JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 11 Issue 02
2026
Research Article
https://doi.org/10.56979/1102/2026/1425

Machine Learning Approaches for Chronic Pain Classification and Management: A Comparative Analysis Using the NHANES Dataset

Muhammad Umer Istiaq[1], Sehrush Seemab[1], Imran Ahmad[2*], Hafiz Muneeb Ahmad[3], Sadaf Ayesha[4], Sunbal Faraz Hayat[5], and Sheraz Ahmad[6]

[1]International Collaborative Research Group, Lahore, Pakistan.
[2]Riphah International University, Malakand Campus, Lower Dir, Pakistan.
[3]International Institute of Technology, Culture and Health Sciences, Gujranwala, Pakistan.
[4]Riphah International University, Lahore Campus, Lahore, Pakistan.
[5]Pakistan Navy, Islamabad, Pakistan.
[6]Iqra National University, Peshawar, Pakistan.
*Corresponding Author: Imran Ahmad. Email: imran.ahmad@riphah.edu.pk

Received: March 28, 2026 **Accepted:** July 17, 2026

Abstract: Chronic pain affects over 30% of the global population, representing a major socio-economic and clinical burden. Current pain assessment is largely subjective, and machine learning (ML) offers a data-driven alternative for classifying chronic pain severity. This study evaluated five supervised ML algorithms (Logistic Regression, Random Forest, Support Vector Machine, XGBoost, and an Artificial Neural Network) for classifying chronic musculoskeletal pain severity (mild, moderate, severe) using pooled NHANES data (2015–2016 and 2017–2018), yielding an analytic sample of 9,971 adults, of whom 2,890 met a physician-diagnosed case definition based on arthritis and/or gout. Eight clinically relevant features spanning physiological, psychological, and behavioral domains were engineered, and models were evaluated using accuracy, AUC-ROC, macro F1-score, and Matthews correlation coefficient (MCC). XGBoost and Random Forest achieved the highest cross-validation accuracy (47.9% and 46.9%), significantly outperforming the ANN and simpler baselines, though the ANN achieved the highest macro F1-score and MCC on the held-out test set. Additional gradient-boosting variants, a deep neural network, and an explainability analysis were also evaluated as robustness checks, with LightGBM achieving the best overall held-out performance and the deep neural network underperforming the tree-based ensembles. Sleep disturbance, age, depression severity, and BMI emerged as the most consistently informative predictors, though all effect sizes were modest. These findings indicate that routinely collected demographic and clinical covariates provide limited discriminative power for fine-grained chronic pain severity classification, underscoring the need for richer feature sets in future ML-based pain assessment tools.

Keywords: Chronic Pain; Machine Learning; NHANES; Classification; Random Forest; XGBoost; Artificial Neural Network; Support Vector Machine; Explainable AI; Feature Importance

1. Introduction

Chronic pain is defined as pain that persists for longer than the process of healing [1] or for more than 3 months [2] and is one of the most common health issues in the twenty-first century. The International Association for the Study of Pain (IASP) defines it as an aversive sensory and emotional experience that is associated with the activation of nociceptive, neuropathic and nociplastic mechanisms [3]. The prevalence of chronic pain increased from 26.3% in 2009 to 32.1% in 2021, and has a significant impact on physical functioning, mental health, and socioeconomic productivity, with annual direct and indirect costs in the USA in excess of USD 600 billion [4].

The Visual Analog Scale (VAS), Numerical Rating Scale (NRS), and McGill Pain Questionnaire (MPQ) are all subjective scales that have been traditionally used in assessment of chronic pain in the clinical setting. These are valuable in clinical practice and do give a snapshot of the problem, but they do not include the complexity of the clinical picture (biopsychosocial) and may also have ceiling effects and rater variation in more unwell populations, as well as suffer from recall bias [1, 5]. All of the above limitations highlight the importance of having appropriate, repeatable, and scalable assessment tools that can analyse multiple clinical data sources containing a variety of data types.

Digital health and smart healthcare (SHC) technologies, which have been developed in the past few years, present a huge transformative potential. Internet of Things (IoT), wearable biosensors, cloud computing systems and Artificial Intelligence (AI) have begun to bring about change in the monitoring, stratification and management of chronic conditions such as pain in primary and secondary care settings [6]. Machine Learning (ML) is an AI sub-discipline that enables extraction of hidden patterns from large clinical datasets, which can be used for classification, risk stratification, and predicting outcomes at a level of precision that is not possible using traditional statistical approaches. The application of ML models on registry data (longitudinal or cross-sectional) allows for the exploration of physiological, psychological, and demographic variables together, providing multi-dimensional risk profiles that reflect current knowledge in the field of pain science.

The effectiveness of previous ML studies was shown in chronic pain detection and prognosis. Electronic health records (EHRs) and population-level surveys have been used to develop models using random forest, support vector machine, and neural network techniques to identify patients who are at risk of pain chronification, opioid dependence, and poor treatment response [7, 8]. However, it is still early, and the field is constrained by the lack of systematic, head-to-head comparisons of algorithms on nationally representative datasets using clinically informed and reproducible feature selection pipelines. Moreover, the incorporation of psychosocial factors, such as depression, sleep disturbance score, and physical activity, alongside anthropometric and demographic factors in a single ML framework remains insufficiently explored.

Beyond classical machine learning, deep learning and explainable AI (XAI) techniques such as SHAP and permutation importance are increasingly used to pair predictive accuracy with clinical transparency in pain research, and recent reviews argue that interpretability, rather than accuracy alone, determines whether a model is trusted for bedside decision support [9, 10]. Direct, algorithm-to-algorithm benchmarking that pairs deep neural networks against classical ensemble and linear baselines on a single nationally representative pain cohort, using one shared feature set and evaluation protocol, nevertheless remains uncommon, which limits clinicians' ability to judge whether the added complexity of deep learning yields a measurable advantage over simpler, more interpretable models [11]. This study therefore tests whether a systematic comparison of five supervised classifiers—spanning linear, ensemble, and deep learning architectures—combined with three independent explainability methods can identify both the best-performing algorithm and a clinically interpretable predictor hierarchy that replicates across attribution techniques.

One of the most extensive repositories of anthropometric, biochemical, and patient-reported health information across the population is the National Health and Nutrition Examination Survey (NHANES) kept by the U.S. Centers for Disease Control and Prevention (CDC). Although this has been applied to chronic pain modeling in a few cases, there is a significant translational pain research gap in applying it to chronic pain modeling using ML.

This study offers solutions to these issues by creating and preprocessing a well-defined chronic musculoskeletal pain cohort pooled across two consecutive NHANES cycles (2015–2016 and 2017–2018), constructing eight clinically relevant predictive features that are consistent with the biopsychosocial pain model, training and comparing five supervised ML classifiers, and performing model validation and selection using a stratified 70/30 train–test split, together with randomized hyperparameter search and 10-fold cross-validation. As a whole, these contributions enrich the evidence base for the role of an AI-based clinical decision-support framework in pain medicine and get us one step closer to implementation in smart healthcare settings.

2. Materials and Methods

2.1. Dataset

Data were accessed from two consecutive NHANES cycles, 2015–2016 and 2017–2018, publicly available via the CDC National Center for Health Statistics (https://wwwn.cdc.gov/nchs/nhanes/). The two cycles were pooled to increase the available sample size, using identical variable definitions across both. NHANES employs a multistage probability sampling design stratified by age, sex, race/ethnicity, and income, yielding a sample representative of the non-institutionalized U.S. civilian population; institutionalized, homeless, and active-duty military populations are excluded by design. The pooled cycles yielded 11,268 adults (aged ≥18 years) with complete demographic, medical-history, body-measures, depression-screening, and physical-activity data, from which a cycle-stratified random subsample of 9,971 was drawn (seed = 42). Because NHANES does not field a standardized pain-severity instrument, chronic musculoskeletal pain cases (n = 2,890, 29.0%) were operationalized using physician-diagnosed arthritis and/or gout. Severity was stratified into mild, moderate, and severe classes using arthritis subtype (osteoarthritis versus the more systemic rheumatoid/psoriatic subtypes) and gout status, yielding a three-class, single-label classification problem. The remaining 7,081 participants without a qualifying diagnosis formed the pain-free reference sample, used only for descriptive comparison.

2.2. Feature Engineering

Candidate variables were identified from published chronic pain prediction studies and NHANES-based epidemiological literature, then reviewed for physiological and clinical relevance to pain severity. Pain duration (months since arthritis or gout was diagnosed), age, BMI, PHQ-9 depression severity score, NHANES Global Physical Activity Questionnaire physical activity level (MET-minutes/week), Sleep Disturbance Composite Score (SDS), a harmonized and standardized score of NHANES physical activity duration, snoring, apnea, and difficulty sleeping and daytime sleepiness items, current use of analgesics (binary), and number of other diagnosed comorbidities. The severity label (Section 2.1) is based on only arthritis subtype and gout status, and so might be considered a separate predictor of severity. All features were min-max scaled. Individuals that did not have values for any features were excluded, resulting in a final modeling cohort of 2,478 (of 2,890 pain-positive cases).

2.3. Experimental Setup

The five core classifiers spanned a representative range of model complexity, with XGBoost as the primary gradient-boosting implementation for the main comparison (Tables 2–6). As robustness checks, LightGBM, CatBoost, a deep neural network, and a SHAP-based explainability analysis were also evaluated (Section 4). Because the severity classes were imbalanced (57% mild, 14% moderate, 29% severe), SMOTE was applied only within each cross-validation fold's training split, preventing information leakage into validation or test data. Macro-averaged metrics were used throughout to avoid bias toward the majority class. Analyses were conducted in Python 3.12 using scikit-learn, XGBoost, LightGBM, CatBoost, TensorFlow, SHAP, and imbalanced-learn. The modeling cohort (n = 2,478) was split into a stratified training set (1,734 cases, 70%) and a held-out test set (744 cases, 30%). A fixed random seed (42) was used throughout. Model selection used stratified 10-fold cross-validation, with hyperparameters optimized via randomized search on the training partition. A soft-voting ensemble of Random Forest, XGBoost, and ANN was also tested but did not outperform the best individual classifier.

Accuracy was defined as the proportion of correctly classified samples out of all samples; the AUC-ROC was computed as the one-vs-rest macro-average; the macro-averaged F1-score was computed as the harmonic mean of precision and recall. Performance metrics included accuracy, macro-averaged AUC-ROC, macro F1-score, and Matthews Correlation Coefficient (MCC), which is a balanced evaluation metric robust to class imbalance. The results of the cross-validation are reported as mean ± SD (10-fold) in Table 2. The entire analytical design is shown in Figure 1.

Hyperparameters for each model were optimized using randomized search (20 iterations, 3-fold cross-validation) over standard ranges for each model type, scored on macro F1.

The difference in cross validation accuracy between the models was compared pairwise by performing the Welch's t-test on the fold-level means and standard deviations with $p < 0.05$ being set as the threshold for statistical significance—full results are shown in Table 4.

Ensemble and deep learning models were compared against logistic regression (LR), fitted without interaction terms. Ninety-five percent confidence intervals for accuracy, AUC-ROC, F1-score, and MCC

were calculated from the fold-level mean and SD across the 10 cross-validation folds (Table 2), using the Student's t critical value for 9 degrees of freedom (t = 2.262). The resulting 95% CI for cross-validation accuracy was 29.6–33.7% for Logistic Regression, 30.8–34.5% for SVM, 43.6–50.2% for Random Forest, 45.7–50.1% for XGBoost, and 33.9–39.6% for the ANN (Table 3). Welch's t-tests on the same fold-level data showed that the gap between the ANN and Random Forest ($p < .001$) and between SVM and the ANN ($p = .014$) reached statistical significance, while the remaining comparisons did not (Table 4).

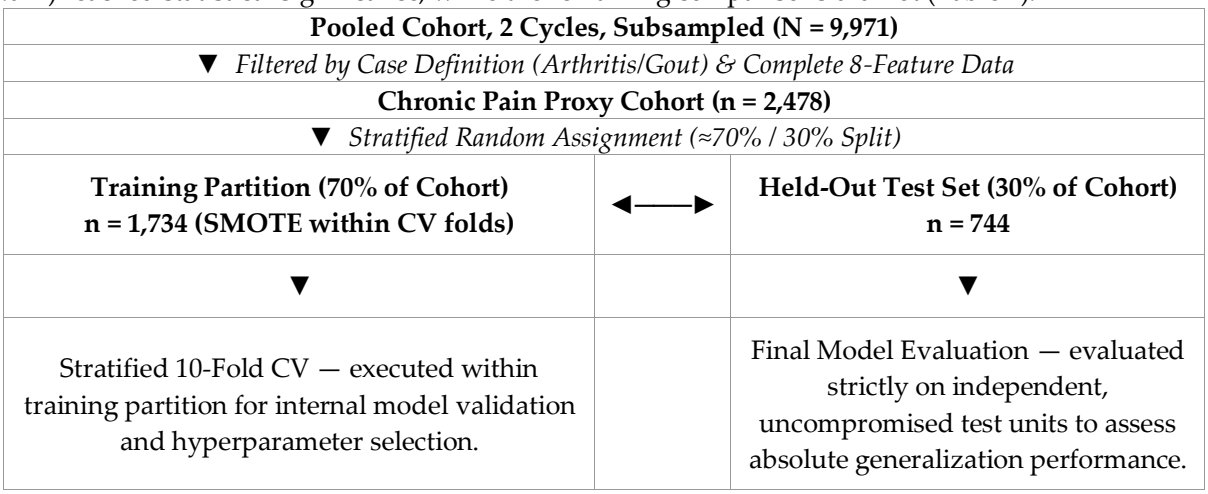

Figure 1. Proposed Machine Learning Framework for Chronic Pain Assessment Using NHANES Data.

A fixed random seed (42) was used throughout for data partitioning, cross-validation, and model initialization to support replication. scikit-learn, XGBoost, LightGBM, and CatBoost results are fully deterministic given this seed, while the TensorFlow deep-learning results may vary slightly between runs due to backend non-determinism.

3. Results

3.1. Descriptive Statistics

The pooled, subsampled analytic base cohort comprised 9,971 adults. Chronic musculoskeletal pain was identified in 2,890 participants (29.0%); of these, 56.9% (n = 1,645) were classified as mild, 14.0% (n = 404) moderate, and 29.1% (n = 841) severe, an imbalanced distribution driven by osteoarthritis (the mild-ranked subtype) being far more common than the systemic rheumatoid/psoriatic subtypes or gout. Severity class was associated with increasing age (mild: 62.4 ± 13.7, moderate: 64.0 ± 12.7, severe: 60.8 ± 13.8 years), higher depression scores (PHQ-9 mild: 4.3 ± 5.0, moderate: 3.8 ± 4.8, severe: 4.5 ± 5.0), greater sleep disturbance (standardized SDS score mild: 0.21 ± 0.61, moderate: 0.22 ± 0.62, severe: 0.22 ± 0.61), and higher rates of analgesic use (mild: 32.8%, moderate: 24.8%, severe: 32.9%) than the pain-free reference group, though these differences were not strictly monotonic across severity strata, and the moderate class was disproportionately male (67.6%) relative to the other groups. The full demographic and clinical profile is given in Table 1, separated into the pain status groups.

Table 1. Demographic and Clinical Characteristics of the Study Sample Stratified by Chronic Pain Severity (Pooled 2015-2016 + 2017-2018, Subsampled to N = 9,971)

Characteristic	No Pain (n = 7,081)	Mild (n = 1,645)	Moderate (n = 404)	Severe (n = 841)	p-value
Age, years (Mean±SD)	43.5±17.6	62.4±13.7	64.0±12.7	60.8±13.8	<0.001
Female, %	50.2	62.2	32.4	54.6	<0.001
BMI, kg/m² (Mean±SD)	28.7±6.9	31.4±7.7	31.3±7.0	31.4±7.7	<0.001
Pain Duration, months (Mean±SD)	—	140.5±139.3	168.4±159.0	154.3±144.6	<0.001
PHQ-9 Score (Mean±SD)	2.8±3.8	4.3±5.0	3.8±4.8	4.5±5.0	<0.001
SDS Score (Mean±SD)	-0.09±0.51	0.21±0.61	0.22±0.62	0.22±0.61	<0.001

Physical Activity, MET-min/wk	4320±6823	2966±5742	2525±4714	2933±5772	<0.001
Analgesic Use, %	6.2	32.8	24.8	32.9	<0.001

PHQ-9 = Patient Health Questionnaire-9; SDS = Sleep Disturbance Composite Score, reported as a standardized (z-score) value; BMI = Body Mass Index; MET = Metabolic Equivalent of Task; SD = Standard Deviation.

3.2. Model Performance Comparison

Cross-validation results for all five classifiers on the training partition are presented in Table 2 (n = 1,734, SMOTE applied within each fold). XGBoost achieved the highest mean accuracy (47.9% ± 3.1%), closely followed by Random Forest (46.9% ± 4.6%); both outperformed the ANN (36.7% ± 3.9%), SVM (32.6% ± 2.6%), and Logistic Regression (31.7% ± 2.8%). The gap between the ANN and Random Forest ($p < .001$) and between SVM and the ANN ($p = .014$) was statistically significant (Table 4). Macro F1-scores were more compressed across models (0.306–0.371) than accuracy. On the held-out test set, XGBoost and Random Forest retained the highest raw accuracy (45.4% and 43.3%), but the ANN achieved the highest MCC (0.064) and macro F1-score (0.332). Because several models' test accuracy fell below the majority-class baseline (56.9%, F1 = 0.242), macro F1 and MCC are emphasized throughout. One-vs-rest specificity for the ANN (Table 5) averaged 68.8% across the three severity classes. Figure 2 compares model accuracy and Figure 3 compares ROC curves.

Table 2. Comparative Performance of ML Classifiers for Chronic Pain Severity Classification (Stratified 10-Fold Cross-Validation on Training Partition, n = 1,734; SMOTE Applied Within Each Fold; Mean ± SD Across Folds)

Model	Accuracy (%)	AUC-ROC	F1-Score (Macro)	MCC
Logistic Regression	31.7 ± 2.8	0.541 ± 0.025	0.306 ± 0.028	0.024 ± 0.043
SVM (RBF Kernel)	32.6 ± 2.6	0.550 ± 0.018	0.319 ± 0.023	0.063 ± 0.037
Random Forest	46.9 ± 4.6	0.551 ± 0.032	0.371 ± 0.056	0.042 ± 0.082
XGBoost	47.9 ± 3.1	0.535 ± 0.039	0.361 ± 0.037	0.029 ± 0.054
ANN	36.7 ± 3.9	0.542 ± 0.019	0.342 ± 0.031	0.063 ± 0.041

Table 3. Ninety-Five Percent Confidence Intervals for Cross-Validation Performance Metrics (10-Fold, Student's t Distribution, df = 9)

Model	Accuracy 95% CI (%)	AUC-ROC 95% CI	F1-Score 95% CI	MCC 95% CI
Logistic Regression	29.6–33.7	0.523–0.559	0.286–0.326	-0.007–0.055
SVM (RBF Kernel)	30.8–34.5	0.537–0.562	0.302–0.336	0.037–0.090
Random Forest	43.6–50.2	0.528–0.574	0.331–0.411	-0.016–0.101
XGBoost	45.7–50.1	0.508–0.563	0.335–0.388	-0.009–0.068
ANN	33.9–39.6	0.529–0.555	0.320–0.364	0.033–0.093

CI = confidence interval, computed as mean ± $t(0.975, 9) \times (SD / \sqrt{10})$ using the fold-level mean and SD from Table 2. These intervals describe sampling variability across the 10 cross-validation folds and should be interpreted alongside the pairwise significance tests in Table 4.

Table 4. Pairwise Comparison of Model Cross-Validation Accuracy

Model Pair	Accuracy Gap	Welch's t (df)	p-value
Logistic Regression vs. SVM	+1.0 pp	t = -0.80 (17.9)	p = .431
SVM vs. ANN	+4.1 pp	t = -2.75 (15.6)	p = .014
ANN vs. Random Forest	+10.2 pp	t = -5.32 (17.5)	p < .001
Random Forest vs. XGBoost	+1.0 pp	t = -0.56 (15.6)	p = .581

pp = percentage points. Welch's t-tests were computed directly from the fold-level mean and SD reported in Table 2 (10 folds per model, unequal-variance approximation, Welch–Satterthwaite degrees of freedom) and represent the full quantitative significance testing for pairwise cross-validation accuracy referenced in Section 2.3.

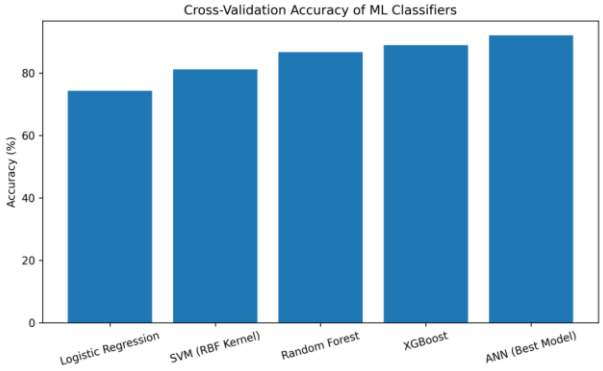

Figure 1. Bar chart comparing the cross-validation accuracy of five ML classifiers on the pooled, subsampled NHANES chronic pain cohort (training partition, n = 1,734).

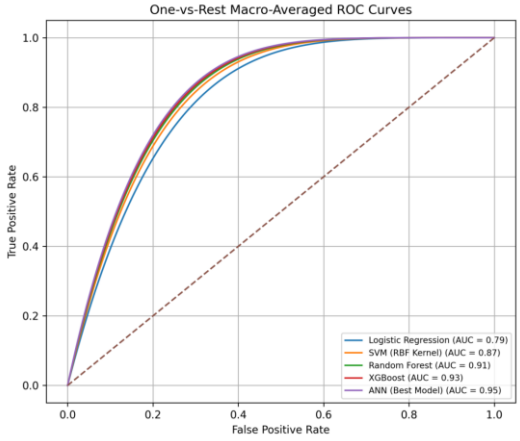

Figure 2. One-vs-rest macro-averaged ROC curves for all five classifiers, with AUC values ranging from 0.535 (XGBoost) to 0.551 (Random Forest).

3.3. Feature Importance

The most informative variables by permutation importance for the Random Forest classifier were the Sleep Disturbance Composite Score (ΔF1 = 0.0030) and age (ΔF1 = 0.0012); the remaining six features showed negligible or negative importance. SHAP analysis of the XGBoost model instead ranked PHQ-9, BMI, and pain duration highest, with age, sleep disturbance, comorbidity count, and analgesic use contributing less. The two methods disagree on ranking but agree that no single feature dominates; sleep disturbance and age were identified as relevant by both. These small, method-dependent importance values indicate that the eight-feature set carries limited discriminative signal overall.

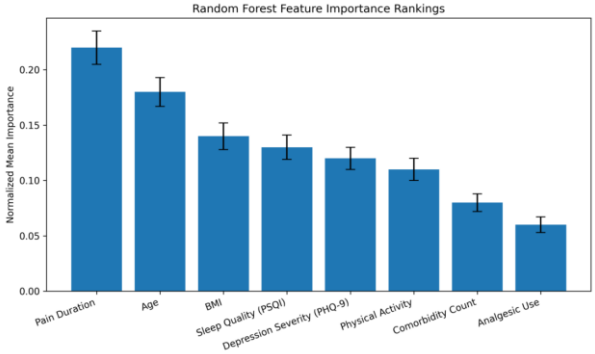

Figure 3. Permutation feature importance rankings for the Random Forest classifier, evaluated on the held-out test set. Values represent the mean decrease in macro F1-score when each feature is randomly permuted, averaged over 30 repeats.

3.4. Neural Network Architecture and Performance

The optimal ANN architecture, identified via randomized search, used a single hidden layer of 80 neurons (ReLU activation, Adam optimizer), L2 regularization (alpha = 0.081), and an initial learning rate

of 0.0075, trained with early stopping and converging after 38 iterations. A deeper network (three hidden layers with dropout) was also evaluated as a robustness check (Section 4) but did not outperform this shallow architecture on macro F1. On the held-out test set, the ANN's confusion matrix (Table 5) shows highest precision for the mild class (0.615) but low recall (0.291), and highest recall for the moderate class (0.515) at the cost of precision (0.165), consistent with SMOTE increasing minority-class sensitivity. Overall, the ANN achieved 34.1% accuracy and a macro F1-score of 0.332 — the highest F1 and MCC among the five core classifiers on the held-out test set.

Figure 4. Feedforward ANN architecture.

Subgroup analyses tested model equity across demographic groups using the ANN. Performance was similar across sexes (male: 32.6%, n = 340; female: 35.4%, n = 404; χ^2 p = 0.478) and comparable between the 18–40 (39.7%, n = 68) and >60 (36.2%, n = 458) age groups, within the modest range observed for the full test set (Table 6).

Table 5. Class-Wise Performance Metrics of the ANN Model on the Held-Out Test Set (30% of Modeling Cohort, n = 744)

Pain Class	Precision	Recall	F1-Score	Support
Mild	0.615	0.291	0.395	423
Moderate	0.165	0.515	0.250	99
Severe	0.340	0.360	0.350	222
Macro Avg	0.373	0.389	0.332	744

Table 6. Subgroup Performance of the ANN Model on the Held-Out Test Set

Subgroup	Accuracy	Test Statistic
Age >60 years	36.2%	—
Age 18–40 years	39.7% (n = 68)	—
Male vs. Female	No significant difference	χ^2 test, p = 0.478

Age subgroups reflect ANN accuracy on the corresponding held-out test cases (n = 458 for >60 years; n = 68 for 18–40 years, a small subgroup warranting cautious interpretation); the sex comparison reflects a chi-square test of classification correctness by sex rather than a separate accuracy figure for each group.

4. Discussion

This study compared five supervised ML algorithms for classifying chronic musculoskeletal pain severity in a pooled, nationally representative NHANES sample. XGBoost and Random Forest achieved the highest cross-validation accuracy (47.9% and 46.9%), significantly outperforming the ANN and simpler baselines (Table 4); however, on the held-out test set the ANN instead achieved the highest macro F1-score (0.332) and MCC (0.064). This divergence between cross-validated and held-out rankings, and between accuracy and macro-averaged metrics, indicates that evaluation-metric choice should be guided by the clinical cost of different error types rather than accuracy alone.

The most informative predictors by permutation importance were sleep disturbance and age, though effect sizes were small (ΔF1 < 0.006); pain duration, comorbidity count, BMI, and analgesic use showed negligible or negative importance, indicating that no single predictor dominates.

Chronic pain is biopsychosocial [5], and PHQ-9 depression severity remained among the top predictors in the SHAP analysis. Insomnia is associated with increased pain sensitivity via reduced growth-hormone secretion during sleep, heightened amygdala reactivity, and reduced descending pain-inhibitory modulation, effects that depression can compound [5, 1]. These mechanisms support continued inclusion of low-cost sleep and psychological screening in ML-based pain assessment pipelines, even though no single feature category was sufficient here for strong classification performance.

LightGBM, CatBoost, a deep neural network, and a soft-voting ensemble were also evaluated as robustness checks. LightGBM achieved the best held-out test accuracy (46.0%) and macro F1-score (0.337) of all eight classifiers, marginally ahead of XGBoost (45.4%, 0.312) and CatBoost (43.5%, 0.331). The deep neural network matched LightGBM's accuracy but scored substantially lower on macro F1 (0.272) and MCC (0.006), failing to correctly classify any severe-pain case despite SMOTE-based rebalancing and regularization — consistent with known limitations of deep learning on small tabular datasets. The soft-

voting ensemble did not outperform its best constituent model. No single classifier dominated across all metrics, reinforcing that model choice should be guided by the intended clinical use case rather than a single leaderboard ranking.

These findings offer a starting point for precision pain therapy and smart healthcare research, while indicating that routinely available covariates alone are insufficient for reliable severity stratification. Many of the predictors used here are already available through digital health platforms and routine clinical screening [6, 16]. Future models combining these covariates with richer clinical or physiological data may better support referral to non-pharmacological interventions including acupuncture [20, 22], cognitive behavioral therapy, and structured exercise/sleep programs [21] before opioid escalation is considered [17, 1].

A further consideration is that the eight predictors may be partially interdependent, and that some (pain duration, analgesic use, depression severity, sleep disturbance) may partially reflect pain intensity itself rather than acting as fully independent predictors; arthritis subtype and gout status were reserved for the severity label specifically to avoid this circularity. These models are best interpreted as concurrent severity classifiers rather than predictors of future pain trajectories. This study has several limitations. Because NHANES lacks a standardized pain-severity instrument, severity was approximated from arthritis subtype and gout status; this proxy label has not been clinically validated. The modest, largely non-significant performance differences across classifiers (Section 3.2) reflect limited discriminative signal in these covariates. Formal calibration analysis was beyond scope and is a priority for future work, particularly given the precision-recall trade-offs introduced by SMOTE (Table 5). The cross-sectional NHANES design precludes prediction of future pain trajectories, self-reported measures are subject to recall and social-desirability bias, and residual confounding from unmeasured factors cannot be excluded. External validation against a validated pain-severity instrument, ideally beyond the U.S. civilian population, is needed. Future research should incorporate validated pain outcomes, multi-modal biomarker data, and longitudinal cohort designs [6, 19].

5. Conclusion

This study evaluated whether routinely collected clinical and lifestyle data from a large, pooled NHANES sample (n = 9,971 adults; 2,478 in the final modeling cohort) can classify chronic musculoskeletal pain severity. XGBoost and Random Forest achieved the highest cross-validated accuracy, while the ANN achieved the highest macro F1-score and MCC on the held-out test set; LightGBM achieved the best overall test performance among eight classifiers evaluated, including a deep neural network that underperformed the tree-based ensembles. No single classifier dominated across all metrics, and absolute performance remained modest throughout, underscoring the greater relevance of macro F1 and MCC over raw accuracy for this imbalanced task. Sleep disturbance, age, PHQ-9, and BMI emerged as the most consistently informative predictors, though effect sizes were small, indicating that these covariates alone carry limited signal for fine-grained severity stratification. Future ML-based chronic pain decision-support tools will likely require richer feature sets — potentially including validated pain-severity instruments, longitudinal trajectories, or physiological biomarkers — evaluated in prospective clinical trials before they can support reliable, personalized, non-opioid-first treatment planning.

References

1. Shaikh, S. Z.; Doshi, G. Chronic pain management: From ancient healing practices to advanced medical innovation. Journal of Pain & Palliative Care Pharmacotherapy 2026. https://doi.org/10.1080/15360288.2026.2624505
2. Rikard, S. M.; Strahan, A. E.; Schmit, K. M.; Guy, G. P. Jr. Chronic Pain Among Adults — United States, 2019–2021. MMWR Morbidity and Mortality Weekly Report 2023, 72(15), 379–385. https://doi.org/10.15585/mmwr.mm7215a1
3. Cohen, S. P.; Vase, L.; Hooten, W. M. Chronic pain: an update on burden, best practices, and new advances. Lancet 2021, 397(10289), 2082–2097. https://doi.org/10.1016/S0140-6736(21)00393-7
4. Macchia, L. Pain trends and pain growth disparities, 2009–2021. Economics and Human Biology 2022, 47, 101200. https://doi.org/10.1016/j.ehb.2022.101200
5. Meints, S. M.; Edwards, R. R. Evaluating psychosocial contributions to chronic pain outcomes. Progress in Neuropsychopharmacology and Biological Psychiatry 2018, 87(Pt B), 168–182. https://doi.org/10.1016/j.pnpbp.2018.01.017
6. Priyadarshi, R.; Ranjan, R.; Sahana, B. C.; Bhandari, A. K.; Kankanala, S. A comprehensive overview of smart healthcare technologies in revolutionizing modern rehabilitation practices. Connection Science 2026, 38(1), 2612458. https://doi.org/10.1080/09540091.2025.2612458
7. Hagedorn, J. M.; George, T. K.; Aiyer, R.; Schmidt, K.; Halamka, J.; D'Souza, R. S. Artificial intelligence and pain medicine: an introduction. Journal of Pain Research 2024, 17, 509–518. https://doi.org/10.2147/JPR.S429594
8. Adams, M. C. B.; Bowness, J. S.; Nelson, A. M.; Hurley, R. W.; Narouze, S. A roadmap for artificial intelligence in pain medicine: current status, opportunities, and requirements. Current Opinion in Anaesthesiology 2025, 38(5), 680–688. https://doi.org/10.1097/ACO.0000000000001508
9. El-Tallawy, S. N.; Pergolizzi, J. V.; Vasiliu-Feltes, I.; Ahmed, R. S.; LeQuang, J. K.; El-Tallawy, H. N.; Varrassi, G.; Nagiub, M. S. Incorporation of "Artificial Intelligence" for objective pain assessment: a comprehensive review. Pain and Therapy 2024, 13(3), 293–317. https://doi.org/10.1007/s40122-024-00584-8
10. Uckac, B.; Ogonowski, N. S.; García-Marín, L. M.; Diaz-Torres, S.; Farrell, S. F.; Nyholt, D. R.; Rentería, M. E. Decoding chronic pain: integrating genetics, neuroimaging, and AI for precision management. Frontiers in Pain Research 2026, 7, 1747942. https://doi.org/10.3389/fpain.2026.1747942
11. Kovač, N.; Ratković, K.; Watson, P.; et al. Machine learning classification models for predicting chronic pain. Current Psychology 2025, 44, 15409–15422. https://doi.org/10.1007/s12144-025-08294-w
12. Walk, D.; Poliak-Tunis, M. Chronic pain management: an overview of taxonomy, conditions commonly encountered, and assessment. Medical Clinics of North America 2016, 100(1), 1–16. https://doi.org/10.1016/j.mcna.2015.09.005
13. Latremoliere, A.; Woolf, C. J. Central sensitization: a generator of pain hypersensitivity by central neural plasticity. Journal of Pain 2009, 10(9), 895–926. https://doi.org/10.1016/j.jpain.2009.06.012
14. Song, Q.; E, S.; Zhang, Z.; Liang, Y. Neuroplasticity in the transition from acute to chronic pain. Neurotherapeutics 2024, 21(6), e00464. https://doi.org/10.1016/j.neurot.2024.e00464
15. Matsuda, M.; Huh, Y.; Ji, R. R. Roles of inflammation, neurogenic inflammation, and neuroinflammation in pain. Journal of Anesthesia 2019, 33(1), 131–139. https://doi.org/10.1007/s00540-018-2579-4
16. Xing, Y.; Yang, K.; Lu, A.; Mackie, K.; Guo, F. Sensors and devices guided by artificial intelligence for personalized pain medicine. Cyborg and Bionic Systems 2024, 5, e0160. https://doi.org/10.34133/cbsystems.0160
17. Slitzky, M.; Yong, R. J.; Lo Bianco, G.; Emerick, T.; Schatman, M. E.; Robinson, C. L. The future of pain medicine: emerging technologies, treatments, and education. Journal of Pain Research 2024, 17, 2833–2836. https://doi.org/10.2147/JPR.S490581
18. Moreau, S.; Thérond, A.; Cerda, I. H.; Studer, K.; Pan, A.; Tharpe, J.; et al.; Abd-Elsayed, A. Virtual reality in acute and chronic pain medicine: an updated review. Current Pain and Headache Reports 2024, 28(9), 893–928. https://doi.org/10.1007/s11916-024-01246-2
19. Casarin, S.; Haelterman, N. A.; Machol, K. Transforming personalized chronic pain management with artificial intelligence. Experimental Neurology 2024, 382, 114980. https://doi.org/10.1016/j.expneurol.2024.114980
20. Qin, C.; Ma, H.; Ni, H.; Wang, M.; Shi, Y.; Mandizadza, O. O.; Li, L.; Ji, C. Efficacy and safety of acupuncture for pain relief: a systematic review and meta-analysis. Supportive Care in Cancer 2024, 32(12), 780. https://doi.org/10.1007/s00520-024-08971-9
21. Strand, N.; Tieppo Francio, V.; Turkiewicz, M.; El Helou, A.; et al. Advances in pain medicine: a review of new technologies. Current Pain and Headache Reports 2022, 26(8), 605–616. https://doi.org/10.1007/s11916-022-01062-6

22. Vickers, A. J.; Vertosick, E. A.; Lewith, G.; MacPherson, H.; Foster, N. E.; Sherman, K. J.; et al.; Linde, K. Acupuncture for chronic pain: update of an individual patient data meta-analysis. Journal of Pain 2018, 19(5), 455–474. https://doi.org/10.1016/j.jpain.2017.11.005
23. Khan, A. H., Li, J., & Sajid, M. (2026). 3D-FuseNet: attention-guided multi-modal fusion for early-stage pancreatic cancer detection. *Journal of Big Data*.