Volume 3, Number 3, 2026 Pages 1890-1910

Bioscience Insights

ISSN: 3080-3489 (Print)
ISSN: 3080-3497 (Online)

BiosciInsights.com

DESIGN AND SIMULATION-BASED EVALUATION OF AN INTERPRETABLE FUZZY LOGIC FRAMEWORK FOR SLEEP APNEA SEVERITY CLASSIFICATION

Mariya Ilyas Khan[*1], Sana Makhdoom[2], Barkha Kumari[3]

[*1] *Institute of Biomedical Engineering (IBET), Liaquat University of Medical and Health Sciences (LUMHS),Pakistan*
[2] *Department of Biomedical Engineering, Air University, Islamabad, Pakistan*
[3] *Faculty of Engineering, Science, Technology and Management, Ziauddin University, Karachi (ZUFESTM), Pakistan*

[*1]mariyak417@gmail.com, [2]sanamakhdoom02@gmail.com, [3]barkhakumari0300@gmail.com

https://doi.org/10.5281/zenodo.24359350

Article Information

Article History

Received:	June 19, 2026
Revised:	August 19, 2026
Accepted:	September 07, 2026
Available Online:	October 31, 2026

Keywords:

Sleep apnea; fuzzy logic; Mamdani inference; interpretable machine learning; severity classification; pulse oximetry; simulation study.

Abstract

Obstructive sleep apnea (OSA) is graded by the apnea–hypopnea index (AHI), yet the grade a patient receives depends on a handful of hard cut-offs that sit awkwardly on top of noisy, partly overlapping physiological signals. Machine-learning classifiers deal with that noise but rarely show their reasoning, which makes them difficult to defend in a clinical setting. We designed a Mamdani fuzzy inference system (FIS) that grades OSA severity from four overnight oximetry and arousal measures, and we asked how much accuracy such a transparent system gives up, if any, relative to crisp threshold rules and standard machine-learning models. A synthetic cohort of 1,000 adults was generated from an explicit data-generating model in which the oxygen desaturation index (ODI), SpO_2 nadir, time below 90% saturation (T90) and arousal index depend on a latent AHI with realistic measurement noise. An expert-specified FIS (28 rules, trapezoidal membership functions) was refined by differential evolution on a 70% training split. Comparators were three threshold classifiers and five machine-learning models. Evaluation used a 30% hold-out set (n = 300) with bootstrap confidence intervals and exact McNemar tests, five-fold cross-validation, feature ablation, and membership-function sensitivity analysis. On the hold-out set the expert FIS reached 70.3% accuracy (95% CI 65.0–75.3) and the refined FIS 80.7% (76.0–85.0), against 74.7% for ODI thresholds at the conventional AHI cut-offs and 81.3–83.7% for logistic

regression, k-nearest neighbors, SVM and random forest. The refined FIS did not differ significantly from any of these four models (McNemar $p \geq 0.24$) and made no error larger than one severity grade. Under five-fold cross-validation the machine-learning models stayed four percentage points ahead (82.2–83.9% versus 79.3%). Accuracy was sensitive to the position of the SpO_2 membership functions and comparatively insensitive to those of T90 and arousal index. In this simulation, a refined fuzzy system recovers most of the accuracy of black-box models while exposing every rule that contributed to a decision. Because the data are synthetic and ODI was generated as a close proxy of AHI, the findings describe the behaviors of the method and should not be read as clinical performance; validation on polysomnography recordings is the necessary next step.

1. INTRODUCTION

1.1 Background and Clinical Motivation

Obstructive sleep apnea is one of the commonest chronic sleep disorders. A widely cited literature-based estimate put the global burden at one billion adults aged 30 to 69 years, with more than 400 million having moderate-to-severe disease [1]. Earlier population work from the Wisconsin Sleep Cohort had already shown that sleep-disordered breathing was far more frequent among middle-aged adults than clinical practice at the time implied [2], and later reviews confirmed that a large share of affected people are never diagnosed [3]. The consequences are not trivial. Repeated airway collapse fragments sleep and produces cycles of intermittent hypoxemia, and has been linked to daytime sleepiness, hypertension and other cardiometabolic problems.

Severity is the quantity that steers management. By the convention used in most clinical guidelines, an AHI below 5 events per hour is considered normal, 5 to under 15 mild, 15 to under 30 moderate, and 30 or more severe [4,28]. These four bands influence who is offered positive airway pressure, who is referred to for oral appliances or surgery, and how urgently follow-up is arranged. The AHI itself comes from attended in-laboratory polysomnography (PSG) or from home sleep apnea testing, and the guideline literature leaves little doubt that PSG remains the reference standard for diagnosis [5].

1.2 Challenges in sleep apnea severity classification

The trouble starts with the index. The AHI counts apneas and hypopnea events, and the number you obtain depends on how hypopnea is defined. Ruehland and colleagues showed that moving from a 4% desaturation criterion to a 3%-or-arousal criterion changed the AHI appreciably in the same recordings, which is enough to move patients across severity boundaries [6]. Night-to-night variability, sleeping position, sleep stage distribution and scorer differences add further spread.

There is also a structural problem. The four bands are cut from a continuous variable, so two patients with AHI values of 14 and 16 land in different classes although their physiology is identical. Any classifier that predicts the band from surrogate measurements inherits that discontinuity: most of its errors fall near the cut-offs, and a model is better judged by how far its mistakes are from the truth than by raw accuracy alone. Finally, access is limited. Full PSG is expensive and slow to schedule, which is why oximetry-based and other simplified measures keep attracting attention as screening and triage tools.

1.3 Fuzzy logic as an interpretable approach

Fuzzy set theory, introduced by Zadeh [10], replaces the sharp in-or-out membership of classical sets with a degree of membership between 0 and 1. A saturation nadir of 86% is then not simply “normal” or

"abnormal"; it can belong partly to a mild-desaturation set and partly to a moderate-desaturation set. Mamdani and Assilian turned the idea into a practical inference scheme in which expert knowledge is written as linguistic IF–THEN rules [11]. The attraction for medicine has been noted for a long time [12,13]: clinicians already reason in qualitative terms such as "markedly elevated" or "borderline", and a rule such as IF ODI is moderate AND arousal index is elevated THEN severity is moderate can be read, challenged and corrected by someone with no machine-learning background.

That readability matters more than it once did. Rudin has argued that for high-stakes decisions one should prefer models that are interpretable by construction instead of explaining opaque ones after the fact [14], and Doshi-Velez and Kim have called for a more rigorous way of deciding what "interpretable" is supposed to mean in a given application [15]. A fuzzy rule base is one of the few model families where the full decision logic can be printed on a page.

1.4 Research gap

Two gaps motivated the present work. First, fuzzy systems are usually presented either as hand-built expert systems with no systematic comparison against modern classifiers, or as heavily tuned neuro-fuzzy hybrids whose rule bases are no longer readable. What is rarely shown is the trade-off itself: how much accuracy is bought, or lost, when an expert-specified fuzzy system is refined only modestly (membership breakpoints, rule weights, and output cut-points), so that the structure stays intelligible. Second, claims about interpretable models are hard to evaluate on clinical data because the true generating mechanism is unknown. A controlled simulation, in which the relationship between the latent AHI and each measured signal is written down, lets one separate the behavior of the method from the idiosyncrasies of a particular dataset.

1.5 Research questions and contributions

We set out to answer four questions.

RQ1: How accurately can an expert-specified Mamdani FIS grade OSA severity from four oximetry and arousal measures, and how much does data-driven refinement improve it?

RQ2: How does the refined FIS compare with crisp threshold rules and with standard machine-learning classifiers on the same data?

RQ3: Which inputs carry the decision, and how fragile is the system to the placement of its membership functions?

RQ4: What does the system reveal about an individual decision that a black-box model does not?

The contributions are as follows. We specify a complete, reproducible simulation framework for studying severity classification, including the data-generating equations. We present a 28-rule Mamdani FIS together with a restricted refinement procedure that preserves the rule structure. We benchmark it against three threshold comparators and five machine-learning models using hold-out evaluation, cross-validation, exact paired tests, ablation, and sensitivity analysis. And we release the dataset and code so that the experiment can be repeated or extended with real recordings.

2. Related Work

2.1 Conventional sleep apnea assessment

Attended polysomnography records electroencephalography, airflow, respiratory effort, oxygen saturation and other channels, and events are scored by trained technologists according to the AASM manual, whose respiratory scoring rules were updated in 2012 [4]. The Sleep Heart Health Study showed that this kind of recording can be conducted at scale in community cohorts [7], and it contributed substantial epidemiological evidence on the cardiovascular correlates of sleep-disordered breathing. The practical limitations are familiar: a night in the laboratory is costly, waiting lists are long in many health systems, and one night may not represent habitual sleep. Home testing and portable monitors have therefore moved into routine use for

patients with a high pre-test probability [5], although their output is again summarized by an AHI-like index and the same classification boundaries apply.

2.2 Physiological signal-based classification

A large body of work tries to estimate apnea or its severity from a smaller set of signals. The PhysioNet Apnea-ECG database, released for a computing-in-cardiology challenge, made single-lead electrocardiogram recordings with expert apnea annotations publicly available [8,9] and has been used in many subsequent studies. Oximetry-derived indices are attractive because pulse oximeters are cheap and ubiquitous. The ODI, the SpO_2 nadir and the percentage of the recording spent below 90% saturation (T90) summaries how often and how deeply saturation falls, while arousal-based measures capture the sleep fragmentation that desaturation alone can miss. Each carries different information, and none of them is a perfect surrogate for the AHI, partly because hypopneas that end in an arousal without desaturation are invisible to oximetry.

2.3 Machine-learning approaches

Supervised classifiers such as logistic regression, support vector machines [20], random forests [19], and gradient boosting [21] are the usual benchmarks for tabular physiological data, and implementations are readily available [17]. They perform well when features are informative and sample sizes are adequate. What they do not give is a rationale a clinician can verify: a random forest of several hundred trees, or a kernel SVM, produces a class label without a compact account of why. Post-hoc explanation tools exist, but they describe the model and not necessarily the physiology, which is part of the reason Rudin cautioned against relying on them in high-stakes settings [14].

2.4 Fuzzy logic and explainable decision systems

Fuzzy inference has been applied in medicine for decades, from diagnostic support to monitoring, and reviews of the area stress both its strengths (transparent handling of vague clinical categories) and its weaknesses (rule bases that are laborious to elicit and membership functions that are chosen somewhat arbitrarily) [12,13]. The usual response to the second weakness is automatic tuning. Evolutionary optimizers such as differential evolution [16] are a natural choice since fuzzy-system accuracy is a non-differentiable function of its parameters. The risk is that unconstrained tuning erases the very interpretability that motivated the approach. In the present study the tuning is deliberately restricted: rule antecedents and consequents are never altered, membership functions are only shifted and rescaled, and rule weights lie in a bounded interval.

3. Materials and Methods

3.1 Study design

The study is a computational simulation with a pre-specified analysis plan. A synthetic cohort was generated once from the data-generating model in Section 3.2 and was then split, with stratification on severity class, into a 70% training set (n = 700) and a 30% hold-out test set (n = 300). All parameter fitting, including fuzzy-system refinement, hyperparameter selection for the machine-learning models and tuning of threshold cut-offs, used the training set only. The hold-out set was touched once for the primary comparison. Five-fold stratified cross-validation on the full cohort, with all fitting repeated inside each fold, provided a second, less split-dependent estimate. Figure 1 summarizes the framework.

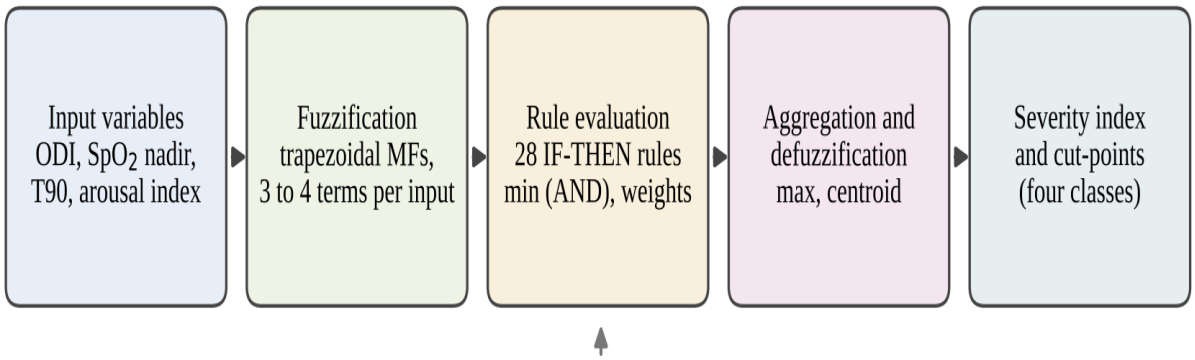


Figure 1. Overview of the fuzzy inference framework. The refinement stage acts on the parameters of the system but not on its rule structure.

3.2 Simulated cohort and data-generation process

We simulated 1,000 adults. The reference variable is a latent AHI drawn from a log-normal distribution (location 2.45, scale 1.0 on the natural-log scale), truncated to 0.2–110 events per hour. This gives a right-skewed distribution with a median near 11.6, which is close to what one sees in clinic-referred samples. The reference class was assigned from the AHI with the cut-offs of 5, 15 and 30 events per hour [4,28]. The measured inputs were then generated conditionally on the AHI, with multiplicative and additive noise added so that no single signal identifies the class without error. The complete model is listed in Table 1. Age, sex, and body mass index were generated as descriptive covariates correlated weakly with AHI; they are reported for completeness and were not given to any classifier.

Table 1. Data-generating model for the simulated cohort (N = 1,000; random seed 2026). $\mathcal{N}(\mu, \sigma^2)$ denotes a normal distribution; all values were rounded to one decimal place after clipping.

Variable	Generating equation	Clipping range
AHI (events/h)	$\exp(\mathcal{N}(2.45, 1.0^2))$	0.2–110
ODI (events/h)	$0.85 \cdot \text{AHI} \cdot \exp(\mathcal{N}(0, 0.28^2)) + \mathcal{N}(0.3, 0.9^2)$	0–100
SpO_2 nadir (%)	$95.5 - 22 \cdot (1 - \exp(-\text{AHI}/28)) + \mathcal{N}(0, 3.0^2)$	50–99
T90 (%)	$0.6 \cdot \text{AHI}^{1.05} \cdot \exp(\mathcal{N}(0, 0.55^2))$, set to 0 with probability 0.07	0–95
Arousal index (events/h)	$6 + 0.8 \cdot \text{AHI} \cdot \exp(\mathcal{N}(0, 0.35^2)) + \mathcal{N}(0, 3.0^2)$	0–100
Age (years)	$\mathcal{N}(40 + 3 \cdot \ln(1 + \text{AHI}), 11^2)$	20–85
Male sex	Bernoulli$(0.45 + 0.08 \cdot \tanh(\ln \text{AHI} - 2.4))$	–
BMI (kg/m²)	$\mathcal{N}(25 + 2.2 \cdot \ln(1 + \text{AHI}), 3.6^2)$	17–55

Two features of this design deserve to be stated plainly. ODI was deliberately made a strong but imperfect proxy of AHI, as it is in practice when most events are accompanied by a desaturation, and the 7% structural

zeros in T90 mimic patients whose saturation never falls below 90% despite repeated events. Because the generator is known, the statistical ceiling for any classifier is also bounded by the noise terms; no model can be expected to reach perfect accuracy on these data.

3.3 Physiological input variables

The FIS receives four inputs. The oxygen desaturation index (ODI) is the number of desaturation events per hour of recording. The SpO_2 nadir is the lowest saturation reached during the night. T90 is the percentage of the recording spent below 90% saturation, and the arousal index is the number of arousals per hour of sleep. These were chosen because they are available from overnight oximetry plus an arousal measure, and because they approximate complementary aspects of disease burden: event frequency (ODI), depth (SpO_2 nadir), duration (T90), and sleep disruption (arousal index). The class to be predicted is the four-level severity band defined from the AHI.

3.4 Mamdani fuzzy inference system

We used a Mamdani-type system [11] with the usual five steps [27]. Each crisp input x is fuzzified through trapezoidal membership functions.

$$\mu(x; a, b, c, d) = \max\left(0,\ \min\left(\frac{x-a}{b-a},\ 1,\ \frac{d-x}{d-c}\right)\right) \tag{1}$$

where triangles are the special case b = c and open shoulders are obtained by setting a = b or c = d. The firing strength of rule k, with antecedent terms A_{k1}, ..., A_{km} combined by the minimum operator for AND and multiplied by a rule weight $w_k \in [0, 1]$, is

$$\alpha_k = w_k \cdot \min_j \ \mu_{A_{kj}}(x_j) \tag{2}$$

Each rule clips its consequent set at height α_k (minimum implication), the clipped sets are combined by the maximum operator, and the aggregated output set is defuzzified with the centroid method over a grid of 241 points on a "severity index" scale from 0 to 60, interpreted as an AHI-equivalent:

$$y^* = \frac{\int y\, \mu_{\text{agg}}(y)\, dy}{\int \mu_{\text{agg}}(y)\, dy} \tag{3}$$

The crisp class is obtained by comparing y* with three cut-points. For the expert system the cut-points were not tuned: they were fixed at the midpoints between the centroids of adjacent output sets (6.0, 16.2 and 34.4), which is the point at which a fully activated output set stop being the closest one. If no rule fired, the index defaulted to fifteen, the middle of the scale; this never happened in the test set.

3.5 Membership functions and rule base

Membership functions were first specified from clinical reasoning and from the conventional severity bands, and then adjusted once after a pilot run, as described below. ODI and the severity output received four terms (normal, mild, moderate, severe), SpO_2 nadir received four terms (normal, mild, moderate, and severe desaturation), and T90 and arousal index received three terms each. Breakpoints are listed in Table 2 and

drawn in Figure 2. The ODI breakpoints were placed slightly below the AHI cut-offs because ODI is typically lower than AHI.

The rule base has twenty-eight rules (Appendix A) and is organized in three tiers. Four anchor rules map each ODI term directly to the matching severity, giving the system a sensible default when other inputs are uninformative. Several corroboration rules strengthen or soften that default when a second signal agrees or disagrees (for example, a moderate ODI combined with a markedly low SpO_2 nadir and a high T90 points to severe disease). The remaining rules handle discordant combinations, such as a normal ODI together with a high T90, which are pushed one grade upward. We should be candid that the expert system went through one revision. A first draft, in which single secondary signals could escalate a class, over-graded mild patients badly (accuracy about 60% on the full cohort). We then re-centered the ODI breakpoints, added the anchor rules, made the escalation rules require two concurrent signals, and derived the output cut-points from the centroids of the output sets as described above. That revision was guided by results on the same simulated cohort and is therefore a source of optimism for the expert system and, indirectly, for the refined one; we report the effect of membership perturbation in Section 4.7 partly for this reason.

Table 2. Membership functions of the expert and refined systems, given as trapezoid breakpoints (a, b, c, d). Refined breakpoints are those obtained by the procedure in Section 3.6 on the training set.

Variable	Linguistic term	Expert (a, b, c, d)	Refined (a, b, c, d)
ODI (events/h)	Normal	(0, 0, 3, 6)	(0, 0, 3.1, 6.5)
	Mild	(3, 7, 11, 16)	(3.1, 7.6, 12.0, 17.6)
	Moderate	(12, 18, 24, 32)	(13.2, 19.9, 26.6, 35.5)
	Severe	(24, 34, 100, 100)	(26.6, 37.7, 100, 100)
SpO_2 nadir (%)	Severe	(50, 50, 76, 82)	(50, 50, 73.2, 79.2)
	Moderate	(78, 84, 84, 89)	(75.2, 81.3, 81.3, 86.4)
	Mild	(86, 90, 90, 93.5)	(83.3, 87.4, 87.4, 90.9)
	Normal	(91, 94, 100, 100)	(88.4, 91.4, 100, 100)
T90 (%)	Low	(0, 0, 2, 5)	(0, 0, 3.9, 7.2)
	Moderate	(3, 10, 10, 20)	(5.0, 12.7, 12.7, 23.5)
	High	(15, 25, 100, 100)	(18.1, 29.0, 100, 100)
Arousal index (events/h)	Low	(0, 0, 10, 16)	(0, 0, 15.3, 24.4)
	Elevated	(12, 22, 22, 34)	(18.3, 33.5, 33.5, 51.7)
	High	(28, 40, 100, 100)	(42.6, 60.9, 100, 100)
Severity index (output)			
	Normal	(0, 0, 2, 6)	unchanged
	Mild	(3, 10, 10, 17)	unchanged
	Moderate	(13, 22, 22, 32)	unchanged
	Severe	(27, 40, 60, 60)	unchanged

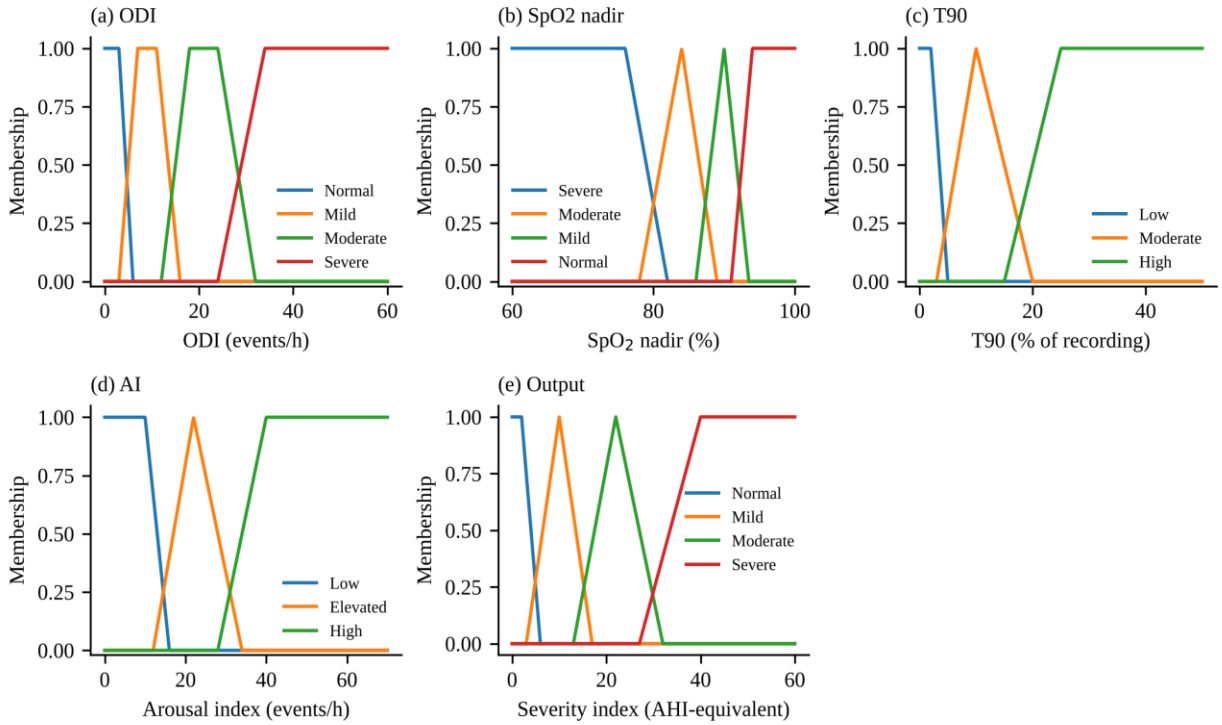


Figure 2. Membership functions of the expert fuzzy system: (a) ODI, (b) SpO_2 nadir, (c) T90, (d) arousal index and (e) the output severity index.

3.6 Refined fuzzy system.

Refinement adjusts parameters but never the rule structure. For each input variable v the interior breakpoints p of all its membership functions were transformed jointly as $p' = s_v \cdot p + d_v$, where s_v is a scale and d_v a shift; breakpoints lying on the boundary of the variable's range were left in place so that open shoulders remained open. Each of the twenty-eight rules received a weight between 0.1 and 1.0, and the three cut-points were allowed to vary in the ranges 3–10, 10–24 and 24–45. This gives thirty-nine free parameters. They were optimized with differential evolution [16], implemented in SciPy [18], minimizing 1 − macro-F1 on the training set, with the expert system injected as one member of the initial population so that the search could not end worse than where it started. The population size was eight times the number of parameters, and the search was stopped after 45 generations for the primary hold-out analysis (30 generations inside each cross-validation fold, to keep running time reasonable). The membership functions of the output variable were not modified.

3.7 Crisp rule-based comparator

Three threshold classifiers served as comparators. The first (ODI thresholds, AASM cut-offs) simply applies the conventional bands of 5, 15 and 30 events per hour to the ODI. The second (ODI thresholds with escalation) does the same but raises the class by one level when the SpO_2 nadir is below 80% or T90 is at least 20%. The third (data-tuned ODI thresholds) chooses the three ODI cut-offs that maximize macro-F1 on the training set by exhaustive search over a grid; the selected values were 4.5, 14.0 and 31.0 events per hour. These comparators matter because they represent what a clinician could do with a pocket calculator, and an FIS must beat them to justify its added complexity.

3.8 Machine-learning baselines

Five standard classifiers were implemented in scikit-learn [17]: multinomial logistic regression, k-nearest neighbors, a support vector machine with an RBF kernel [20], a random forest [19], and gradient-boosted trees [21]. For logistic regression, k-NN and SVM inputs were log (1 + x)-transformed and standardized. Hyperparameters were chosen by five-fold grid search on the training data using macro-F1; the selected configurations are in Appendix B. No classifier received age, sex, or BMI.

3.9 Evaluation metrics and statistical analysis

The primary metrics were overall accuracy, balanced accuracy (the mean of the four per-class sensitivities) and macro-average F1. Because severity is ordinal, we also report Cohen's kappa [22], the quadratic-weighted kappa [23], and the proportion of errors that were adjacent-class and non-adjacent. Per-class sensitivity, specificity and F1 were computed one-versus-rest. Agreement was described with the conventional descriptors of Landis and Koch were useful [26]. For the continuous outputs we computed the area under the ROC curve (AUC) for the three clinically relevant thresholds (AHI ≥ 5, ≥ 15, ≥ 30), using the defuzzified severity index for the FIS, the summed probability of the relevant classes for the machine-learning models, and raw ODI as a one-variable reference. Spearman's rank correlation between each continuous score and the AHI was also computed.

Ninety-five percent confidence intervals on the hold-out set were obtained by non-parametric bootstrap with 2,000 resamples of the test cases [25]. Because every model was applied to the same three hundred test cases, pairwise differences from the refined FIS were assessed with the exact McNemar test on the discordant pairs [24]. No correction for multiple comparisons was applied; with nine comparisons, readers should treat p-values between 0.01 and 0.05 with caution. Cross-validation results are given as mean ± standard deviation across the five folds.

Feature ablation. To see what each input contributes, a variable was removed from the FIS by treating its antecedents as always satisfied and deleting rules that referenced only that variable. For the random forest, the variable was simply excluded and the model retrained. Sensitivity analysis. Breakpoints of one variable at a time were scaled by ±5%, ±10% and ±20% (for SpO_2 nadir, shifted by ±1, ±2 and ±4 percentage points); all breakpoints were also perturbed jointly at random with standard deviations of 5%, 10% and 20% (300 draws each), and a variant with smooth S-shaped instead of linear membership edges was tested. The expert system was assessed on the full cohort, since it uses no training data, and the refined system on the hold-out set.

All analyses were run in Python 3 with NumPy, SciPy and scikit-learn. The fuzzy inference engine was written from scratch so that every step could be inspected.

4. Results

4.1 Simulated cohort characteristics

The cohort contained two hundred participants with normal AHI (20.0%), 392 with mild (39.2%), 233 with moderate (23.3%) and 175 with severe OSA (17.5%). Table 3 gives the characteristics by class. All four physiological inputs rose or fell monotonically with severity, as designed. The mean ODI went from 3.1 events/h in the normal group to 47.5 in the severe group, and the mean SpO_2 nadir fell from 93.1% to 77.5%. The training and hold-out sets contained 700 (140/274/163/123) and 300 (60/118/70/52) participants in the four classes, respectively.

Table 3. Characteristics of the simulated cohort by reference severity class (mean ± SD unless stated).

Characteristic	Overall (n = 1000)	Normal (n = 200)	Mild (n = 392)	Moderate (n = 233)	Severe (n = 175)
Share of cohort, %	100.0	20.0	39.2	23.3	17.5
Age, years	46.8 ± 11.0	42.2 ± 10.9	46.2 ± 11.0	48.4 ± 10.6	51.3 ± 9.8

Characteristic	Overall (n = 1000)	Normal (n = 200)	Mild (n = 392)	Moderate (n = 233)	Severe (n = 175)
Male sex, %	47.4	32.5	44.9	56.7	57.7
BMI, kg/m²	30.8 ± 4.2	27.9 ± 3.5	30.2 ± 3.8	31.9 ± 3.7	34.0 ± 3.5
AHI, events/h (median [IQR])	12.0 [5.9–23.6]	3.2 [2.0–4.1]	9.3 [6.8–11.8]	20.9 [17.8–24.8]	46.8 [36.9–65.0]
ODI, events/h	16.9 ± 18.4	3.1 ± 1.6	8.8 ± 3.8	19.5 ± 6.8	47.5 ± 23.0
SpO_2 nadir, %	86.7 ± 6.3	93.1 ± 3.1	89.3 ± 3.4	83.7 ± 3.5	77.5 ± 3.8
T90, %	13.9 ± 17.9	2.2 ± 1.8	6.7 ± 4.7	15.8 ± 11.0	40.8 ± 24.9
Arousal index, events/h	21.9 ± 18.4	8.5 ± 3.2	13.9 ± 5.2	23.6 ± 7.8	52.6 ± 21.9

4.2 Expert fuzzy-system performance

Without any use of the training data, the expert FIS classified 70.3% of the hold-out cases correctly (95% CI 65.0–75.3; 68.4% on the full cohort), with a macro-F1 of 70.2% and a quadratic kappa of 0.837 (Table 4). Its confusion matrix (Figure 4) shows a consistent bias: 33 of 60 normal cases were graded as mild and 35 of 118 mild cases as moderate, so sensitivity for the normal class was only 43.3% (Table 5). Higher grades were recognized well, with sensitivities of 78.6% for moderate and 94.2% for severe disease. Only one case out of three hundred was misclassified by more than one grade. In short, the expert system had good ordinal behavior but was tilted towards over-grading, which is the sort of systematic offset that a modest recalibration should be able to remove.

4.3 Refined-system performance

The refinement stage reached a training macro-F1 of 0.843. On the hold-out set the refined FIS achieved an accuracy of 80.7% (95% CI 76.0–85.0), a macro-F1 of 81.1% and a quadratic kappa of 0.899, an improvement of 10.4 accuracy points over the expert system (McNemar $b = 54$, $c = 23$, $p < 0.001$). Sensitivity for the normal class rose from 43.3% to 81.7% without losing specificity (96.7%), and no hold-out case was misclassified by more than one grade. Adjacent-class errors fell from 29.3% to 19.3% of all cases.

The parameters that changed most are informative. The SpO_2 breakpoints moved about four percentage points downward (shift −3.96), which means the refined system waits for a lower nadir before regarding desaturation as moderate or severe. The arousal-index breakpoints were stretched by 50% (scale 1.52), the ODI breakpoints by 12%, and the T90 breakpoints by 9% with a small upward shift. The cut-points moved to 6.6, 12.9 and 31.8, so the mild-to-moderate threshold dropped by more than three units relative to the expert setting. The optimizer also reduced many rule weights: thirteen of the twenty-eight rules ended with a weight below 0.5, and the mean weight was 0.57 (Appendix A). The refined breakpoints are listed in Table 2 and compared with the expert ones in Figure 3.

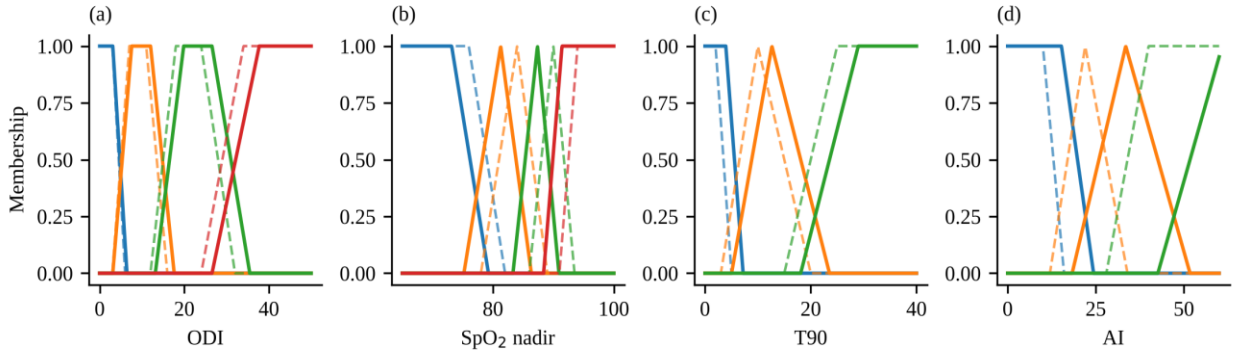


Figure 3. Expert (dashed) and refined (solid) membership functions for (a) ODI, (b) SpO_2 nadir, (c) T90 and (d) arousal index.

Table 4. Classification performance on the hold-out set (n = 300). Accuracy is shown with a 95% bootstrap confidence interval; the last two columns give the share of all cases misclassified by one grade and by two or more grades.

Model	Accuracy, % (95% CI)	Balanced acc., %	Macro-F1, %	Cohen's κ	Quadratic κ	Adjacent-class errors, %	Errors ≥2 classes, %
Expert FIS	70.3 (65.0–75.3)	71.2	70.2	0.589	0.837	29.3	0.3
Refined FIS	80.7 (76.0–85.0)	80.6	81.1	0.731	0.899	19.3	0.0
ODI thresholds (AASM cut-offs)	74.7 (69.7–79.3)	73.5	74.4	0.643	0.867	25.3	0.0
ODI thresholds + SpO_2/T90 escalation	78.0 (73.0–82.7)	78.6	77.3	0.695	0.894	22.0	0.0
ODI thresholds (data-tuned)	76.7 (72.0–81.3)	75.2	76.5	0.672	0.875	23.3	0.0
Logistic regression	82.7 (78.3–87.0)	81.4	82.6	0.756	0.908	17.3	0.0
k-NN	83.7 (79.3–87.7)	82.4	83.6	0.771	0.913	16.3	0.0
SVM (RBF)	82.0 (77.7–86.3)	80.5	82.0	0.746	0.902	18.0	0.0
Random forest	81.3 (76.7–85.7)	80.4	81.3	0.738	0.902	18.7	0.0
Gradient boosting	78.3 (73.3–83.0)	77.6	78.5	0.697	0.885	21.7	0.0

Table 5. Per-class sensitivity, specificity and F1 on the hold-out set for the expert FIS, the refined FIS, and the random forest.

Model	Severity class	Sensitivity, %	Specificity, %	F1, %
Expert FIS	Normal	43.3	99.2	59.1
	Mild	68.6	81.3	69.5
	Moderate	78.6	83.0	67.1
	Severe	94.2	94.4	85.2
Refined FIS	Normal	81.7	96.7	83.8
	Mild	82.2	87.4	81.5
	Moderate	75.7	90.4	73.1
	Severe	82.7	98.0	86.0
Random forest	Normal	73.3	97.1	79.3
	Mild	87.3	84.6	82.7
	Moderate	74.3	93.5	75.9
	Severe	86.5	97.6	87.4

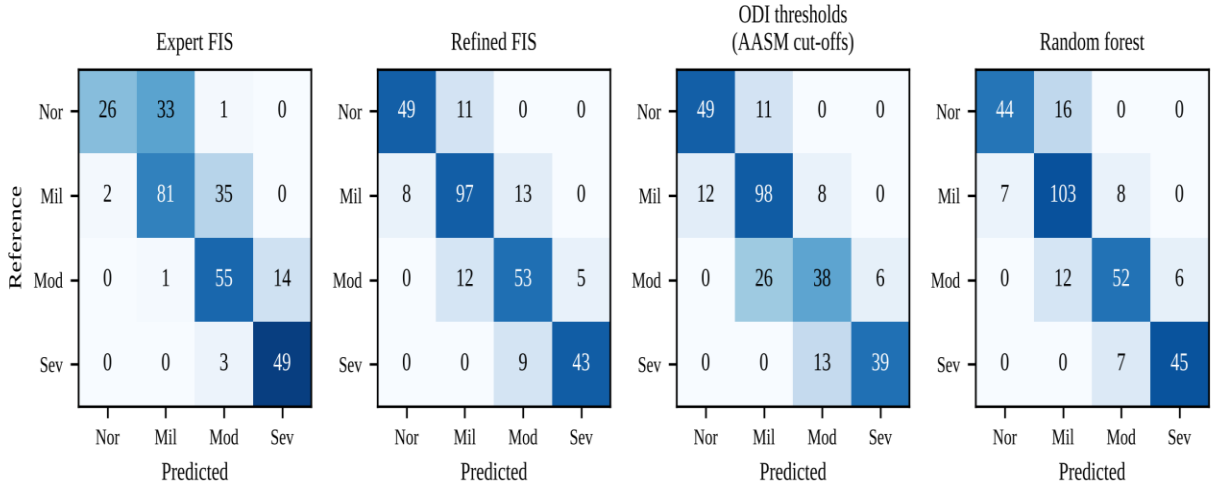


Figure 4. Confusion matrices on the hold-out set (rows: reference class; columns: predicted class; Nor = normal, Mil = mild, Mod = moderate, Sev = severe).

4.4 Comparison with machine-learning models

Four of the five machine-learning models fell in a narrow band on the hold-out set: logistic regression 82.7%, k-NN 83.7%, SVM 82.0% and random forest 81.3%, against 80.7% for the refined FIS (Table 4). Gradient boosting reached 78.3%. The confidence intervals overlap completely, and the exact McNemar tests did not find a significant difference between the refined FIS and any of these five models (p between 0.24 and 0.87; Table 8). For example, against the random forest the two models disagreed on thirty-eight cases, of which the FIS got 18 right and the forest 20.

The picture is a little less flattering under cross-validation (Table 6, Figure 5). Averaged over five folds the refined FIS had an accuracy of 79.3 ± 3.5%, whereas the five machine-learning models scored between 82.2%

and 83.9%, so the gap is closer to four percentage points than to the one to three seen on the hold-out split. Part of this is the shorter optimization budget used inside the folds (30 instead of 45 generations), and part is that a 700-case hold-out set is simply a noisy instrument. We would therefore describe the refined FIS as slightly less accurate than the best black-box models, not as equivalent to them. On the other hand, the quadratic kappa tells a consistent and encouraging story: 0.894 ± 0.019 for the refined FIS against 0.908–0.916 for the machine-learning models.

Table 6. Five-fold cross-validated performance on the full cohort (mean ± SD across folds). All fitting, including fuzzy refinement, was repeated within each training fold.

Model	Accuracy, %	Balanced acc., %	Macro-F1, %	Quadratic κ
Expert FIS	68.4 ± 2.8	69.1 ± 3.1	68.2 ± 2.9	0.834 ± 0.018
Refined FIS	79.3 ± 3.5	79.7 ± 3.3	79.7 ± 3.2	0.894 ± 0.019
ODI thresholds (AASM cut-offs)	75.9 ± 3.0	75.6 ± 3.1	75.9 ± 3.0	0.875 ± 0.013
ODI thresholds + SpO_2/T90 escalation	76.0 ± 3.8	77.2 ± 3.8	74.8 ± 4.0	0.888 ± 0.018
ODI thresholds (data-tuned)	76.8 ± 2.6	76.9 ± 2.7	76.9 ± 2.8	0.880 ± 0.013
Logistic regression	83.7 ± 2.2	83.4 ± 1.6	83.9 ± 2.1	0.916 ± 0.010
k-NN	83.9 ± 2.9	83.5 ± 3.1	84.0 ± 3.0	0.916 ± 0.016
SVM (RBF)	83.4 ± 3.3	83.0 ± 3.0	83.6 ± 3.1	0.914 ± 0.017
Random forest	82.2 ± 2.7	81.8 ± 2.2	82.2 ± 2.5	0.908 ± 0.013
Gradient boosting	82.9 ± 2.7	83.1 ± 2.6	83.1 ± 2.8	0.913 ± 0.014

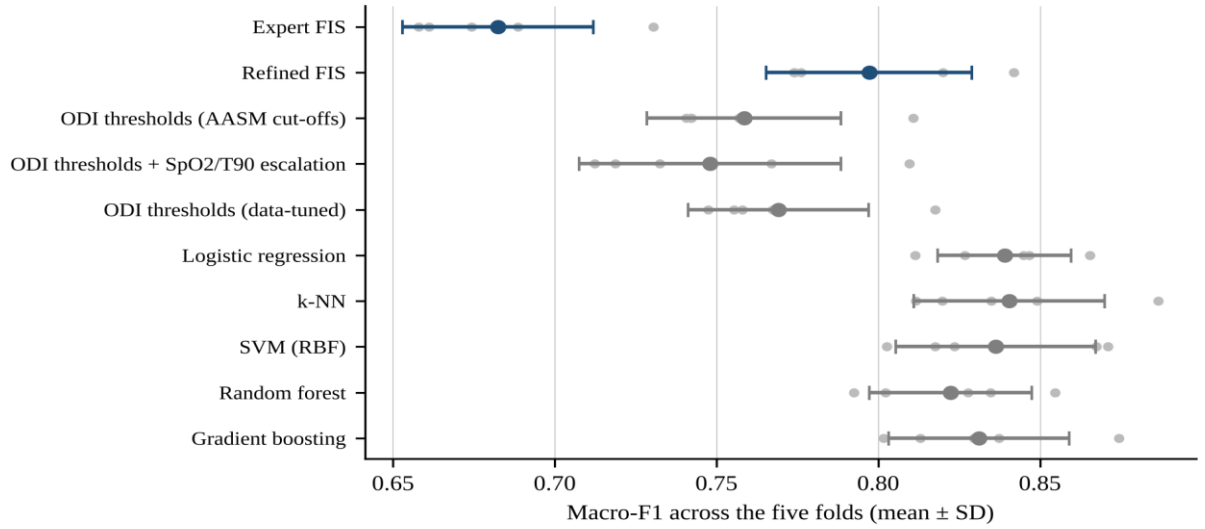


Figure 5. Macro-F1 of all models across the five cross-validation folds. Small grey points are individual folds; the filled marker is the mean with ±1 SD.

The continuous outputs were also very similar on discriminative terms (Table 7). For the clinically most relevant threshold of AHI ≥ 15, AUC values ranged from 0.975 (raw ODI) to 0.988 (logistic regression and

SVM), with the refined FIS at 0.982. For AHI ≥ 5 the raw ODI (0.973) outperformed both fuzzy systems (0.960 and 0.962), a reminder that when one variable is sufficient, combining several inputs can add noise at the low end. Spearman correlations of the defuzzified index with the AHI were 0.948 (expert) and 0.955 (refined), against 0.958 for ODI alone.

Table 7. Area under the ROC curve on the hold-out set for three severity thresholds.

Model / score	AUC, AHI ≥ 5	AUC, AHI ≥ 15	AUC, AHI ≥ 30
Expert FIS	0.960	0.979	0.985
Refined FIS	0.962	0.982	0.987
ODI (continuous)	0.973	0.975	0.980
Logistic regression	0.979	0.988	0.990
k-NN	0.961	0.984	0.981
SVM (RBF)	0.978	0.988	0.990
Random forest	0.974	0.985	0.989
Gradient boosting	0.967	0.982	0.974

Table 8. Paired comparison of the refined FIS with each comparator on the hold-out set (exact McNemar test on discordant cases).

Comparator	Refined correct, comparator wrong	Comparator correct, refined wrong	Exact McNemar p
Expert FIS	54	23	<0.001
ODI thresholds (AASM cut-offs)	29	11	0.006
ODI thresholds + SpO_2/T90 escalation	25	17	0.280
ODI thresholds (data-tuned)	21	9	0.043
Logistic regression	17	23	0.430
k-NN	19	28	0.243
SVM (RBF)	20	24	0.652
Random forest	18	20	0.871
Gradient boosting	25	18	0.360

4.5 Threshold-based classification

The crisp comparators were informative. Applying the conventional AHI bands directly to the ODI gave 74.7% accuracy (95% CI 69.7–79.3) and a notable weakness for the moderate class, with a sensitivity of only 54.3%: of 70 moderate cases, 26 were assigned to mild and 6 to severe. Tuning the cut-offs on the training data (4.5, 14.0 and 31.0) improved accuracy to 76.7%, and adding escalation by SpO_2 nadir and T90 gave 78.0%. The refined FIS was significantly better than the first and third of these (McNemar $p = 0.006$ and 0.043) but not clearly better than the escalated rule ($p = 0.28$). The honest reading is that a great deal of what the FIS achieves can be approximated by a very simple two-step crisp rule, and that the fuzzy system's real advantage is a smoother, graded treatment of cases near boundaries, together with a rule base that can be

adjusted without re-deriving a decision tree. In cross-validation the threshold comparators clustered at 75.9–76.8%, 2.5 to 3.4 points below the refined FIS.

4.6 Feature ablation

Removing an input produced very different outcomes for the fuzzy systems and for the random forest (Table 9). The random forest barely noticed: with all four inputs it scored 81.3%, and without ODI 76.7%, without SpO_2 nadir 79.0%, and without T90 or arousal index 80.7%. A forest using only the ODI reached 75.0%. The inputs carry overlapping information, since all four are generated from the same latent AHI, so each can partially stand in for the others.

The fuzzy systems collapsed. Without SpO_2 nadir the expert FIS fell to 36.7% accuracy and without T90 to 42.7%, and the refined FIS fell to between 48.3% and 59.3% depending on which variable was removed. This is not evidence that the fuzzy system extracts more from each input; it shows that an FIS cannot simply lose a variable, because deleting an antecedent turns several rules into duplicates with conflicting consequents (for instance, ODI normal AND SpO_2 normal → normal and ODI normal AND SpO_2 mild → mild both become ODI normal). The rule base would need to be re-elicited or re-tuned for a reduced sensor set. We consider this a real practical limitation of the approach and return to it in Section 5.

Table 9. Accuracy on the hold-out set when one input is removed. The fuzzy systems were evaluated without re-tuning; the random forest was retrained.

Input set	Expert FIS acc., %	Refined FIS acc., %	Random forest acc., %	Random forest quad. κ
All four inputs	70.3	80.7	81.3	0.902
Without ODI	61.3	54.3	76.7	0.880
Without SpO_2 nadir	36.7	59.3	79.0	0.889
Without T90	42.7	52.7	80.7	0.900
Without arousal index	50.0	48.3	80.7	0.899

4.7 Membership-function sensitivity

The expert system's accuracy depended strongly on where the SpO_2 membership functions sat, and hardly at all on T90 and arousal index (Table 10, Figure 6). Shifting the SpO_2 breakpoints down by 4 points raised accuracy from 68.4% to 77.5%, while shifting them up by 4 points lowered it to 50.9%. ODI also mattered: scaling its breakpoints by −20% reduced accuracy to 61.7%, whereas +20% raised it to 71.3%. By contrast, perturbing T90 or arousal-index breakpoints by up to ±20% moved accuracy by less than one percentage point. That the expert SpO_2 placement was suboptimal is also what the refinement stage found (the −3.96 shift), so the two analyses agree.

The refined system was less sensitive in the neighborhood of its optimum, as one would expect, but it was not immune: a 20% upward shift of the SpO_2 breakpoints reduced accuracy from 80.7% to 72.3%, and a 20% reduction of the ODI scale reduced it to 76.3%. When all breakpoints of the expert system were jittered randomly (Table 11), mean accuracy was 68.2% at a standard deviation of 5%, 66.7% at 10% and 61.5% at 20%, with the spread widening from ±2.0 to ±6.8 points. Replacing linear membership edges with smooth S-shaped ones had little effect: the expert system's accuracy rose from 68.4% to 70.0%, and the refined systems stayed at 80.7%.

Table 10. Accuracy (%) of the expert FIS on the full cohort when the breakpoints of one input are rescaled (for SpO_2 nadir the breakpoints are shifted by −4, −2, −1, +1, +2 and +4 points, matching the column order). The unperturbed accuracy is 68.4%.

Input perturbed (expert FIS)	-20%	-10%	-5%	+5%	+10%	+20%
ODI	61.7	65.0	67.4	69.9	71.1	71.3
SpO_2 (shift ±1, 2, 4 points)	77.5	74.6	72.7	64.6	59.4	50.9
T90	68.0	68.3	68.4	68.7	68.8	68.9
AI	67.8	68.1	68.3	68.8	69.0	69.3

Table 11. Accuracy of the expert FIS under random joint perturbation of all membership-function breakpoints (300 draws per level).

Perturbation SD	Mean accuracy, %	SD, % points	2.5th–97.5th percentile, %
5%	68.2	2.0	63.9–71.9
10%	66.7	3.6	59.1–72.6
20%	61.5	6.8	47.6–73.6

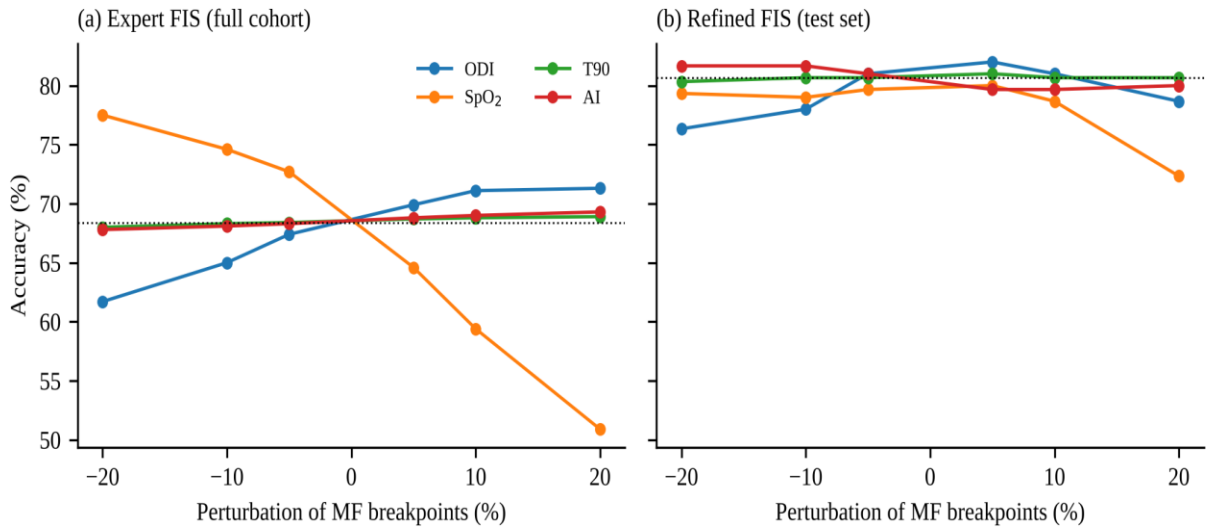


Figure 6. Accuracy as a function of the perturbation applied to the breakpoints of each input: (a) expert FIS on the full cohort and (b) refined FIS on the hold-out set. The dotted line marks the unperturbed accuracy.

5. Discussion

5.1 Interpretation of the findings

Three results stand out. The first is that the expert system, built from clinical reasoning with no access to training data, was already a reasonable classifier (70.3%) with excellent ordinal behavior, a quadratic kappa of 0.837 and no gross errors. The second is that restricted refinement moved it to the level of the standard machine-learning models on the hold-out set and to within about four points of them under cross-validation. The third, and the one we did not expect when we began, is how much of the performance is explained by just two ingredients: a well-placed set of ODI cut-offs and a correction for low SpO_2 nadir. The refined FIS outperformed the plain AASM-band ODI classifier by 6 percentage points on the hold-out set, but its margin over an escalated threshold rule was 2.7 points and not statistically distinguishable.

The per-class results (Table 5) show where the extra accuracy came from. The refined FIS fixed the expert system's tendency to over-grade normal and mild cases, and it recovered the moderate class far better than the plain ODI rule (sensitivity 75.7% versus 54.3%). Moderate disease is the hardest class, because it is bordered on both sides, and it is where a graded system should help; it is also where the refined FIS and the random forest performed alike (75.7% and 74.3%).

5.2 Performance versus interpretability

The practical question is whether four percentage points of cross-validated accuracy are worth the transparency. We think the answer depends on the setting. In a screening pathway where a misclassification costs a repeat test, the black-box models are the sensible choice. In a decision-support tool that clinicians must be able to question, or where regulators ask for a rationale, a gap of that size seems tolerable, especially since the FIS never made an error larger than one grade on the hold-out set.

A worked example shows what the FIS offers. For one hold-out participant (ODI 22.3 events/h, SpO_2 nadir 82.6%, T90 14.2%, arousal index 27.7 events/h; true AHI 22.8), the refined system produced a severity index of 22.4 and graded the case as moderate. The decision was driven by four rules: ODI is moderate AND arousal index is elevated (R22, firing strength 0.62), ODI is moderate AND T90 is moderate (R21, 0.60), ODI is moderate AND SpO_2 nadir is moderate (R19, 0.43) and the anchor rule ODI is moderate (R3, 0.17). Every number in that account can be traced to a membership degree and a weight. On average only 4.7 of the 28 rules in the refined system had a firing strength above 0.05 for a given hold-out case (5.0 for the expert system), which is small enough for a person to read. A random forest of 300 trees offers nothing comparable without an additional explanation layer.

The refinement also made some of the rule's base redundant in an informative way. On the hold-out set, 16 of the 28 rules never reached a firing strength of 0.5 in the refined system, compared with only 2 in the expert system. A clinician reviewing the system could prune or re-examine those rules. We did not do so here, since pruning decisions made on hold-out evidence would contaminate the evaluation.

5.3 Implications of simulation design

The simulation lets us be specific about what the findings do and do not mean. Because the generator is known, accuracy cannot reach 100%: ODI has multiplicative noise of about 28%, and a patient whose AHI is 14 can easily have an ODI above 15 or below 10, which is all that is needed to flip a class. The ceiling is therefore not a property of the classifiers but of the noise we put in, and all comparisons are relative. The rank-order of the models is more transferable than their absolute accuracies, but even that cannot be taken for granted.

The design also affects how the ablation results should be read. All four inputs depend on the same latent AHI, which makes them partly redundant by construction, and it is redundant inputs that a flexible model can compensate when one is dropped. In real recordings the inputs are related through physiology that is less tidy than a smooth parametric function, and ODI will not be as faithful a proxy of AHI as it is here.

5.4 Limitations and threats to validity

Several limitations deserve emphasis. Construct validity: all data are synthetic, so nothing in this paper is evidence of clinical accuracy. The noise model is Gaussian or log-normal, whereas real oximetry has artefact, probe displacement, baseline saturation differences, lung disease, and altitude effects, none of which we simulated. Circularity: the expert membership functions were written by the authors who also wrote the generator, and the one revision of the expert system was guided by the same simulated cohort; this favors the expert system and the refined system to an unknown degree. A cleaner design would have separate teams, or an expert panel that sees only the clinical definitions. Internal validity: the hold-out set has 300 cases, which leaves a confidence interval of ±4.5 points on accuracy; most pairwise differences among the top models are inside that noise. The refinement used a stochastic optimizer and a single random seed for the primary analysis, so some variation across seeds is to be expected, and we did not quantify it. No correction was made for multiple testing. Scope: we considered only four inputs, no demographic or symptom information, and no sleep-stage-dependent or positional AHI. Central sleep apnea, mixed events, and treatment effects are outside the framework. The ablation of the fuzzy systems was done without re-tuning, which exaggerates the cost of removing a variable.

5.5 Future evaluation using real PSG recordings.

The obvious next step is to repeat the pipeline on recorded data. Candidate sources include community cohorts with PSG and oximetry, such as the Sleep Heart Health Study [7], and the PhysioNet databases [8,9], although the latter provide ECG-derived rather than oximetry-derived measures and would need a different input set. A real-data study should (i) lock the rule base and membership functions with a panel of sleep physicians before looking at outcomes, (ii) use an external test cohort from a different site, (iii) evaluate calibration and not only classification metrics, (iv) report the effect of hypopnea-scoring criteria [6], and (v) include a prospective comparison of how clinicians use a rule-level explanation compared with a bare class prediction. A further technical question is whether a modest increase in rule complexity, for example adding a body-mass or symptom input, brings the FIS fully level with the black-box models without destroying its readability.

6. Conclusion

We designed and tested, in simulation, a Mamdani fuzzy system for grading OSA severity from ODI, SpO_2 nadir, T90 and arousal index. A hand-specified version was sound but biased towards over-grading (70.3% hold-out accuracy). Restricted data-driven refinement of membership breakpoints, rule weights and cut-points raised accuracy to 80.7% on the hold-out set and 79.3% under cross-validation, without any error larger than one severity grade on the hold-out data and with a decision trail of about five active rules per case. Standard machine-learning models were slightly more accurate under cross-validation (82.2–83.9%), and a simple oximetry rule with escalation came within 2.7 points on the hold-out set. The fuzzy system's value, on this data, lies in a favorable balance between accuracy and transparency, not in superior accuracy. It also showed clear weaknesses: strong dependence on how the SpO_2 membership functions is placed and a rule base that cannot lose input without re-tuning. Because every observation in the study was simulated, these conclusions describe the method and must be tested on real polysomnography and oximetry data before any clinical inference is drawn.

REFERENCES

Benjafield AV, Ayas NT, Eastwood PR, et al. Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. Lancet Respir Med. 2019;7(8):687–698.

Young T, Palta M, Dempsey J, Skatrud J, Weber S, Badr S. The occurrence of sleep-disordered breathing among middle-aged adults. N Engl J Med. 1993;328(17):1230–1235.

Punjabi NM. The epidemiology of adult obstructive sleep apnea. Proc Am Thorac Soc. 2008;5(2):136–143.

Berry RB, Budhiraja R, Gottlieb DJ, et al. Rules for scoring respiratory events in sleep: update of the 2007 AASM manual for the scoring of sleep and associated events. J Clin Sleep Med. 2012;8(5):597–619.

Kapur VK, Auckley DH, Chowdhuri S, et al. Clinical practice guideline for diagnostic testing for adult obstructive sleep apnea: an American Academy of Sleep Medicine clinical practice guideline. J Clin Sleep Med. 2017;13(3):479–504.
Ruehland WR, Rochford PD, O'Donoghue FJ, Pierce RJ, Singh P, Thornton AT. The new AASM criteria for scoring hypopneas: impact on the apnea hypopnea index. Sleep. 2009;32(2):150–157.
Quan SF, Howard BV, Iber C, et al. The Sleep Heart Health Study: design, rationale, and methods. Sleep. 1997;20(12):1077–1085.
Penzel T, Moody GB, Mark RG, Goldberger AL, Peter JH. The Apnea-ECG database. Computers in Cardiology. 2000;27:255–258.
Goldberger AL, Amaral LAN, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation. 2000;101(23):e215–e220.
Zadeh LA. Fuzzy sets. Information and Control. 1965;8(3):338–353.
Mamdani EH, Assilian S. An experiment in linguistic synthesis with a fuzzy logic controller. Int J Man-Machine Studies. 1975;7(1):1–13.
Steimann F. On the use and usefulness of fuzzy sets in medical AI. Artif Intell Med. 2001;21(1–3):131–137.
Phuong NH, Kreinovich V. Fuzzy logic and its applications in medicine. Int J Med Inform. 2001;62(2–3):165–173.
Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206–215.
Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608. 2017.
Storn R, Price K. Differential evolution: a simple and efficient heuristic for global optimization over continuous spaces. J Glob Optim. 1997;11(4):341–359.
Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–2830.
Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17(3):261–272.
Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–297.
Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;29(5):1189–1232.
Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20(1):37–46.
Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull. 1968;70(4):213–220.
McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika. 1947;12(2):153–157.
Efron B, Tibshirani RJ. An Introduction to the Bootstrap. New York: Chapman & Hall; 1993.
Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–174.
Ross TJ. Fuzzy Logic with Engineering Applications. 3rd ed. Chichester: Wiley; 2010.
Flemons WW, Buysse D, Redline S, et al. Sleep-related breathing disorders in adults: recommendations for syndrome definition and measurement techniques in clinical research. Sleep. 1999;22(5):667–689.

APPENDICES

Appendix A. Complete Fuzzy Rule Base

Table A1. The 28 rules of the fuzzy system with expert weights (all 1.00) and refined weights. Antecedent terms are combined with the minimum operator.

Rule	Antecedent	Consequent	Expert weight	Refined weight
R1	ODI is Normal	Normal	1.00	0.44

Rule	Antecedent	Consequent	Expert weight	Refined weight
R2	ODI is Mild	Mild	1.00	0.77
R3	ODI is Moderate	Moderate	1.00	0.17
R4	ODI is Severe	Severe	1.00	0.24
R5	ODI is Normal AND SpO_2 is Normal	Normal	1.00	0.32
R6	ODI is Normal AND T90 is Low	Normal	1.00	0.82
R7	ODI is Normal AND AI is Low	Normal	1.00	0.49
R8	SpO_2 is Normal AND T90 is Low AND AI is Low	Normal	1.00	0.89
R9	ODI is Normal AND SpO_2 is Mild	Mild	1.00	0.37
R10	ODI is Normal AND T90 is High	Mild	1.00	0.97
R11	ODI is Mild AND SpO_2 is Normal	Mild	1.00	0.54
R12	ODI is Mild AND T90 is Low	Mild	1.00	0.49
R13	ODI is Mild AND AI is Elevated	Mild	1.00	0.99
R14	ODI is Mild AND SpO_2 is Mild	Mild	1.00	0.55
R15	ODI is Moderate AND AI is Low	Mild	1.00	0.61
R16	ODI is Mild AND SpO_2 is Moderate	Moderate	1.00	0.24
R17	ODI is Mild AND T90 is High	Moderate	1.00	0.31
R18	ODI is Mild AND AI is High	Moderate	1.00	0.84
R19	ODI is Moderate AND SpO_2 is Moderate	Moderate	1.00	0.58
R20	ODI is Moderate AND SpO_2 is Mild	Moderate	1.00	0.10
R21	ODI is Moderate AND T90 is Moderate	Moderate	1.00	0.70
R22	ODI is Moderate AND AI is Elevated	Moderate	1.00	1.00
R23	ODI is Moderate AND SpO_2 is Severe	Severe	1.00	0.45
R24	ODI is Moderate AND T90 is High AND AI is High	Severe	1.00	0.67
R25	ODI is Moderate AND SpO_2 is Severe AND T90 is High	Severe	1.00	0.38
R26	SpO_2 is Severe AND T90 is High	Severe	1.00	0.77
R27	SpO_2 is Severe AND AI is High	Severe	1.00	0.85
R28	T90 is High AND AI is High	Severe	1.00	0.46

Appendix B. Machine-Learning Configurations

Table B1. Configurations selected by five-fold grid search on the training set (parameters not listed were scikit-learn defaults; k-NN, SVM and logistic regression used log(1 + x) transformation and standardisation of the inputs).

Model	Selected Configuration
Logistic regression	Logistic Regression (C=10, max_iter=2000)
k-NN	KNeighborsClassifier(n_neighbors=15)
SVM (RBF)	SVC(C=1, probability=True, random_state=0)
Random forest	RandomForestClassifier(n_estimators=300, random_state=0)

Model	Selected Configuration
Gradient boosting	GradientBoostingClassifier(learning_rate=0.05, random_state=0)

Appendix C. Dataset Dictionary

Table C1. Columns of the dataset used in simulation.

Column	Description
ID	Participant identifier (S0001–S1000)
Age	Age in years
Sex	1 = male, 0 = female
BMI	Body mass index (kg/m²)
ODI	Oxygen desaturation index (events/h)
SpO2_nadir	Lowest oxygen saturation (%)
T90	Percentage of recording below 90% saturation
Arousal index	Arousals per hour of sleep
AHI	Latent apnea–hypopnea index (events/h), used only to define the reference class
Severity class	0 = normal, 1 = mild, 2 = moderate, 3 = severe
Severity label	Text label of the severity class