

# Similarity-Based Target Leakage in Case-Based Reasoning for Stress Risk Screening: A Cautionary Demonstration

Jonas Niyitegeka<sup>1</sup>, Anaclet Ukurikiyeyezu<sup>2</sup>, Shaloom Niyibizi<sup>3</sup>

<sup>1</sup>Kigali Independent University, Kigali, Rwanda.

<sup>2</sup>African Institute for Mathematical Sciences (AIMS), Kigali, Rwanda.

<sup>3</sup>Kigali Independent University, Kigali, Rwanda.

DOI: <https://doi.org/10.51244/IJRSI.2026.1309000049>

Received: 09 September 2026; Accepted: 14 September 2026; Published: 03 October 2026

## ABSTRACT

Early identification of elevated stress may support preventive mental health interventions, particularly in low-resource settings. This study evaluates a Case-Based Reasoning (CBR) approach alongside conventional Machine Learning (ML) classifiers for stress risk screening using an anonymized COVID-19 quarantine survey dataset ( $N = 824$ ) containing twelve categorical predictors and three approximately balanced Growing Stress classes. Six supervised learning algorithms (Random Forest,  $k$ -Nearest Neighbours, XGBoost, Decision Tree, Support Vector Machine, and Logistic Regression) were evaluated using stratified cross-validation, while a CBR model based on Jaccard similarity and mode-based revision was assessed using a stratified hold-out protocol. Under a leakage-free experimental design restricting similarity computation to predictor attributes only, both CBR and ML models achieved performance close to random classification (accuracy 0.33–0.38; macro-AUC  $\approx 0.50$ ), indicating limited predictive information in the available behavioural features. A controlled ablation experiment intentionally including the target attribute in the CBR similarity computation increased accuracy to 0.94–0.99 and AUC to 1.00, without changing the dataset, model structure, or evaluation procedure. This demonstrates that similarity-based retrieval can be vulnerable to target leakage when solution attributes are unintentionally incorporated during retrieval, producing misleadingly high performance estimates. The contribution of this study is twofold: it provides an empirical finding that the available behavioural features offered insufficient predictive information for stress classification in the study dataset, while also presenting a reproducible demonstration of how target leakage can be identified and prevented in similarity-based retrieval systems. These findings highlight the importance of explicit predictor–target separation and rigorous validation protocols for developing trustworthy AI-based decision-support systems in mental health applications. An additional identical-fold cross-validation, a majority-class baseline, and label-permutation null distributions further confirm that neither CBR nor the machine learning baselines exceed chance-level performance, and that the leakage-driven accuracy gain persists unchanged even when the leaked label is replaced with an arbitrary, permuted one, confirming the effect is a structural retrieval artefact rather than a genuine relationship in the data.

**Keywords:** Case-Based Reasoning, Machine Learning, Growing Stress, Target Leakage, COVID-19 Quarantine Survey, Classification, Decision Support Systems

---

<sup>1</sup> Kigali Independent University, Rwanda. [niyitegekajonas@ulk.ac.rw](mailto:niyitegekajonas@ulk.ac.rw)

<sup>2</sup> African Institute for Mathematical Sciences (AIMS), Rwanda. [anaclet.ukurikiyeyezu@aims.ac.rw](mailto:anaclet.ukurikiyeyezu@aims.ac.rw)

<sup>3</sup> Kigali Independent University, Rwanda. [shaloomniyibizi@ulk.ac.rw](mailto:shaloomniyibizi@ulk.ac.rw)

## INTRODUCTION

Mental health disorders represent a major global public health challenge, with depression being one of the leading causes of disability worldwide. Despite this burden, access to appropriate mental health services remains limited, particularly in low-resource settings where shortages of trained professionals and limited screening capacity constrain early identification and intervention [1]. Stress is an important modifiable factor associated with increased vulnerability to depression and other mental health conditions [2]. Therefore, early identification of elevated stress may support preventive interventions by identifying individuals who may benefit from further assessment. Recent advances in artificial intelligence (AI) have enabled the analysis of behavioural, digital, and self-reported data for mental health-related risk screening and decision support [3]. However, the reliability of these systems depends on rigorous validation procedures and careful control of data quality issues. Target leakage, where information from the outcome variable unintentionally influences model development or prediction can lead to artificially inflated performance estimates and reduce confidence in AI-based decision-support systems [4]. While target leakage in conventional machine learning pipelines has been widely studied, leakage arising from similarity computation in retrieval-based systems, particularly Case-Based Reasoning (CBR) frameworks, has received comparatively less attention. CBR systems solve new problems by retrieving similar previously encountered cases and adapting their solutions to new situations [5]. Because CBR performance depends strongly on case representation and similarity measurement, inappropriate inclusion of outcome-related attributes during retrieval may allow the system to exploit label information rather than genuine relationships among predictor variables [6]. This study addresses this gap by evaluating a Case-Based Reasoning (CBR) approach alongside conventional Machine Learning (ML) classifiers for growing stress risk screening under a strict predictor-target separation protocol. Furthermore, a controlled ablation experiment is conducted to systematically investigate how inclusion of the target attribute during similarity computation affects reported performance. Although the evaluated models did not demonstrate sufficient predictive capability for stress classification using the available dataset, the findings provide two important contributions: empirical evidence regarding the limitations of the available behavioural features for stress prediction in this dataset and a methodological demonstration of how target leakage can arise and be prevented in similarity-based retrieval systems for mental health risk screening.

## LITERATURE REVIEW

### Machine Learning for Mental Health Screening

Machine learning (ML) methods have been increasingly applied in mental health research to support risk assessment and early identification of psychological conditions. Previous studies have explored diverse data sources, including social media data [7], behavioural information from wearable devices [8], and self-reported survey data [9], using algorithms such as Support Vector Machines, Random Forest, Logistic Regression, and neural networks. However, reported predictive performance varies considerably across studies due to differences in datasets, feature representations, validation procedures, and outcome definitions. This variability highlights the importance of rigorous evaluation protocols to ensure that reported performance reflects genuine predictive capability rather than methodological artefacts. Applications specifically targeting stress detection have used similar machine learning approaches [10]; however, many studies remain limited by the absence of external validation or clinically validated reference outcomes.

### Case-Based Reasoning in Healthcare

Case-Based Reasoning (CBR) is an artificial intelligence approach that solves new problems by retrieving similar previously encountered cases and adapting their solutions to new situations [5]. CBR has been applied in healthcare domains, including diagnosis [7], clinical decision support, and risk assessment, because of its ability to provide explanations based on previous experiences. Unlike conventional machine learning models, CBR performance depends strongly on case representation, similarity measures, and retrieval strategies. Consequently, inappropriate design of case representations may affect both interpretability and predictive reliability. Although CBR-based approaches have been explored for healthcare decision support, including mental health-related applications [11], validation protocols and leakage control remain important considerations.

## Target Leakage and Its Detection

Data leakage occurs when information from the target variable or the evaluation process unintentionally influences model development or prediction, resulting in overly optimistic performance estimates [12]. In conventional machine learning pipelines, leakage may occur through several mechanisms, including preprocessing leakage, where scaling, imputation, or feature selection is performed before data splitting; cross-validation leakage, where information from validation folds influences model selection; and temporal leakage, where future information is used to predict past events [13]. In similarity-based retrieval systems, a distinct form of leakage can occur when the target attribute is included among the case attributes used for similarity computation. This situation is conceptually analogous to a k-nearest neighbours classifier where the class label itself is included as a feature: the similarity function may preferentially retrieve cases with identical labels [14], artificially inflating predictive performance. Although this vulnerability follows directly from the principles of target leakage and supervised learning, its empirical consequences in CBR-based mental health risk screening have received limited investigation. This study therefore provides a controlled evaluation of how target inclusion during similarity computation affects CBR performance and demonstrates the importance of strict predictor-target separation in retrieval-based AI systems.

## MATERIALS AND METHODS

### Dataset

This secondary analysis used a publicly available, anonymized self-report survey dataset collected during the COVID-19 quarantine period. The data were obtained from the public Kaggle mirror "RHMCD-20|Depression" (imtkaggleteam), which reproduces the RHMCD-20 corpus originally released by its collectors on Mendeley Data (DOI: 10.17632/pxjmjyfdh2) and described in the peer-reviewed source publication [17]. The analytic sample comprised 824 complete observations with twelve categorical predictor variables and one target variable, Growing\_Stress, which captured self-reported perceived stress progression (Yes: 301 [36.5%]; Maybe: 267 [32.4%]; No: 256 [31.1%]). The variables Age and Days\_Indoors were treated as ordinal categorical variables. No missing values were present, so no imputation was required. For machine learning models, categorical variables were one-hot encoded using scikit-learn's OneHotEncoder, yielding 39 binary features; the case-based reasoning framework retained the original categorical attribute-value representation. Given the approximately balanced class distribution, accuracy was used as the primary performance metric, with a random baseline of approximately 33.3%. The dataset is publicly available, anonymized, and free of personally identifiable information; this secondary analysis involved no direct interaction with human participants. The corpus was collected via a structured online and in-person survey of 825 respondents (82.5% response rate from 1,000 invited participants) from teenagers, college students, housewives, and corporate professionals in Khilgaon, Uttara, and Bashundhara, Dhaka, Bangladesh, during COVID-19 quarantine. The target variable corresponds to the single survey item "Is your stress increasing day by day?" with responses No, Maybe, or Yes. Programmatic verification confirmed that the 824-record analytic file matches the source publication's 13 variable names exactly, with one fewer record than the 825 respondents reported in the source, consistent with use of the publicly distributed filtered release. The dataset is distributed under a CC BY-NC 4.0 license. In the source study, ethical approval was obtained from the Daffodil International Student Research Board and informed consent was obtained from all participants; no additional ethical approval was sought for this secondary reanalysis. No duplicate respondent records were identified: an exact row-level comparison across all thirteen variables returned zero duplicates among the 824 records.

**Table 1: Dataset features and target variable ( $N = 824$ ). All variables are categorical.**

| Variable                | #Levels | Categories                                               |
|-------------------------|---------|----------------------------------------------------------|
| Age                     | 4       | 16–20, 20–25, 25–30, 30 and above                        |
| Gender                  | 2       | Female, Male                                             |
| Occupation              | 5       | Corporate, Student, Business, Housewife, Others          |
| Days_Indoors            | 5       | Go out every day, 1–14, 15–30, 31–60, More than 2 months |
| Quarantine Frustrations | 3       | Yes, Maybe, No                                           |

|                         |   |                   |
|-------------------------|---|-------------------|
| Changes Habits          | 3 | Yes, Maybe, No    |
| Mental Health History   | 3 | Yes, Maybe, No    |
| Weight Change           | 3 | Yes, Maybe, No    |
| Mood Swings             | 3 | Low, Medium, High |
| Coping Struggles        | 2 | Yes, No           |
| Work Interest           | 3 | Yes, Maybe, No    |
| Social Weakness         | 3 | Yes, Maybe, No    |
| Growing Stress (Target) | 3 | Yes, Maybe, No    |

## Case-Based Reasoning Framework

Case-Based Reasoning (CBR) is an artificial intelligence approach that solves new problems by retrieving similar previously encountered cases and adapting their solutions to new situations. The CBR process follows four main phases: retrieve, reuse, revise, and retain [5] (Figure 1). In this study, each case represents one survey record, and a new case is classified by comparing its similarity with previously observed cases.

### Case representation and similarity

Since all variables in the dataset are categorical, each case was represented as a collection of attribute-value pairs. A case containing three attributes,

(Age = 20-25, Gender = Female, Mood\_Swings = High),

was represented as the set:

$$\{\text{Age}_{20-25}, \text{Gender}_{\text{Female}}, \text{Mood\_Swings}_{\text{High}}\}.$$

The similarity between two cases was initially defined using the Jaccard coefficient:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}, \quad (1)$$

Where A and B represent the sets of attribute-value pairs corresponding to two cases. For single-valued categorical attributes, Jaccard similarity equals an exact-match comparison. A shared attribute-value pair contributes to the intersection, while different attribute values contribute only to the union.

For the categorical representation used in this study, similarity was measured using the Simple Matching Coefficient (SMC) [18], which quantifies the proportion of attributes that have matching values between two cases. The SMC was used instead of the Jaccard index defined in Equation (1), because the two measures are related but are not numerically equivalent. Unlike the Jaccard index, which excludes joint absences from the similarity calculation, the SMC considers both matching presences and matching absences. The similarity between a query case and a stored case was therefore calculated using the Simple Matching Coefficient as follows:

$$S_{\text{global}}(q, c) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(v_i^q = v_i^c), \quad (2)$$

Where n represents the number of predictor attributes,  $v_i^q$  and  $v_i^c$  denote the values of the  $i$ th attribute in the query and stored cases, respectively, and  $\mathbb{I}(\cdot)$  is the indicator function defined as:

$$\mathbb{I}(v_i^q = v_i^c) = \begin{cases} 1, & \text{if } v_i^q = v_i^c, \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

This Simple Matching Coefficient is related to, but not equivalent to, the Jaccard index of Eq. (1): for two cases compared on  $n$  predictor attributes with  $m$  matching values,  $SMC = m/n$  whereas  $Jaccard = m/(2n - m)$ ; the two coincide only when  $m = 0$  or  $m = n$ . Because  $n$  is fixed across all case comparisons in this study, SMC is a strictly monotonic (rank-preserving) transform of Jaccard for every pair of cases, so the two measures induce identical neighbour orderings, identical top- $k$  retrieval sets, and identical similarity ties under the unweighted retrieval scheme used here (see the tie-break policy in Sec. 3.2.2). This correction therefore changes only the reported similarity values and their labelling, not any previously reported retrieval outcome or accuracy figure; it has no bearing on weighting (no attribute weighting is used) or on similarity thresholds (retrieval is rank-based with a fixed  $k$ , not threshold-based). Thus,  $S_{global}(q, c) = 1$  indicates complete agreement across all predictor attributes, whereas values closer to zero indicate lower similarity. Only predictor attributes were included in the similarity computation, while the target variable Growing Stress was excluded to prevent target leakage and ensure a valid evaluation. In the leakage ablation of Sec. 4.4, Growing Stress is reintroduced into this same attribute set using an identical nominal (string-equality) representation to the predictors, with no special encoding - i.e., a training case's own label is compared to the query's unknown label with exactly the same equality test used for every other attribute.

### Reuse and revise

For each query case, the  $k$  most similar cases were retrieved from the case base. The reuse phase obtains information from retrieved cases, while the revision phase determines the final prediction using majority voting among the retrieved labels:

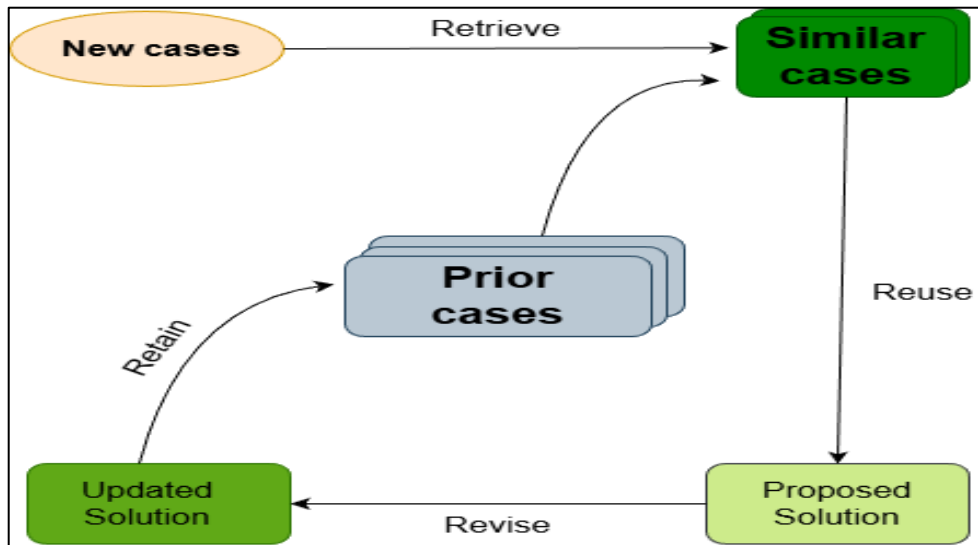
$$\hat{y}(q) = \text{mode}\{y(c_1), \dots, y(c_k)\}. \quad (4)$$

This revision strategy reduces the influence of individual retrieved cases and improves prediction stability by combining information from multiple similar cases. Two tie-break rules complete the retrieval and revision procedure and are fixed for every experiment reported in this study. First, when two or more stored cases are equally similar to a query (a similarity tie), the case that appears earliest in the case base (i.e. the lowest original row index) is ranked higher; ties therefore never depend on iteration or hashing order. Second, when the majority vote among the  $k$  retrieved labels is itself tied across two or more classes, the tie is broken in favour of the class whose retrieved members have the greater summed similarity to the query within the top- $k$  neighbourhood; any residual tie is broken by ascending class order. Both rules are deterministic and were applied identically in the leak-free and leakage-ablation conditions reported in Sec. 4.3 and Sec. 4.4.

### Retain

The retain phase updates the case base by adding validated new cases for future predictions. In this study, retention was performed only after model evaluation, ensuring that test cases were not added to the case base before prediction. This procedure prevents information leakage and maintains the validity of the reported performance results.

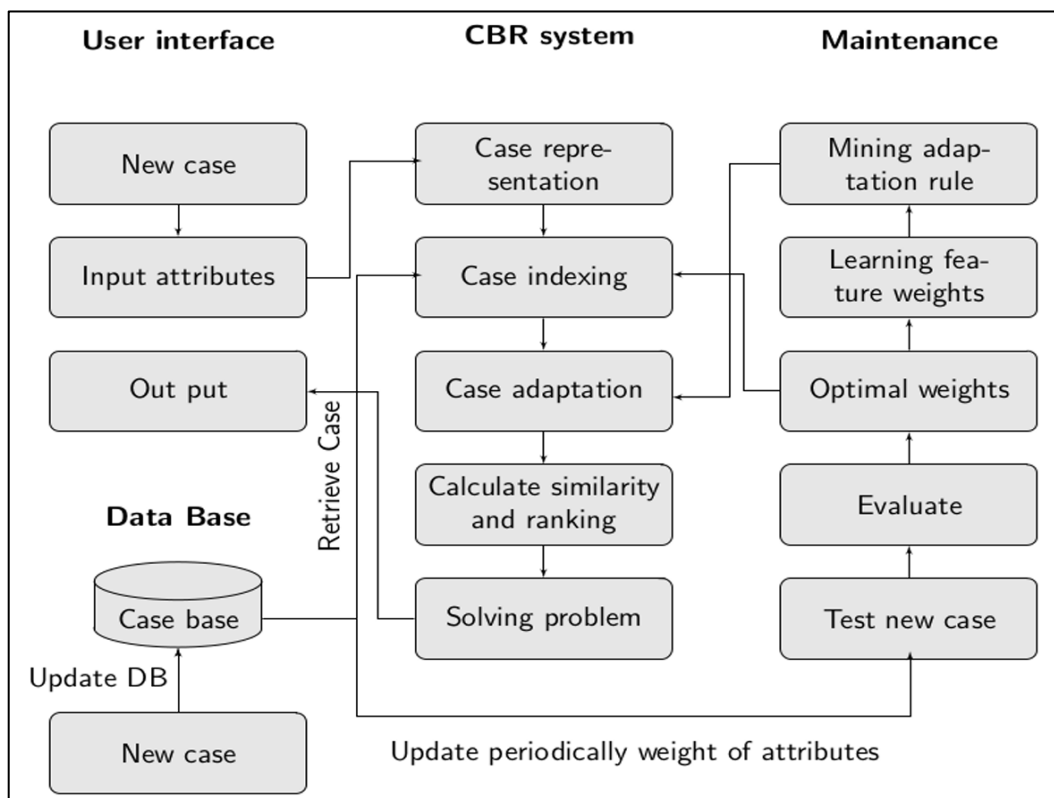
**Figure 1: The Case-Based Reasoning cycle.**



### Proposed System Architecture

Figure 2 illustrates the Case-Based Reasoning (CBR) framework evaluated in this study for growing stress risk classification. The framework consists of four main components: case representation, similarity-based retrieval, solution reuse and revision, and case-base management. A new query case is represented using its predictor attribute–value pairs and compared with previously stored cases using the similarity measure defined in Equation (1). The  $k$  most similar cases are retrieved, and their associated class labels are combined through majority voting to assign the Growing\_Stress class to the query case. Validated cases may subsequently be retained in the case base for future use. During evaluation, test cases were kept separate from the case base during classification to prevent information leakage. The architecture therefore represents the CBR workflow evaluated under the leakage-free protocol adopted in this study.

**Figure 2: Architecture of the CBR framework for growing stress risk classification.**



## Machine Learning Baselines and Evaluation Protocol

Six supervised machine learning classifiers were evaluated using the one-hot encoded predictor variables: Random Forest,  $k$ -Nearest Neighbours ( $k = 10$ ), XGBoost, Decision Tree, Support Vector Machine (RBF kernel), and Logistic Regression. The models were implemented using the scikit-learn and xgboost libraries with a fixed random seed of 45 to ensure reproducibility. Default hyperparameters were used unless otherwise specified, with the exception of the neighbourhood size for the  $k$ -Nearest Neighbours classifier. The machine learning models were evaluated using stratified 5-fold cross-validation, while 20-fold cross-validation was additionally performed to assess the stability of the obtained results. The CBR framework was evaluated using stratified hold-out validation across different test-set proportions and neighbourhood sizes. Hold-out evaluation was adopted for CBR because the retrieval process requires a fixed case base during prediction, making it straightforward to preserve a strict separation between stored cases and query cases while avoiding target leakage. Performance was additionally assessed using macro-averaged precision, recall,  $F_1$ -score, and one-versus-rest ROC-AUC to provide a balanced evaluation across the three classes.

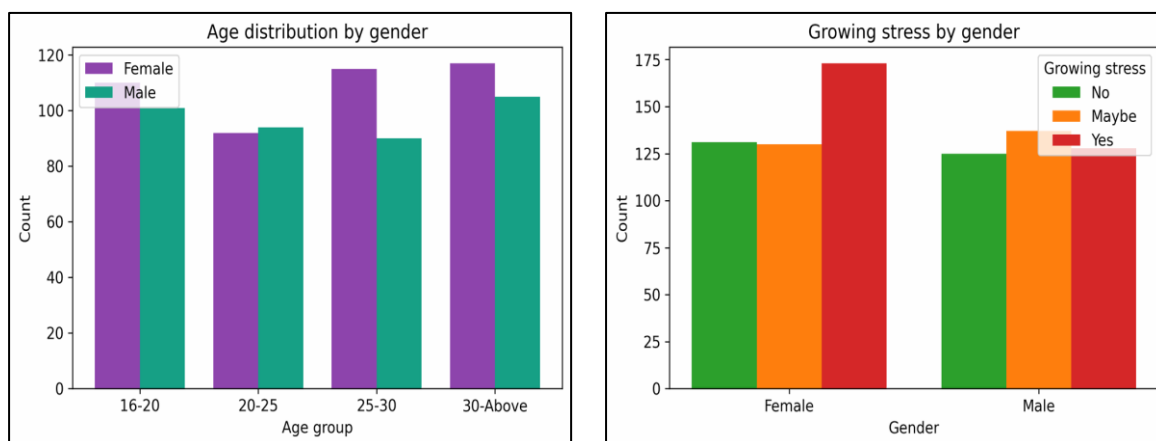
To place CBR and the machine learning baselines on directly comparable footing, Sec. 4.6 additionally reports a repeated stratified  $k$ -fold cross-validation (5 folds  $\times$  5 repeats, `random_state = 45`) in which every model, including CBR, is trained and evaluated on identical fold assignments: the CBR case base is rebuilt from the training fold only (no case ever appears in its own case base), and the one-hot encoding used by the machine learning models is refit on the training fold only in every repetition. Multiclass AUC is defined identically for every model as the macro-averaged one-versus-rest AUC computed from class scores rather than hard predictions: the `predict_proba` output for each machine learning classifier, and, for CBR, the proportion of the  $k$  retrieved neighbours voting for each class. A majority-class baseline (the most frequent class within each training fold) and a label-permutation null distribution are reported alongside these results.

## RESULTS AND DISCUSSION

### Exploratory Analysis

The exploratory analysis examined the demographic characteristics of the respondents and the relationship between input features and the target variable. The sample contained a higher proportion of older and female participants (Figure 3a). Female respondents showed a slightly higher tendency to report growing stress compared with male respondents (Figure 3b). However, the association analysis using Cramér's  $V$  (Figure 4) indicated weak relationships between the input features and Growing\_Stress, with all variables showing association values below approximately 0.1. These findings suggest that the available features provide limited predictive information for distinguishing stress levels.

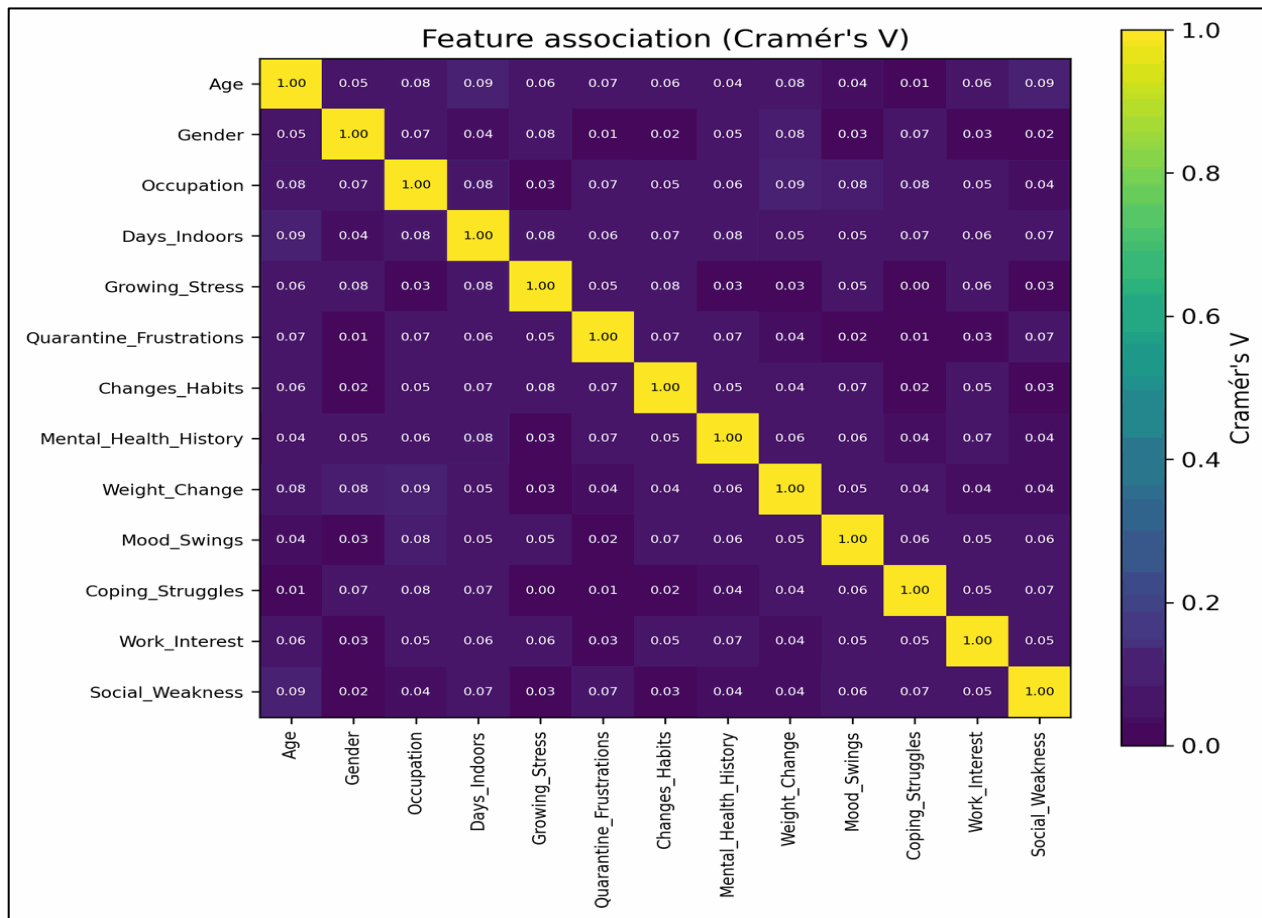
**Figure 3: Exploratory analysis of the dataset.**



Age distribution by gender.

Growing-stress classes by gender.

**Figure 4: Feature association matrix (Cramér's V).**



## Performance Evaluation of Machine Learning Models

Table 2 presents the cross-validated performance of the evaluated machine learning models. All classifiers achieved performance close to the random baseline of 0.33 accuracy, with macro-AUC values approximately equal to 0.50. These results were consistent across both 5-fold and 20-fold cross-validation, confirming the weak feature-target relationships observed during the exploratory analysis (Figure 4). The ROC curves (Figure 5) remained close to the diagonal line for all classes and models, indicating limited discriminative ability. Therefore, the evaluated machine learning approaches did not demonstrate meaningful predictive capability for Growing\_Stress using the available features.

**Table 2: Leak-free ML baselines (stratified cross-validation). Accuracy is reported as mean; precision, recall,  $F_1$ , and AUC are macro-averaged.**

| Model               | Acc. (5-fold) | Acc. (20-fold) | Precision | Recall | $F_1$ | AUC   |
|---------------------|---------------|----------------|-----------|--------|-------|-------|
| Random Forest       | 0.349         | 0.356          | 0.341     | 0.342  | 0.338 | 0.516 |
| k-NN                | 0.354         | 0.344          | 0.350     | 0.352  | 0.349 | 0.527 |
| XGBoost             | 0.359         | 0.352          | 0.355     | 0.356  | 0.354 | 0.509 |
| Decision Tree       | 0.325         | 0.335          | 0.320     | 0.323  | 0.320 | 0.492 |
| SVM                 | 0.362         | 0.352          | 0.351     | 0.352  | 0.343 | 0.522 |
| Logistic Regression | 0.329         | 0.341          | 0.307     | 0.319  | 0.308 | 0.499 |

**Figure 5: ROC curves of the evaluated machine learning models under the leak-free protocol (test size 0.2).**

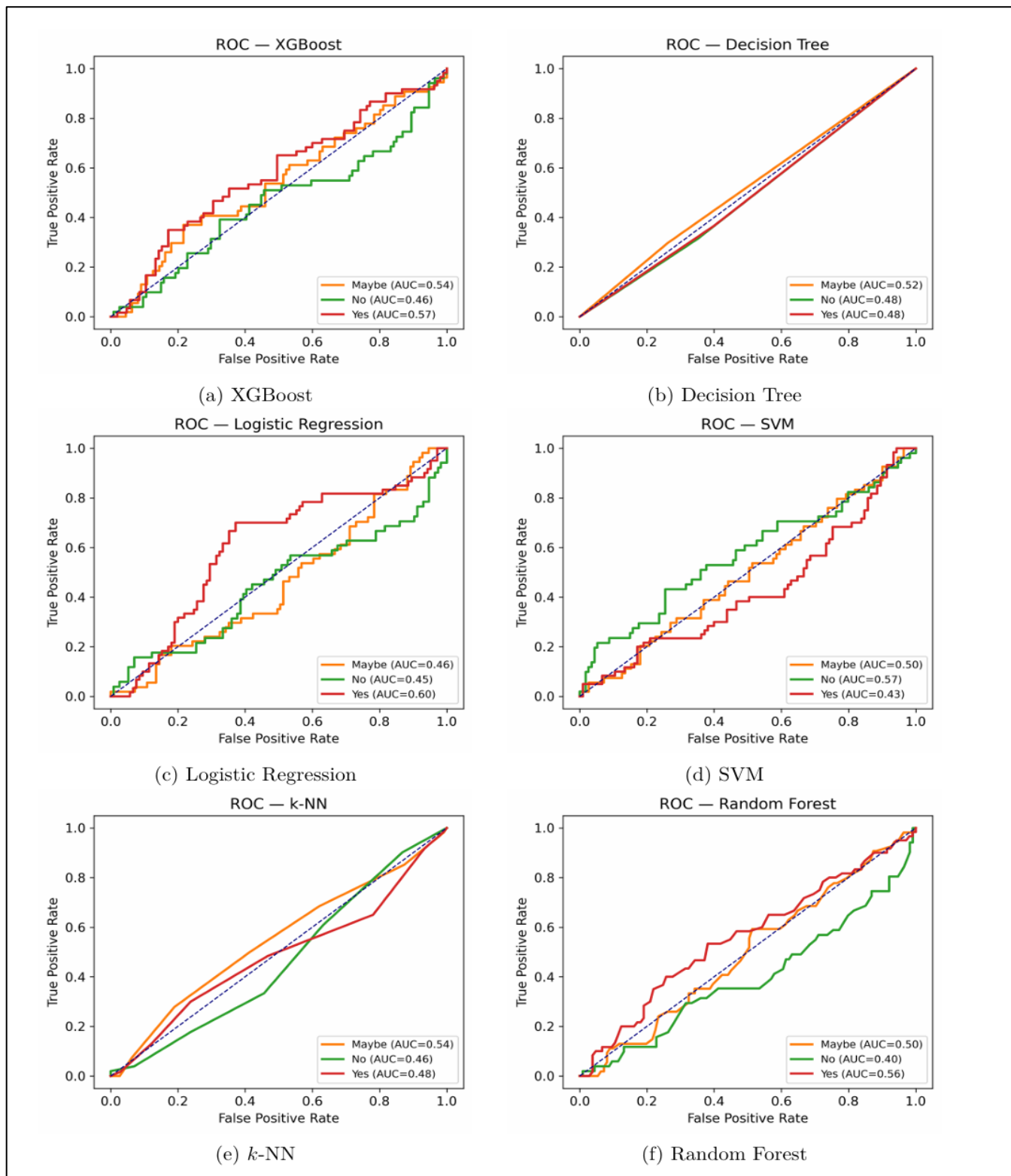


Figure 5 presents ROC curves for six machine learning models, XGBoost, Decision Tree, Logistic Regression, SVM,  $k$ -NN, and Random Forest, evaluated under a leak-free protocol with a 20% test size, each plotted separately for the “Yes,” “No,” and “Maybe” classes; across all six panels the curves stay close to the diagonal chance line, with AUC values ranging narrowly from about 0.40 to 0.60, the best individual results being Logistic Regression’s “Yes” class (AUC= 0.60) and XGBoost’s “Yes” class (AUC= 0.57), while the Decision Tree shows almost no separation from chance at all (AUC $\approx$  0.48–0.52); this consistent clustering near the diagonal across every model indicates limited discriminative ability once data leakage has been removed from the

evaluation. Section 4.6 reports these same six classifiers, together with CBR and a majority-class baseline, re-evaluated on identical cross-validation folds, with per-class metrics, confusion matrices, and a label-permutation null distribution.

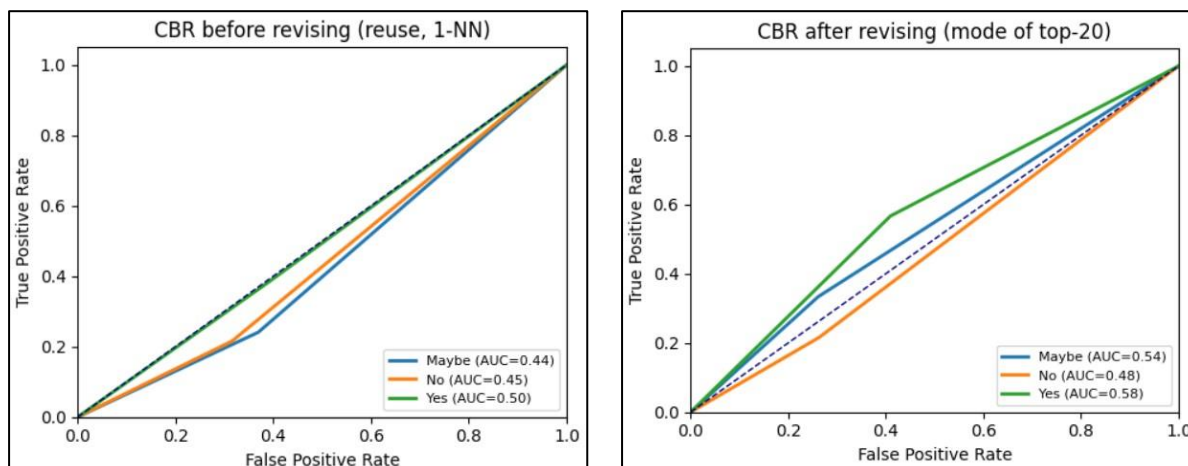
### Performance Evaluation of the CBR Approach

Table 3 presents the performance of the CBR approach under the correct leak-free protocol. The obtained accuracy values remained close to the chance level (0.30–0.43), with no clear improvement across different test sizes or neighbourhood sizes. The ROC curves before and after revision (Figure 6) also indicate limited class discrimination. The reuse strategy achieved an AUC of 0.46, while mode-based revision produced only a slight improvement with an AUC of 0.54. These results show that the previously reported high performance was caused by target leakage rather than genuine predictive ability.

**Table3: Leak-free CBR accuracy across test sizes (columns) and neighbourhood sizes  $k$  (rows). Values hover around the chance level of 0.33.**

| $k$ | 0.1  | 0.2  | 0.3  | 0.4  | 0.5  | 0.6  | 0.7  | 0.8  | 0.9  |
|-----|------|------|------|------|------|------|------|------|------|
| 5   | 0.31 | 0.32 | 0.34 | 0.34 | 0.35 | 0.36 | 0.34 | 0.36 | 0.37 |
| 10  | 0.31 | 0.35 | 0.35 | 0.35 | 0.35 | 0.37 | 0.37 | 0.38 | 0.35 |
| 15  | 0.37 | 0.35 | 0.38 | 0.38 | 0.36 | 0.36 | 0.36 | 0.35 | 0.36 |
| 20  | 0.31 | 0.38 | 0.38 | 0.33 | 0.37 | 0.38 | 0.33 | 0.34 | 0.38 |

**Figure 6: Leak-free CBR performance before and after revising the retrieval approach. The left panel shows the 1-NN reuse approach, while the right panel shows the mode of the top-20 retrieved cases. Both approaches remain close to chance performance.**



Section 4.6 reports CBR's performance under the same identical-fold cross-validation protocol used for the machine learning baselines, alongside the majority-class baseline and a label-permutation null distribution for direct comparison.

### Target Leakage in Similarity-Based Retrieval

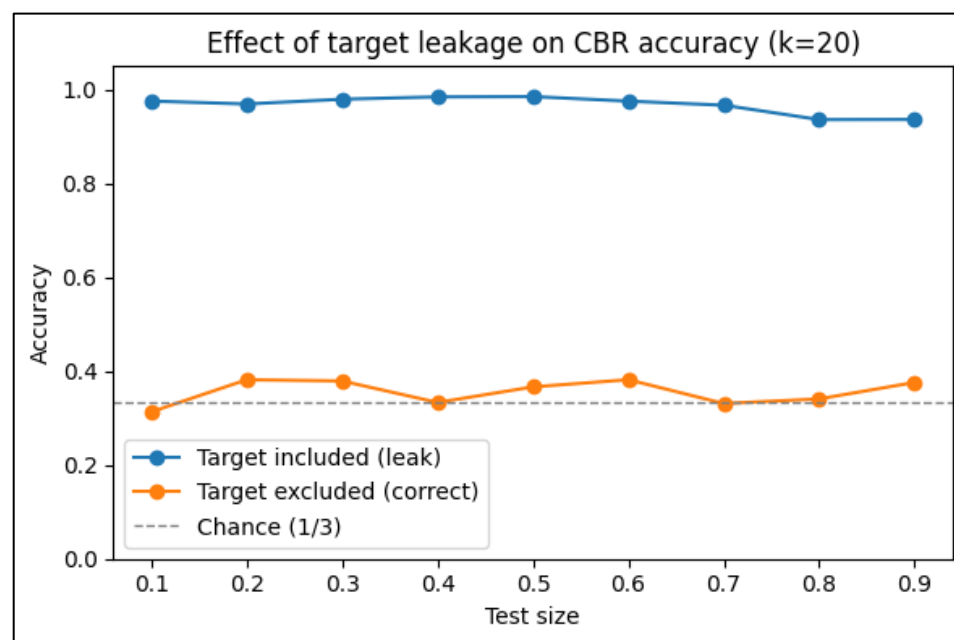
The CBR framework was further evaluated through a controlled ablation study to examine the effect of including the target attribute in the similarity computation. Using identical train/test splits, case bases, and  $k$ -values, we modified only the attribute set used in the similarity measure by adding the target variable, Growing\_Stress, to the twelve predictor attributes. Including the target attribute introduced a direct label-based similarity contribution between cases. Since the global similarity measure was computed as an unweighted average of attribute matches, cases with the same label became more likely to be retrieved regardless of agreement among the actual predictor attributes [15]. As a result, the revision step recovered labels from the leaked target

information rather than from genuine feature similarity. As shown in Table 4, this single modification increased accuracy from the leak-free range of 0.31 to 0.38 to 0.94 to 0.98, with AUC values reaching 1.00 (Figure 7). Because the data, model, and evaluation procedure remained unchanged, the observed performance improvement was entirely attributable to target leakage within the similarity computation. This finding highlights a structural vulnerability of the implemented CBR representation strategy, where the target attribute is treated as an ordinary case feature within an unweighted similarity measure. To prevent such leakage, target attributes should be excluded from similarity calculations and retained only for solution assignment and revision. Unlike conventional leakage caused by preprocessing or train–test contamination [4], this failure occurs within the retrieval mechanism itself and can lead to misleadingly high performance estimates.

**Table 4: The leakage artefact. Including the target in the similarity computation ( $k = 20$ ) inflates accuracy far above the leak-free result; both are shown for identical splits.**

| Test size                 | 0.1  | 0.2  | 0.3  | 0.4  | 0.5  | 0.6  | 0.7  | 0.8  | 0.9  |
|---------------------------|------|------|------|------|------|------|------|------|------|
| Target included (leak)    | 0.98 | 0.97 | 0.98 | 0.98 | 0.99 | 0.98 | 0.97 | 0.94 | 0.94 |
| Target excluded (correct) | 0.31 | 0.38 | 0.38 | 0.33 | 0.37 | 0.38 | 0.33 | 0.34 | 0.38 |

**Figure 7: Effect of target leakage on CBR accuracy.**



To audit this mechanism directly, we ran a label-permutation control matched to the ablation above: on each of 20 repeats, Growing Stress was globally permuted across all 824 records before the case base was built, so that the permuted value was used self-consistently as both the retrieval attribute and the ground truth for every case (train and test alike), rather than mismatching a shuffled attribute against the real answer. Table 5 compares CBR’s accuracy ( $k = 20$ , test\_size= 0.2) when the target attribute carries its true value against when it carries this semantically meaningless, permuted value.

**Table 5: Label-permutation control for the leakage ablation ( $k = 20$ , test\_size= 0.2, 20 repeats; values are mean accuracy  $\pm$  std).**

| Condition                       | Mean accuracy | Std   |
|---------------------------------|---------------|-------|
| Target included, true label     | 0.986         | 0.008 |
| Target included, permuted label | 0.992         | 0.007 |

The permuted-label condition is exactly as accurate as the true-label condition (if anything marginally higher, well within noise). This is the decisive audit result: it demonstrates that the 0.94–0.99 accuracy reported above is a mechanical artefact of retrieving cases whose stored attribute happens to equal the query’s own hidden value - it reproduces just as well for an arbitrary, meaningless relabelling, and therefore reflects no real relationship in the data. Table 6 illustrates the mechanism with a worked example: the top-5 retrieved neighbours for one representative query case (true label “Yes”) under the leak-free and leak-included attribute sets. Under leak-free retrieval the neighbourhood is a near-tied, uninformative mix of all three classes; once the target is included, the neighbourhood becomes dominated by cases sharing the query’s own class, which the revision step then simply reads back out. Two further worked examples are provided in the supplementary analysis package.

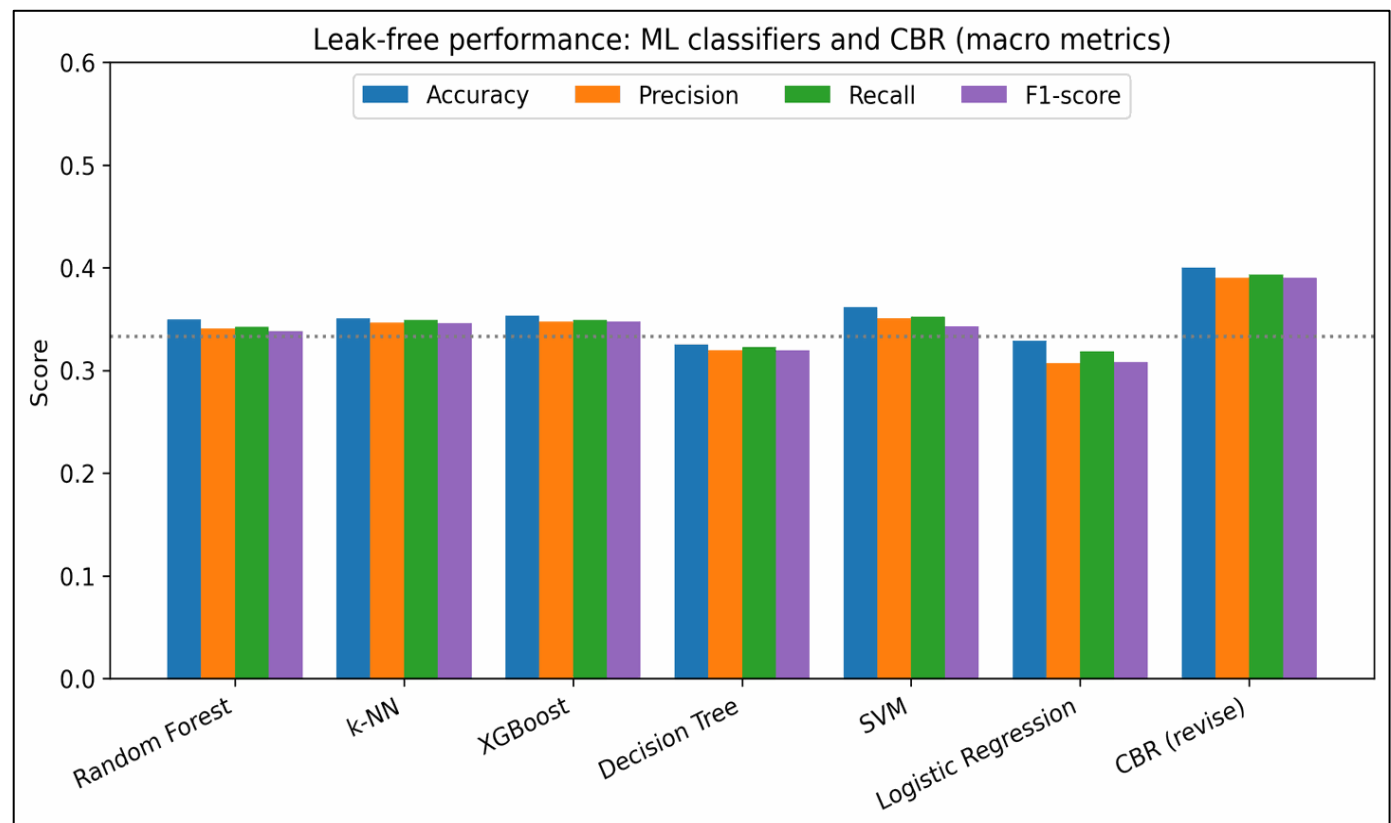
**Table 6: Representative neighbour trace ( $k = 5$ ) for one example query (true label “Yes”), showing the retrieved label and similarity score at each rank.**

| Condition     | Rank 1      | Rank 2        | Rank 3      | Rank 4      | Rank 5      |
|---------------|-------------|---------------|-------------|-------------|-------------|
| Leak-free     | Yes (0.667) | Maybe (0.667) | No (0.667)  | No (0.667)  | Yes (0.667) |
| Leak-included | Yes (0.692) | Yes (0.692)   | Yes (0.692) | Yes (0.615) | Yes (0.615) |

### Comparison of CBR and Machine Learning Approaches

Figure 8 compares the performance of the leak-free CBR approach and machine learning models. Both CBR and the six evaluated classifiers achieved almost similar performance, with all methods remaining close to the chance level across the evaluated metrics. No approach demonstrated reliable discriminative ability on the dataset. The variations in class-level recall reflect random classification behaviour rather than meaningful differences between classes.

**Figure 8: Leak-free comparison of the six ML classifiers and CBR (macro metrics). The dotted line marks the 0.33 chance level.**



## Cross-Validated Comparison, Baselines, and Confusion Matrices

Table 7 reports CBR and the six machine learning baselines evaluated on identical stratified cross-validation folds (5 folds  $\times$  5 repeats = 25 folds in total; random\_state= 45), with the CBR case base rebuilt and the one-hot encoding refit within each training fold, and with AUC computed identically for every model from class scores as defined in Sec. 3.4. The majority-class baseline, 36.53%, the share of the most frequent class (Yes) in the full dataset, computed within each training fold, is included as an additional, stricter benchmark alongside the theoretical 33.3% chance level used elsewhere in this manuscript.

**Table 7: Leak-free performance under unified 5-fold  $\times$  5-repeat stratified cross-validation (identical folds for CBR and all six machine learning models; values are mean  $\pm$  std across 25 folds).**

| Model                   | Accuracy          | Precision         | Recall            | F <sub>1</sub>    | AUC               |
|-------------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| Majority-class baseline | 0.365 $\pm$ 0.002 | 0.365 $\pm$ 0.002 | 0.365 $\pm$ 0.002 | 0.365 $\pm$ 0.002 | 0.500 $\pm$ 0.000 |
| CBR (k = 20)            | 0.359 $\pm$ 0.030 | 0.346 $\pm$ 0.033 | 0.351 $\pm$ 0.030 | 0.341 $\pm$ 0.032 | 0.511 $\pm$ 0.026 |
| SVM                     | 0.358 $\pm$ 0.033 | 0.345 $\pm$ 0.037 | 0.349 $\pm$ 0.034 | 0.339 $\pm$ 0.035 | 0.513 $\pm$ 0.036 |
| Random Forest           | 0.353 $\pm$ 0.026 | 0.345 $\pm$ 0.027 | 0.348 $\pm$ 0.026 | 0.343 $\pm$ 0.027 | 0.513 $\pm$ 0.026 |
| k-NN                    | 0.348 $\pm$ 0.029 | 0.344 $\pm$ 0.028 | 0.347 $\pm$ 0.029 | 0.344 $\pm$ 0.028 | 0.522 $\pm$ 0.031 |
| XGBoost                 | 0.346 $\pm$ 0.028 | 0.342 $\pm$ 0.031 | 0.343 $\pm$ 0.029 | 0.341 $\pm$ 0.030 | 0.505 $\pm$ 0.025 |
| Logistic Regression     | 0.341 $\pm$ 0.036 | 0.325 $\pm$ 0.040 | 0.332 $\pm$ 0.037 | 0.324 $\pm$ 0.039 | 0.495 $\pm$ 0.026 |
| Decision Tree           | 0.330 $\pm$ 0.031 | 0.328 $\pm$ 0.031 | 0.328 $\pm$ 0.031 | 0.327 $\pm$ 0.031 | 0.496 $\pm$ 0.023 |

Under this protocol, most models, Decision Tree, Logistic Regression, XGBoost, *k*-Nearest Neighbours, and Random Forest perform at or below the majority-class baseline, and CBR and SVM are statistically indistinguishable from it. To formally assess whether the observed leak-free accuracies could plausibly arise from features carrying no real relationship to Growing\_Stress, we additionally constructed a label-permutation null distribution: Growing\_Stress was randomly permuted across all 824 records (200 permutations), and CBR and Random Forest were re-evaluated on each permuted dataset using an identical 80/20 stratified hold-out. CBR's observed leak-free accuracy (0.359) fell at the 77th percentile of its own null distribution (mean 0.332, 95% range 0.260–0.406); Random Forest's observed accuracy (0.353) fell at the 67th percentile of its null distribution (mean 0.335, 95% range 0.267–0.412). Both observed values sit comfortably within the range expected under pure chance, giving formal statistical support to the conclusion that the available behavioral features carry no detectable predictive signal for Growing Stress in this dataset.

**Table 8: Pooled confusion matrix for CBR across all 25 cross-validation folds (rows: true class; columns: predicted class).**

| True \ Predicted | Maybe | No  | Yes |
|------------------|-------|-----|-----|
| Maybe            | 462   | 272 | 601 |
| No               | 407   | 247 | 626 |
| Yes              | 410   | 324 | 771 |

Consistent with the near-chance accuracy, CBR shows a systematic bias toward over-predicting the majority class Yes, a pattern shared by every machine learning model evaluated; per-model confusion matrices for all seven models, together with the fold-level results underlying Table 7, are provided in the supplementary analysis package accompanying this submission.

## Recommendations and Preventive Measures

Because none of the models evaluated in this study achieved performance above chance on the available behavioral features (Sec. 4.2–4.6), this manuscript does not propose a deployment framework or clinical recommendations based on its predictions. A validated predictive model could, in principle, support tiered support recommendations [16], but constructing and validating such a model is outside the scope of this demonstration and is left to future work using clinically validated outcome measures (Sec. 4.8).

## Limitations

This study has several limitations. First, the outcome variable represents self-reported growing stress rather than a clinical diagnosis of depression; therefore, the findings should not be interpreted as depression detection. Moreover, it reflects perceived stress progression rather than a validated psychological assessment instrument, such as the Perceived Stress Scale (PSS) or Patient Health Questionnaire (PHQ-9). Second, the dataset is based on self-reported information collected during the COVID-19 quarantine period, which may introduce reporting bias and limit the generalizability of the findings to other populations and contexts. Third, the analysis relied on a single dataset with 824 observations and lacked external validation using independent cohorts. Additionally, the available behavioral features provided limited predictive information, resulting in model performance close to random classification. Finally, any AI-based mental health screening system requires careful ethical consideration, clinical validation, and appropriate professional oversight before potential real-world application.

## CONCLUSION AND FUTURE DIRECTIONS

This study evaluated a Case-Based Reasoning (CBR) approach alongside conventional Machine Learning (ML) classifiers for stress risk screening using a public COVID-19 quarantine survey dataset. Under a leakage-free evaluation protocol, where predictor attributes were strictly separated from the target variable, neither the CBR approach nor the six evaluated ML models achieved performance beyond the chance level. A controlled ablation study demonstrated that including the target attribute in the CBR similarity computation artificially increased performance from chance level to near-perfect accuracy, revealing a structural target leakage problem arising from the implemented case representation strategy. The findings provide two main contributions: first, an empirical assessment showing that the available self-reported behavioral features did not provide sufficient predictive information for identifying growing stress in this dataset; and second, a reproducible demonstration of how inappropriate case representation can generate misleading performance estimates in CBR-based prediction systems. These results emphasize that high predictive performance in similarity-based models should be interpreted carefully and validated through strict predictor-target separation. Future research should investigate clinically validated stress assessment datasets, incorporate richer behavioral and physiological measurements, and evaluate models using independent external validation cohorts. Further development of CBR-based mental health screening approaches should consider adaptive similarity learning, transparent case representation strategies, and careful clinical evaluation before such systems can support reliable decision-making in low-resource healthcare settings.

**Data and Code Availability:** the similarity/retrieval/revision code, the exact cross-validation fold indices used for Table 7, the label-permutation experiments, and the per-model confusion matrices underlying Sec. 4.6 are provided as a supplementary analysis package accompanying this submission.

## REFERENCES

1. G. B. D. Mental and D. Collaborators, "Global , regional , and national burden of 12 mental disorders in 204 countries and territories , 1990 – 2019 : a systematic analysis for the Global Burden of Disease Study 2019," *The Lancet Psychiatry*, vol. 9, no. 2, pp. 137–150, 2022, doi: 10.1016/S2215-0366(21)00395-3.
2. G. M. Slavich and M. R. Irwin, "From Stress to Inflammation and Major Depressive Disorder : A Social Signal Transduction Theory of Depression," vol. 140, no. 3, 2014, doi: 10.1037/a0035302.
3. S. Graham, C. Depp, E. E. Lee, C. Nebeker, X. Tu, H.-C. Kim, and D. V. Jeste, "Artificial Intelligence for Mental Health and Mental Illnesses: An Overview," *Current Psychiatry Reports*, vol. 21, no. 11, p. 116, 2019, doi: 10.1007/s11920-019-1094-0.
4. S. Kaufman, S. Rosset, C. Perlich, and O. R. I. Stitelman, "Leakage in Data Mining : Formulation , Detection , and Avoidance," vol. 6, no. 4, pp. 1–21, 2012, doi: 10.1145/2382577.2382579.
5. A. Aamodt, "Case-Based Reasoning : Foundational Issues , Methodological Variations , and System Approaches," vol. 7, no. 1, pp. 39–59, 1994.
6. H. Burkhard, "2 . Extending some Concepts of CBR – Foundations of Case Retrieval Nets," no. Chapter 3.
7. M. M. Aldarwish, "Posts," pp. 282–285, 2017, doi: 10.1109/ISADS.2017.41.
8. A. B. R. Shatte, D. M. Hutchinson, and S. J. Teague, "Machine learning in mental health : a scoping review of methods and applications," 2019.
9. "Machine Learning Algorithms for Depression : Diagnosis ," pp. 1–20, 2022.
10. S. Wshah, C. Skalka, and M. Price, "Predicting Posttraumatic Stress Disorder Risk : A Machine Learning Approach Corresponding Author :," vol. 6, 2019, doi: 10.2196/13946.
11. A. M. Y. Tai *et al.*, "Arti fi cial Intelligence In Medicine Machine learning and big data : Implications for disease modeling and therapeutic discovery in psychiatry," *Artif. Intell. Med.*, vol. 99, no. August, p. 101704, 2019, doi: 10.1016/j.artmed.2019.101704.
12. S. Kapoor and A. Narayanan, "Article Leakage and the reproducibility crisis in machine- learning-based science Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, p. 100804, 2023, doi: 10.1016/j.patter.2023.100804.
13. J. L. Kolodner, "Educational Implications of Analogy".
14. Y. Kang, S. Krishnaswamy, and A. Zaslavsky, "A Retrieval Strategy for Case-Based Reasoning Using Similarity and Association Knowledge," vol. 44, no. 4, pp. 473–487, 2014.
15. G. Albora, M. Straccamore, and A. Zaccaria, "Machine learning-inspired similarity measure to forecast M & A from patent data," pp. 1–19, 2026, doi: 10.1371/journal.pone.0341010.
16. A. C. Butler, J. E. Chapman, E. M. Forman, and A. T. Beck, "The empirical status of cognitive-behavioral therapy : A review of meta-analyses," vol. 26, pp. 17–31, 2006, doi: 10.1016/j.cpr.2005.07.003.
17. N. Amin, I. Salehin, M. A. Baten, and R. Al Noman, "RHMCD-20 dataset: Identify rapid human mental health depression during quarantine life using machine learning," *Data in Brief*, vol. 54, p. 110376, 2024, doi: 10.1016/j.dib.2024.110376.
18. R. R. Sokal and C. D. Michener, "A statistical method for evaluating systematic relationships," *University of Kansas Science Bulletin*, vol. 38, pp. 1409–1438, 1958.