *et al.*, "Expanding the taxonomy of the diagnostic criteria for temporomandibular disorders," *Journal of Oral Rehabilitation*, vol. 41, no. 1, pp. 2-23, 2014, doi: 10.1111/joor.12132.

[2] M. Sánchez y W. Becerra, "Osteoartritis (artrosis) de la articulación temporomandibular", *Revista de Otorrinolaringología y Cirugía de Cabeza y Cuello*, vol. 80, no. 4, pp. 540-553, 2020, doi: 10.4067/s0718-48162020000400540.

[3] O. Almăşan, D.C. Leucuţa, M. Hedeşiu, S. Mureşanu, and S.L. Popa, "Temporomandibular joint osteoarthritis diagnosis employing artificial intelligence: Systematic review and meta-analysis," *Journal the Clinical Medicine*, vol. 12, no. 3, Art. No. 942, 2023, doi: 10.3390/jcm12030942.

[4] J. Bianchi *et al.*, "Decision support systems in temporomandibular joint osteoarthritis: A review of data science and artificial intelligence applications," *Seminar in Orthodontics*, vol. 27, no. 2, pp. 78-86, Jun. 2021, doi: 10.1053/j.sodo.2021.05.004.

[5] C. Le *et al.*, "TMJOAI: An artificial web-based intelligence tool for early diagnosis of the temporomandibular joint osteoarthritis," in *Clinical Image-Based Procedures, Distributed and Collaborative Learning, Artificial Intelligence for Combating COVID-19 and Secure and Privacy-Preserving Machine Learning*, C. Oyarzún Laura *et al.*, Eds., vol. 12969 Cham, Switzerland: Springer, 2021, pp. 78-87, doi: 10.1007/978-3-030-90874-4_8.

[6] J. Bianchi *et al.*, "Osteoarthritis of the temporomandibular joint can be diagnosed earlier using biomarkers and machine learning," *Scientific Reports*, vol. 10, Art. no. 8012, 2020, doi: 10.1038/s41598-020-64942-0.

[7] N. Al Turkestani *et al.*, "A comprehensive patient-specific prediction model for temporomandibular joint osteoarthritis progression," *Proceedings of the National Academy of Sciences,* vol. 121, no. 8, Art. no. e2306132121, 2024, doi: 10.1073/pnas.2306132121.

[8] L. Xu, J. Chen, K. Qiu, F. Yang, and W. Wu, "Artificial intelligence for detecting temporomandibular joint osteoarthritis using radiographic image data: A systematic review and meta-analysis of diagnostic test accuracy," *PLoS ONE*, vol. 18, no. 7, Art. no. e0288631, 2023, doi: 10.1371/journal.pone.0288631.

[9] W. Jung, K.E. Lee, B.J. Suh, H. Seok, and D. W. Lee, "Deep learning for osteoarthritis classification in temporomandibular joint," *Oral Diseases*, vol. 29, no. 3, pp. 1050-1059, 2021, doi: 10.1111/odi.14056.

[10] E. Choi, D. Kim, J.Y. Lee, and H.K. Park, "Artificial intelligence in detecting temporomandibular joint osteoarthritis on orthopantomogram," *Scientific Reports*, vol. 11, no. 1, Art. no. 10246, 2021, doi: 10.1038/s41598-021-89742-y.

[11] S. Peñafiel, N. Baloian, H. Sanson, and J. Pino. "Applying Dempster-Shafer theory for developing a flexible, accurate and interpretable classifier," *Expert Systems with Applications*, vol. 148, Art. no. 113262, 2020, doi: 10.1016/j.eswa.2020.113262.

[12] L. Abualigah, M.A. Elaziz, P. Sumari, Z.W. Geem, and A.H. Gandomi, "Reptile search algorithm (RSA): A nature-inspired meta-heuristic optimizer," *Expert Systems with Applications*, vol. 191, Art. no. 116158, 2022, doi: 10.1016/j.eswa.2021.116158.

[13] J.C. Westland, "Information loss and bias in likert survey responses," PLoS One, vol. 17, no. 7, Art. no. e0271949, 2022, doi: 10.1371/journal.pone.0271949.

[14] G. Shafer, *A Mathematical Theory of Evidence*, 1st ed. Princeton, New Jersey, USA: Princeton University Press, 1976.

[15] B. Sasmal, A.G. Hussien, A. Das, K.G. Dhal, and R. Saha, "Reptile search algorithm: Theory, variants, applications, and performance evaluation," *Archives of Computational Methods in Engineering*, vol. 31, no. 1, pp. 521-549, 2024, doi: 10.1007/s11831-023-09990-1.

[16] A. Dahou *et al.*, "Intrusion detection system for IoT based on deep learning and modified reptile search algorithm," *Computational Intelligence and Neuroscience*, vol. 2022, Art. no. 6473507, pp. 1-15, 2022, doi: 10.1155/2022/6473507.

[17] Z. Elgamal, A.Q.M. Sabri, M. Tubishat, D. Tbaishat, S.N. Makhadmeh, and O.A. Alomari, "Improved reptile search optimization algorithm using chaotic map and simulated annealing for feature selection in medical field," *IEEE Access*, vol. 10, pp. 51428-51446, 2022, doi: 10.1109/access.2022.3174854.

[18] M.K. Khan, M.H. Zafar, S. Rashid, M. Mansoor, S.K.R. Moosavi, and F. Sanfilippo, "Improved reptile search optimization algorithm: application on regression and classification problems," *Applied Sciences*, vol. 13, no. 2, Art. no. 945, 2023, doi: 10.3390/app13020945.

[19] E. Flores Araya, L. Ramos-Rojas, A. Veloz, and R. Olivares, November 26, 2025, "Enhanced Dempster-Shafer classifier with metaheuristics optimization", distributed by Zenodo, doi:10.5281/zenodo.17718302.