Deployment-Stage Validation and Web-Based Screening of Hybrid Ensemble Models for Workplace Depression and Anxiety Prediction

Akib Tanzim, Ramisha Binte Helal, Nyme Ahmed*
Department of Computer Science, American International University-Bangladesh, Dhaka-1229, Bangladesh

***Abstract*—This study assesses the deployment stage model refitting and consistency of implementation for hybrid stacking classifiers to classify the PHQ-9 and GAD-7 severity for 406 office workers from Bangladesh using questionnaires. The original audit compared evaluation models built with 284 training records with archived deployment models built with 345 development records, on test sets containing 61 records per target. The observed accuracy for the archived models was higher for PHQ-9 at 91.80 percent versus 88.52 percent and GAD-7 at 93.44 percent versus 91.80 percent, but neither exact McNemar comparison was statistically significant, with p-values of 0.500 and 1.000, respectively. Paired bootstrap intervals contained zero. A direct questionnaire-scoring baseline recovered all 406 labels for each target, and the original 35 predictors included the 16 questionnaire items that were used to define the targets. A repeated evaluation of 960 model fits was conducted across 20 stratified seeds, six model families, two feature sets, and evaluation or refit versions. Refit HybridStack accuracy was 89.10 percent and 94.26 percent with 35 features, but just 39.34 percent and 49.92 percent with 19 context-only features for PHQ-9 and GAD-7, respectively. Probability quality was evaluated without the use of a calibration model. The new application contains 298 baseline, 26 extra CSV, and 20 documentation or chart parity checks, and maintains feature, prediction, and full probability-vector parity on the available records. These findings do not support clinical prediction or generalizability to new workplaces, but instead support deployment auditing and implementation consistency.**

Keywords*—Workplace mental health; depression-anxiety severity; PHQ-9; GAD-7; model retraining; hybrid ensemble learning

I. Introduction

Psychosocial working conditions, including high job demands, limited workplace control, and inadequate organizational support, have been associated with depressive and anxiety symptoms among employees [1]–[7]. When clinician-administered assessments are impractical for large-scale workplace research, standardized self-report instruments such as the Patient Health Questionnaire-9 (PHQ-9) [8] and Generalized Anxiety Disorder-7 (GAD-7) [9] can provide structured assessments of symptom severity. Machine-learning approaches have also been applied to questionnaire-based mental-health data to classify self-reported symptom-severity categories. However, the evaluation of model performance after retraining and the verification of implementation consistency remain important methodological considerations.

This literature tends to neglect the methodological steps that occur after a model has been selected. By this point, the standard practice in many cases is to fit the same architecture on the entire training and validation data. Refitting a selected model architecture using additional development data can change its performance; however, the direction and magnitude of any change requires empirical evaluation. Because of the interaction between the steps of model selection and performance estimation, several evaluation frameworks warn against blurring the line between model selection and true assessment [10], [11]. In the context of modern guidance for prediction models, the distinction between the development and evaluation phases is also still valid [12], [13]. Refitting after model selection is an established practice, but performance estimates for the selected model should not automatically be attributed to the newly refitted artifact. In practice, the performance figures reported in a research paper are usually for the validation-selected model and not for the refit model that is ultimately deployed. This distinction is not cosmetic.

A recent scoping review found that workplace mental-health machine-learning studies predominantly used cross-sectional designs and emphasized the limited investigation of temporal dynamics and future mental-health outcomes [14]. More generally, human-computer interaction research on mental health and machine learning emphasizes that predictive accuracy is an inadequate design foundation for models without describing how those models will behave once they leave the training environment [15], and the engineering research on production machine learning has documented the gap in model behaviour between reported metrics and the behaviour of the deployed system as technical debt [16].

This study builds on two previously developed hybrid stacking ensembles, HybridStack_PHQ and HybridStack_GAD, for classifying questionnaire-defined PHQ-9 depression severity and GAD-7 anxiety severity among 406 Bangladeshi office workers. The investigation addresses four related questions. First, in each target, how different is the model reconstructed from the evaluation-stage data from the archived model at the deployment stage, and how uncertain are these differences? Second, how well can the provided severity labels be reproduced using a deterministic scoring of the questionnaires, and what is the impact of adding target-defining questionnaire items on the interpretation of the original classification task? Third, are the differences between evaluation and refitting stable across repeated data partitions, model families, and the context-only feature set, which omits all

*Corresponding author.

PHQ-9 and GAD-7 questionnaire items? Fourth, what can be learned about the deployed artifacts and their implementation from the probability-quality assessments and software verification? These questions are related to classification, evaluation at deployment, and consistency of implementation as defined by the questionnaires, but not clinically verified diagnoses or to prospective prediction of mental health.

This study has four major contributions. First, a paired evaluation compares reconstructed evaluation-stage models with archived deployment-stage models on target-specific held-out partitions, incorporating accuracy, macro-F1, exact McNemar testing, and bootstrap uncertainty analysis. Second, a feature–target audit and deterministic questionnaire-scoring baseline examine the relationship between the original predictors and questionnaire-defined severity labels. Third, a second repeated evaluation is done for six model families, 20 stratified partitions, two sets of predictors (the original 35 encoded features and 19 context-only features), and evaluation-stage versus refitted training procedures. This analysis examines the stability of observed differences and the extent to which performance depends on the availability of questionnaire responses. Fourth, probability-quality assessments and additional software verification examine model-generated probabilities, preprocessing consistency, input validation, and documented application behavior. These contributions provide a more comprehensive deployment audit without establishing clinical superiority, independent diagnostic validity, or external generalizability.

The web-based research prototype provides individual and batch questionnaire-screening interfaces and a model-documentation view. The revised implementation is publicly available, together with a live demonstration of the application workflow. Access details are provided in Data Availability.

The remainder of this study is organized as follows. Section II reviews related research on workplace mental-health classification, ensemble learning, and deployment-stage evaluation. Section III describes the dataset, questionnaire-scoring baseline, model architectures, original and repeated evaluation protocols, probability assessment, and implementation verification. Section IV reports the original paired model comparisons, class-wise results, direct-scoring findings, repeated and context-only model evaluations, probability assessments, and software verification results. Section V discusses the implications, Section VI addresses limitations, and Section VII presents the conclusions and future research directions.

II. Related Work

A. Machine Learning for Workplace Mental Health

Occupational health research concludes that anxiety and depressive symptoms can be attributed to psychosocial work conditions and organizational factors, although research is limited to the extent to which causal relationships can be drawn across different labour markets [1]–[6]. Within this research context, machine-learning methods have been applied to structured workplace survey data to classify questionnaire-derived mental-health severity categories. As previously noted, the machine learning framework presented in this study builds on studies that have developed and tested a classifier of the PHQ-9 and GAD-7 among Bangladeshi office employees using hybrid stacking ensembles. It has been shown that even with the limited existing studies employing machine learning for mental health in the workplace, most of the research has largely been conducted from a computer science perspective and has predominantly employed cross-sectional, single-measure studies [14]. It is for this reason that the current study focuses on what happens after a model has been developed, and not on the development of a model.

B. Ensemble Learning for Structured Health Classification

Stacked generalization is a technique where a meta-learner integrates base learner predictions via a function beyond simple voting. Wolpert formalized this method to take advantage of complementary error patterns across different models [17]. With respect to structured tabular data, XGBoost and CatBoost are competitive baselines [18], [19], and scikit-learn provides stacking frameworks and cross-validation tools to combine these models [20]. In the original model-development study, the PHQ-9 stacking ensemble combined CatBoost, XGBoost, and Logistic Regression, while the GAD-7 ensemble combined Logistic Regression, CatBoost, and Random Forest. These architectures were selected using validation macro-F1 within the original candidate-model comparison. The additional six-model experiments presented in this revision assess their performance under repeated partitions and alternative predictor availability rather than assuming that the originally selected ensembles are universally superior.

C. The Research-to-Deployment Gap in Applied Machine Learning

Sculley et al. stated that the cost and maintenance of a machine learning system during deployment are considerable, and that assumptions that hold during the development of a system erode once the system is being used to fulfil a need [16]. Statistical evaluation research states that the credibility of performance claims can be materially impacted by the procedure of classifier comparison and model selection [10], [11]. Furthermore, when reporting risk of bias is emphasized, frameworks for transparency stress that model development and evaluation must be clearly separated [12], [13]. The specific case of this gap addressed here is rather circumscribed and definite: Although refitting a validation-selected architecture using combined training and validation data is an established machine-learning practice, the distinction between evaluation-stage performance and the performance of the artifact deployed remains methodologically important. The present study addresses this distinction through a paired comparison of reconstructed evaluation-stage and archived deployment-stage models on target-specific held-out partitions. The additional scoring and repeated-evaluation analyses further clarify the interpretation and stability of the observed results.

D. Human-Computer Interaction Perspectives on Deployable Mental-Health ML

Thieme et al. conducted a systematic review of human-computer interaction studies on machine-learning mental-health systems and expressed that predictive performance metrics in isolation do not justify the deployability of a system. Indeed, in the real world, a system's deployability hinges on whether and

how model outputs are made available, trusted, and integrated into a person's or organization's given workflow [15]. Structural power asymmetries and uncertainty about how the data will be used further complicate worker consent [6]. AI risk-management guidance emphasizes evaluating systems within their intended contexts [21], while WHO guidelines provide recommendations for protecting and supporting mental health at work [22]. In this regard, the application detailed in Section III-D is offered as a research prototype for the analysis of preprocessing and prediction consistency of the records provided, not as evidence of clinical validity or operational suitability in actual workplaces.

E. Web-Based Deployment Precedents in Health Machine Learning

Utilizing lightweight web frameworks allows a trained health screening model to become an inspectable and reproducible artifact. As a precedent, Buga et al. developed a Streamlit application that hosted a brain-metastasis deep learning model for use in monitoring patients post-Gamma Knife surgery [23]. Buga et al. utilized Streamlit to present a means for clinical review of the model's output without the need for a deployment stack. In a workplace context, Assumang et al. proposed a framework for employee mental-health dashboards through a desk-based review [24]. The framework discusses combining employee-related information into summaries for organizational decision-making. It provides a conceptual precedent for aggregate dashboard presentation, rather than an empirical benchmark for the present application. Together, the Streamlit implementation described by Buga et al. [23] and the dashboard framework proposed by Assumang et al. [24] informed the individual and batch presentation modes of the research prototype.

III. Materials and Methods

Fig. 1 summarizes the original deployment-stage model comparison and implementation-verification workflow. Further analysis explored deterministic scoring in questionnaires, re-performing six-model comparisons with the original and context-only feature sets, probability quality, and revised application input handling, as described below.

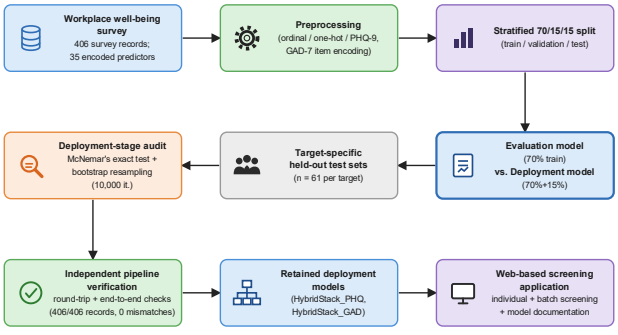

Fig. 1. A model evaluation and implementation verification workflow for the deployment stage.

A. Dataset and Original Modeling Framework

The dataset comprises 406 cross-sectional survey records from office workers in the banking and information technology sectors of Bangladesh. The raw dataset has 29 columns, one of which is a timestamp column, or 28 columns if the timestamp column is not counted. The processed data comprises 35 encoded predictor features and two questionnaire-defined severity targets. Demographic and occupational variables were transformed using the original documented ordinal mappings and one-hot encoding procedures. The original feature set includes all nine PHQ-9 questionnaire items and all seven GAD-7 questionnaire items, each encoded as an integer response ranging from 0 to 3. Accordingly, the predictor matrix contains the same questionnaire responses from which the severity targets are calculated. The dataset does not provide sufficient participant or workplace identifiers to establish clustering by individual or organization.

The PHQ-9 severity target was derived from the total PHQ-9 questionnaire score and categorized into five levels: None-Minimal, Mild, Moderate, Moderately Severe, and Severe. Similarly, the GAD-7 severity target was derived from the total GAD-7 questionnaire score and categorized into four levels: Minimal, Mild, Moderate, and Severe. Importantly, the questionnaire items used to calculate these severity labels were also included among the model predictors. Consequently, the classification task evaluates the reproduction of questionnaire-defined severity categories from the corresponding questionnaire responses rather than the independent prediction of clinically diagnosed mental-health conditions. Although previous studies have examined measurement invariance of PHQ-9 and GAD-7 across demographic groups [25], [26], those findings do not establish equivalent measurement properties within the present Bangladeshi workplace sample. The preprocessing consistency checks undertaken in this study are described in Section III-E.

A deterministic questionnaire-scoring baseline was implemented using the published scoring instructions [27]. PHQ-9 totals were calculated by summing nine item responses, each scored from 0 to 3. Scores of 0–4, 5–9, 10–14, 15–19, and 20–27 were assigned to None-Minimal, Mild, Moderate, Moderately Severe, and Severe categories, respectively. GAD-7 totals summed seven responses scored from 0 to 3; ranges of 0–4, 5–9, 10–14, and 15–21 corresponded to Minimal, Mild, Moderate, and Severe categories. The implementation required complete integer responses within the permitted range and rejected missing or invalid inputs without imputation. Scoring requires no model fitting. Agreement was assessed against the supplied labels for all available records and each original partition. This baseline evaluates recovery of questionnaire-defined categories and is not an equal-input comparator for context-only models, which exclude the questionnaire items.

To examine classification performance without the questionnaire items used to construct the target labels, a separate context-only feature set was defined. All nine PHQ-9 and seven GAD-7 questionnaire items were excluded from the original 35-feature matrix, leaving 19 encoded demographic, occupational, and workplace-context features. These included age group, professional experience, salary range, commuting duration, job-security and growth satisfaction, weekday sleep duration, and

encoded gender, marital-status, stress-relief, and organizational-support categories. The original questionnaire-defined targets were retained unchanged. Therefore, the context-only experiment estimates associations between contemporaneous workplace-context variables and questionnaire-derived severity categories; it does not evaluate future mental-health outcomes or independently confirmed diagnoses. The 19-feature models were evaluated separately and did not replace the archived deployment models.

B. Model Architectures

Two stacking ensembles from the original study are included with unmodified architectures. HybridStack_PHQ stacks the base learners CatBoost, XGBoost, and Logistic Regression, coupled with a Logistic Regression meta-learner trained with five-fold internal cross-validation on out-of-fold base-learner predictions (stack_method="predict_proba", cv=5, passthrough=False). HybridStack_GAD implements the same configuration with Logistic Regression, CatBoost, and Random Forest as base learners. Both were originally selected as the best configuration, in terms of validation macro-F1, among seven single-model baselines and two ensemble strategies (soft voting and stacking) tested for each target.

The additional experimental comparison included six model families: a majority-class baseline, Logistic Regression, Random Forest, CatBoost, XGBoost, and the previously selected HybridStack architecture. All six models were evaluated for both severity targets using both the original 35-feature matrix and the separate 19-feature context-only matrix. The recorded model configurations were retained without conducting a new hyperparameter search. Where feature scaling was applicable, its parameters were learned within the corresponding training pipelines, including within the internal folds used by the stacking implementation. The stacking ensembles retained five-fold out-of-fold training, probability-based stacking, and no direct feature passthrough. The additional model fits were used for comparative sensitivity analysis and did not replace the archived deployment artifacts.

C. Deployment-Oriented Retraining Protocol

The original evaluation used a two-stage stratified data-partitioning procedure with random seed 42. First, 70% of records were assigned to training and 30% to a temporary partition. The temporary partition was then divided equally into validation and test subsets using the same random seed and target-specific stratification. This produced 284 training, 61 validation, and 61 test records for each target. PHQ-9 and GAD-7 were partitioned separately according to their respective target labels; consequently, their test partitions were not identical, with only eight records appearing in both target-specific test sets. Within each target, however, the evaluation-stage and deployment-stage variants were compared on the same 61 held-out records.

Table I reports severity-class counts in the original target-specific training, validation, and test partitions, together with the combined training-and-validation counts used for refitting.

TABLE I. ORIGINAL TARGET-SPECIFIC CLASS COUNTS BY PARTITION

Target/class	Training	Validation	Test	Refit
PHQ None-Minimal	120	26	26	146
PHQ Mild	97	20	21	117
PHQ Moderate	31	7	7	38
PHQ Moderately Severe	20	4	4	24
PHQ Severe	16	4	3	20
GAD Minimal	131	28	29	159
GAD Mild	97	21	20	118
GAD Moderate	30	6	7	36
GAD Severe	26	6	5	32

[a] Refit combines the training and validation partitions and is not an additional disjoint partition. Each target has 284 training, 61 validation, and 61 test records; refitting uses 345 records. PHQ-9 and GAD-7 test partitions share eight records.

The original evaluation-stage model artifacts were unavailable and were therefore reconstructed from archived model configurations using the corresponding 284-record training partitions. Their reproduced performance estimates agreed with the originally reported values after rounding. The deployment-stage models were retained as archived artifacts fitted using the combined training and validation partitions of 345 records. The original test records were not used to fit either model variant. The archived deployment artifacts remained the models served by the research prototype, while the additional models trained during the repeated experiments were separate experimental fits. Reconstruction of the evaluation-stage scores is not necessarily indicative of the recovery of all previous historical model-selection and preprocessing decisions.

Both models are tested on the same test rows in terms of accuracy and macro-averaged F1, the same as the macro-F1 used in the original study as the main indicator in the case of class imbalance. Beyond this point-estimate comparison, two sets of paired predictions on the same test partition are used to evaluate both models [28], [29]:

- McNemar's exact test. A binary correctness indicator is calculated for each model for each test row, and the number of disagreements between the models is counted. Since the number of discordant pairs is small in both targets (Section IV-B), the exact two-sided binomial form of McNemar's test is used instead of the chi-square approximation, which is not valid at this sample size [28], [30].
- Bootstrap resampling. A documented paired bootstrap procedure was applied separately to each target's original 61-record held-out test partition. For each target, 10,000 row-level bootstrap resamples were drawn with replacement using NumPy's default_rng(42), reset independently for each target. Both versions of the model were recalculated using the same resampled row indices to obtain accuracy and macro-F1. Differences were calculated as deployment minus evaluation. The primary macro-F1 calculation used the full list of classes for the target, with undefined class F1 values set to zero (zero_division=0). The 95% percentile intervals reported were the 2.5th and 97.5th percentiles of the bootstrap

differences. A second sensitivity calculation was performed using an average of macro-F1 across classes within resamples. The percentage of resamples in favor of the deployment model was kept as a descriptive resampling frequency, rather than as an effect size or the probability of true superiority. As the original bootstrap code was not available, these calculations are documented as a recovery rather than a recovery of the original resampling implementation [29], [30].

- Repeated evaluation protocol. A further sensitivity analysis was performed with 20 random seeds (1001–1020). The same two-stage stratified splitting procedure resulted in 284 training, 61 validation, and 61 test records for each seed and target. Six model families have been tested, using the 35-feature full feature set and the 19-feature context-only set. Each architecture was configured in an evaluation-stage configuration with 284 training records and in a newly refitted configuration with the combined 345-record training + validation partition. The corresponding variants were evaluated on identical target-specific test records. Across two targets, two feature sets, six models, two fitting versions, and 20 seeds, the protocol comprised 960 completed model fits. Model configurations were fixed before the repeated evaluation, and all predetermined seeds were retained irrespective of performance. Accuracy, macro-F1, class-wise metrics, and probability-quality metrics were recorded for each complete fit. Results were summarized using means, sample standard deviations, and matched refit-minus-evaluation differences. As repeated partitions reused the same 406 survey records, these summaries represent descriptive internal sensitivity analyses rather than independent validation cohorts. Furthermore, the historical model-selection process had already used this dataset, and a fully nested re-evaluation of that selection process was not performed.

D. System Implementation

The deployment-stage models are integrated into a web-based research screening prototype consisting of three main components: preprocessing, model inference, and the application interface. The preprocessing component independently implements the ordinal category mappings, nominal one-hot encoding, and PHQ-9/GAD-7 item-to-numeric conversions used during model development. It produces encoded feature vectors in the order expected by the trained models. The inference component loads the serialized deployment-stage models and uses them to generate questionnaire-defined severity classifications and model-generated class probabilities. The application interface provides individual screening, batch screening, and model-documentation views.

The individual-screening interface allows for input of demographic and occupational information and questionnaire responses and outputs the predicted severity category along with model class probabilities. These probabilities are shown as model outputs, not calibrated estimates of clinical certainty, as probability calibration has not been determined in this current study. The batch-screening interface screens multiple records, checks the inputs that are submitted, marks malformed records, and provides summarisation of the predicted severity distributions for the successful records. The model-documentation view presents the ensemble architectures, training-data characteristics, and deployment-versus-evaluation performance comparisons.

The application retained the original archived deployment model weights and feature schema. The context-only experimental models were not substituted for these artifacts. Following the initial audit, input and CSV parsing were revised to enforce the documented input contract, including validation of questionnaire item types, ranges, missing fields, and malformed batch records. Automated preprocessing, inference-parity, parser, and documentation/chart checks were performed locally. The revised implementation was subsequently released at https://github.com/akib1437/workplace-mental-health-screening-revised and deployed at https://workplace-mental-health-screening.streamlit.app/. The reported automated checks concern the locally evaluated implementation and do not constitute a separate validation of the hosted runtime.

E. Verification Methodology

Implementation verification was organized into preprocessing parity, inference parity, application input validation, documentation/chart checks, and manual interface observations. Preprocessing parity was examined using the 406 available survey records by reconstructing the corresponding questionnaire and contextual inputs and re-encoding them through the application preprocessing pipeline. The resulting feature vectors were compared element by element with the original encoded records. Inference parity compared archived-model predictions and complete model-generated class-probability vectors obtained from the original feature representations and reconstructed application inputs.

Application input validation was evaluated against a documented set of expected behaviors. The initial application was tested for valid and malformed individual and batch inputs, including invalid questionnaire values, missing fields, and CSV formatting conditions. The updated application was then tested against the original test suite and extra CSV parser tests. The revised model-documentation and visualization components were analysed by separate automated checks of the data and configuration contracts. Manual interface observations encompassed individual-screening boundary examples, valid and mixed validity batch submissions, downloaded valid-row results, and updated documentation views. These were considered smoke checks and not full browser or usability testing.

These procedures evaluate the consistency of implementation and handling of documented inputs on the available records and test cases. They do not validate clinically, do not demonstrate temporal stability, do not demonstrate robustness to data distributions not tested, and do not demonstrate reliability in new workplaces.

F. Probability Quality Assessment

Multiclass log loss and the multiclass Brier score, two well-known scoring rules for probabilistic predictions [31], were used to assess the model-generated class probabilities. The Brier

score was calculated as the average over the records assessed of the sum of squared differences between the predicted probability vector and the one-hot encoded true class vector. No adjustment was made to the number of classes. Probability-quality metrics were calculated for the original held-out evaluation and deployment predictions and for the repeated experiments.

Descriptive one-versus-rest reliability summaries were constructed using five fixed predicted-probability bins: [0, 0.2), [0.2, 0.4), [0.4, 0.6), [0.6, 0.8), and [0.8, 1.0]. The last bin contained the endpoint 1.0. Observations were made for each bin, and empty bins were not assigned an observed event frequency. Pooled reliability summaries were repeated appearances of predictions of existing records, not independent participants, for repeated evaluations. No calibration model for the probability was fitted. The analyses thus represent patterns of probability-quality and reliability of the evaluated samples and not calibrated clinical certainty.

IV. Results

A. Deployment Vs. Evaluation-Stage Model Performance

Table II presents the performance of the reconstructed evaluation-stage models, archived deployment-stage models, and deterministic questionnaire-scoring references on the original target-specific held-out test partitions. Each target contained 61 test records, with model variants compared on identical records within that target. For PHQ-9, the reconstructed evaluation-stage model achieved 88.52% accuracy and 0.8133 macro-F1, compared with 91.80% accuracy and 0.8827 macro-F1 for the archived deployment-stage model. This corresponds to 54 and 56 correct predictions, respectively. For GAD-7, the corresponding accuracies were 91.80% and 93.44%, with macro-F1 scores of 0.8852 and 0.8955, representing 56 and 57 correct predictions. Both deployment-stage models therefore achieved higher observed point estimates, although statistical testing did not establish a significant accuracy improvement. Direct questionnaire scoring achieved 100.00% accuracy and 1.0000 macro-F1 for both targets on their original held-out test partitions. This perfect agreement reflects deterministic recovery of questionnaire-defined severity labels from their defining items, not independent clinical predictive validity.

TABLE II. Original Held-Out Performance of Reconstructed Evaluation Models, Archived Deployment Models, and Direct Questionnaire Scoring

Target	Model version	Accuracy	Macro-F1
PHQ-9	Reconstructed evaluation	88.52%	0.8133
PHQ-9	Archived deployment	91.80%	0.8827
PHQ-9	Direct questionnaire scoring	100.00%	1.0000
GAD-7	Reconstructed evaluation	91.80%	0.8852
GAD-7	Archived deployment	93.44%	0.8955
GAD-7	Direct questionnaire scoring	100.00%	1.0000

[b.] Each target uses its own 61-record test partition. Direct scoring uses the questionnaire items and requires no model training.

B. Paired Significance Testing

Table III summarizes the paired comparison of the evaluation-stage and deployment-stage models on the shared held-out test records. For PHQ-9, the models differed in correctness on only two of the 61 records, with both disagreements favouring the deployment-stage model. The exact McNemar test yielded $p = 0.500$. For GAD-7, three discordant records were observed, with two favouring the deployment-stage model and one favouring the evaluation-stage model. The corresponding exact McNemar test yielded $p = 1.000$. Neither comparison was statistically significant at the conventional 0.05 significance level.

Bootstrap resampling was used to characterize uncertainty in the original paired model comparisons. For PHQ-9, the deployment-minus-evaluation accuracy difference was 0.0328, corresponding to approximately 3.28 percentage points. Its 95% paired bootstrap interval was [0.000, 0.082]. Under the prespecified full-class averaging policy, the corresponding macro-F1 difference was 0.0695, and the 95% interval was [0.000, 0.172]. The deployment model had a higher accuracy and macro-F1 in 87.01% of bootstrap resamples, but these frequencies do not establish true model superiority. For GAD-7, the observed accuracy difference was 0.0164, or approximately 1.64 percentage points, with a 95% interval of [-0.033, 0.082]. The macro-F1 difference was 0.0103, with an interval of [-0.024, 0.049]. Deployment-stage scores were higher in 60.69% of accuracy resamples and 61.46% of macro-F1 resamples. Both GAD-7 intervals also included zero. An alternative observed-class macro-F1 averaging policy produced a PHQ-9 upper interval endpoint of 0.195, rather than 0.172, illustrating sensitivity to the treatment of classes absent from individual bootstrap samples.

TABLE III. Paired Comparison of Archived Deployment and Reconstructed Evaluation Models

Measure	PHQ-9	GAD-7
Discordant correctness pairs	2	3
Deployment fixes	2	2
Deployment regressions	0	1
Exact McNemar p-value	0.500	1.000
Bootstrap 95% CI, Δ accuracy	[0.000, 0.082]	[-0.033, 0.082]
Bootstrap 95% CI, Δ macro-F1	[0.000, 0.172]	[-0.024, 0.049]
Resamples favouring deployment (Accuracy%)	87.01%	60.69%
Resamples favouring deployment (Macro-F1%)	87.01%	61.46%

C. Confusion Matrix and Classwise Performance Analysis

Fig. 2 and 3 compare the reconstructed evaluation-stage and archived deployment-stage confusion matrices for PHQ-9 and GAD-7, respectively. Within each target, both models were evaluated on the same 61 test records; the test partitions differed between targets. For PHQ-9, the deployment-stage model correctly classified 56 records. Among the five misclassifications, one Mild case was classified as None-Minimal, three Mild cases were classified as Moderate, and one Moderately Severe case was classified as Severe. For GAD-7, the deployment-stage model correctly classified 57 records. Its

four misclassifications included one Minimal case classified as Mild, one Mild case classified as Minimal, and two Moderate cases classified as Severe. All displayed errors occurred between adjacent severity categories. These observations describe the error patterns within the available held-out test partition and do not establish performance in independent workplace populations.

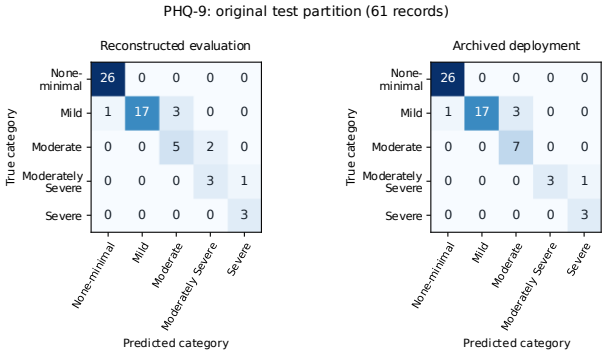

Fig. 2. PHQ-9 confusion matrices for the reconstructed evaluation-stage model and archived deployment-stage model on the same target-specific 61-record test partition. Cell values represent record counts.

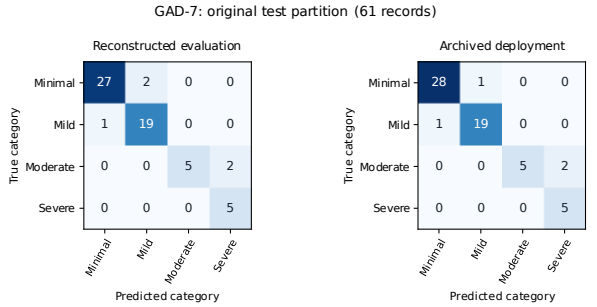

Fig. 3. GAD-7 evaluation-stage and deployment-stage confusion matrices on the same 61-record held-out test partition. Cell values represent record counts.

Table IV presents classwise precision, recall, specificity, F1 scores, and test support for the reconstructed evaluation-stage and archived deployment-stage models. The class-wise comparison showed that the PHQ-9 deployment-stage model corrected two Moderate cases that the evaluation-stage model had classified as Moderately Severe, without introducing additional incorrect predictions. As a result, the Moderate class recall improved from 5/7 (71.43%) to 7/7 (100.00%) on the original held-out test partition. For GAD-7, deployment refitting corrected two predictions and added one new error, for a net gain of one correct prediction. These changes were just on a few records and not a general benefit for rare severity categories. The original test partitions contained only three PHQ-9 Severe cases and five GAD-7 Severe cases. Although the deployment models correctly classified the available Severe cases, the sparse support and wide recall uncertainty prevent reliable conclusions about severe-case detection in a broader population.

TABLE IV. CLASSWISE PERFORMANCE OF EVALUATION AND DEPLOYMENT MODELS FOR PHQ-9 AND GAD-7 SEVERITY

PHQ-9 classwise performance					
Evaluation					
Class	n	P	R	Sp	F1
NM	26	0.963	1.000	0.971	0.981
Mild	21	1.000	0.810	1.000	0.895
Mod	7	0.625	0.714	0.944	0.667
MS	4	0.600	0.750	0.965	0.667
Severe	3	0.750	1.000	0.983	0.857
Deployment					
Class	n	P	R	Sp	F1
NM	26	0.963	1.000	0.971	0.981
Mild	21	1.000	0.810	1.000	0.895
Mod	7	0.700	1.000	0.944	0.824
MS	4	1.000	0.750	1.000	0.857
Severe	3	0.750	1.000	0.983	0.857
GAD-7 classwise performance					
Evaluation					
Class	n	P	R	Sp	F1
Min	29	0.964	0.931	0.969	0.947
Mild	20	0.905	0.950	0.951	0.927
Mod	7	1.000	0.714	1.000	0.833
Severe	5	0.714	1.000	0.964	0.833
Deployment					
Class	n	P	R	Sp	F1
Min	29	0.966	0.966	0.969	0.966
Mild	20	0.950	0.950	0.976	0.950
Mod	7	1.000	0.714	1.000	0.833
Severe	5	0.714	1.000	0.964	0.833

[c] n = test support; P = precision; R = recall/sensitivity; Sp = one-versus-rest specificity; F1 = F1 score; NM = None-Minimal; Min = Minimal; Mod = Moderate; MS = Moderately Severe. Each target-specific test partition contains 61 records. Metrics are rounded to three decimals.

D. Direct Questionnaire Scoring and Feature–Target Analysis

The feature–target audit confirmed that the original 35-feature predictor matrix included all nine PHQ-9 items and all seven GAD-7 items used to calculate the respective severity targets. Direct application of the questionnaire scoring rules reproduced the supplied PHQ-9 severity category for all 406 records and the supplied GAD-7 category for all 406 records, with no disagreements. The scoring implementation also passed all 74 recorded function tests, covering scoring boundaries and invalid-input conditions. Direct scoring without model fitting was 100.00% accurate and 1.0000 macro-F1 on each target's original 61-record test partition.

These results prove that the labels in the provided data set can be recovered deterministically from the answers to the questionnaire, which are provided as part of the original model predictors. Thus, high classification scores in the original feature

configuration should be considered primarily as the recovery of the categories defined by the questionnaire rather than as evidence of independent prediction of mental-health conditions. The separate context-only experiment was thus kept to investigate performance in the absence of the target-defining questionnaire items. Direct questionnaire scoring is not an equal-input predictive baseline for that limited experiment since it involves the same questionnaire items that are omitted from the context-only predictors.

E. Repeated Evaluation and Model Comparison

This repeated evaluation was able to process all 960 planned fits in 20 stratified random seeds, two targets, two feature sets, six model families, and two versions of fitting. None of the seeds were excluded based on performance, and all configurations were evaluated on matched target-specific held-out partitions. The HybridStack performance for the repeated partitions is summarized in Table V. Results showed that with the original 35-feature configuration, mean refitted HybridStack accuracy was 89.10% for PHQ-9 and 94.26% for GAD-7, while the evaluation-stage fits were 87.70% and 93.44%, respectively. Mean macro-F1 increased from 0.8463 to 0.8528 for PHQ-9 and from 0.9180 to 0.9212 for GAD-7. But the differences between the refit and the evaluation were not always positive and varied for each partition.

All 16 questionnaire items were excluded, which substantially lowered performance. For the context-only configuration (19 features), the mean refitted HybridStack accuracy was 39.34% for PHQ-9 and 49.92% for GAD-7, and macro-F1 scores were 0.3227 and 0.4273, respectively. The same original-feature refits, by comparison, achieved 89.10% and 94.26% accuracy. This distinction shows the reliance of the original classification results on the questionnaire responses included. The context-only evaluation is still restricted to the classification of categories defined by a questionnaire and derived from the same time-context inputs: demographic and workplace-context.

HybridStack was not consistently better with limited predictors. For PHQ-9, Logistic Regression had higher macro-F1 (0.3535 versus 0.3227), while HybridStack's accuracy was below the majority baseline (39.34% versus 42.62%). Context-only CatBoost had higher accuracy but lower macro-F1 than HybridStack. Ranking depends on features and metrics; repeated partitions represent the same 406-record survey, not independent cohorts.

TABLE V. REPEATED HYBRIDSTACK PERFORMANCE ACROSS 20 PARTITIONS - MEAN ± SAMPLE SD

Target/features	Evaluation accuracy	Refit accuracy	Evaluation macro-F1	Refit macro-F1
PHQ-9/35	87.70 ± 4.10%	89.10 ± 3.99%	0.8463 ± 0.0578	0.8528 ± 0.0584
PHQ-9/19	38.52 ± 5.79%	39.34 ± 5.45%	0.3154 ± 0.0545	0.3227 ± 0.0640
GAD-7/35	93.44 ± 4.19%	94.26 ± 3.03%	0.9180 ± 0.0542	0.9212 ± 0.0484
GAD-7/19	47.79 ± 6.21%	49.92 ± 4.53%	0.4113 ± 0.0635	0.4273 ± 0.0504

[d.] Accuracy values are percentages; macro-F1 values are unitless. Standard deviations describe variability across overlapping partitions, not confidence intervals.

Table VI compares refitted macro-F1 across all six model families under the original and context-only feature configurations.

TABLE VI. REFITTED MODEL MACRO-F1 ACROSS 20 PARTITIONS

PHQ-9		
Model	**Original 35**	**Context-only 19**
Majority	0.120 (0.000)	0.120 (0.000)
Logistic regression	0.780 (0.086)	0.354 (0.077)
Random forest	0.755 (0.082)	0.324 (0.081)
CatBoost	0.800 (0.075)	0.322 (0.078)
XGBoost	0.777 (0.069)	0.343 (0.083)
HybridStack	0.853 (0.058)	0.323 (0.064)
GAD-7		
Model	**Original 35**	**Context-only 19**
Majority	0.159 (0.002)	0.159 (0.002)
Logistic regression	0.881 (0.067)	0.394 (0.056)
Random forest	0.843 (0.049)	0.367 (0.063)
CatBoost	0.888 (0.045)	0.382 (0.062)
XGBoost	0.865 (0.044)	0.368 (0.046)
HybridStack	0.921 (0.048)	0.427 (0.050)

[e.] Values are mean macro-F1 (sample SD) across 20 overlapping partitions. Original 35 includes questionnaire items; context-only 19 excludes them. All models were refitted on 345 records per partition. SDs are not confidence intervals.

F. Probability Assessment Results

The probability-quality assessment for the original reconstructed evaluation-stage and archived deployment-stage models is summarized in Table VII. For PHQ-9, log loss decreased from 0.297838 to 0.252674, and the multiclass class-summed Brier score decreased from 0.155692 to 0.130849. For GAD-7, log loss decreased from 0.248064 to 0.182997, and the Brier score decreased from 0.138218 to 0.098301. The lower observed scores are in favour of the deployment-stage models on the respective original held-out test partitions. The results, however, are based on only 61 records per target and have not resulted in population-level superiority or calibrated clinical confidence.

Probability-quality measures were also examined in all model fits that were repeated. The repeated original-feature assessments for the HybridStack models tended to give lower mean log loss and Brier score for the re-fit models. Fixed probability bins were used for the one-versus-rest reliability summaries along with the retained bin counts. The results were analysed descriptively owing to the small number of observations in some severity classes and probability bins. No calibration model was fitted, and multiple prediction appearances of existing records were combined, not independent participants, in reliability summaries. So, the assessment enables reporting probability quality but does not resolve formal calibration of the assessment and external probability validation.

TABLE VII. ORIGINAL HELD-OUT PROBABILITY-QUALITY ASSESSMENT

Target/version	Original test log loss	Original test class-summed Brier
PHQ evaluation	0.297838	0.155692
PHQ deployment	0.252674	0.130849
GAD evaluation	0.248064	0.138218
GAD deployment	0.182997	0.098301

[f] The Brier score is the mean class-summed squared probability error per record. Lower scores indicate better observed probability quality; no calibrator was fitted.

G. Preprocessing and Inference Verification Results

The original application audit, revised input-validation checks, documentation/chart checks, and existing-record inference-parity results are summarized in Table VIII. No differences were detected between the output from the original-record preprocessing and the output from the application of the preprocessing to the records by an independent processor out of the 406 records available. Archived-model predicted labels and complete class-probability vectors also agreed with those obtained after reconstruction and re-encoding for the examined records in both targets. After the input-handling changes were documented, all 298 baseline checks passed. These results show that the input-contract faults found in the original application have been corrected and that inference parity is preserved on the available survey records.

Selected individual-screening boundary cases, valid batch processing, mixed valid/invalid batch processing, downloading of valid-row results, and updating of documentation/chart displays were accomplished with the help of separate manual interface observations. The observations are not a complete browser test or a formal usability evaluation. Similarly, the successful input-contract tests do not provide protection against arbitrary, unknown data in the workplace or future distributional shifts.

TABLE VIII. ORIGINAL AND REVISED APPLICATION VERIFICATION SUMMARY

Stage	Result
Original audit	228 passed; 70 failed out of 298
Revised baseline checks	298/298 passed
Additional CSV checks	26/26 passed
Documentation/chart checks	20/20 passed
Original-record parity	Features, labels, and full probability vectors agree on 406 records per target

H. System Overview

The verified model artifacts and preprocessing components were integrated into a web-based research prototype comprising individual screening, batch screening, and model-documentation views, as described in Section III-D. Fig. 4 presents the individual-screening interface. The application implementation is publicly available at https://github.com/akib1437/workplace-mental-health-screening-revised, with a live demonstration at https://workplace-mental-health-screening.streamlit.app/.

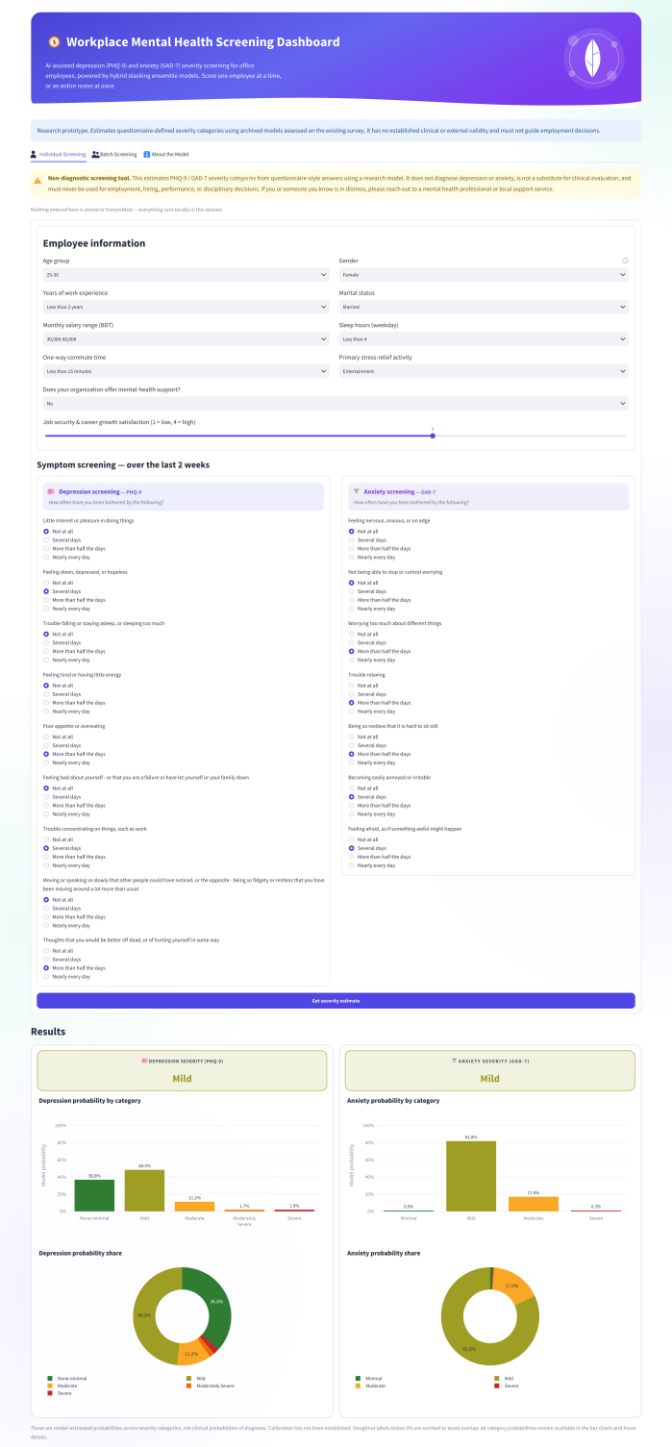

Fig. 4. Workplace mental health screening dashboard for individual employee assessment.

As shown in Fig. 4, the individual-screening interface accepts demographic, occupational, lifestyle, and questionnaire responses. The application generates questionnaire-defined PHQ-9 and GAD-7 severity classifications and corresponding model-generated class probabilities. The interface identifies the system as a non-diagnostic research prototype.

Fig. 5 illustrates the batch-screening interface, which processes multiple questionnaire records within a single submission. The application displays predicted PHQ-9 and GAD-7 severity categories and associated model-generated class probabilities for successfully processed records. The aggregated visualizations provide an overview of the severity categories predicted across records submitted. The given probabilities are not individually calibrated.

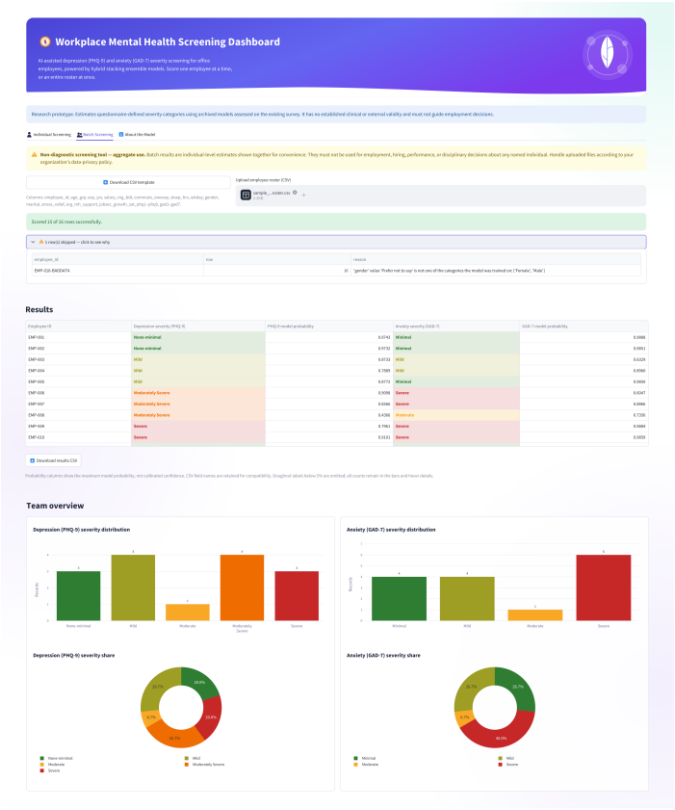

Fig. 5. Batch screening interface and team-level mental health severity overview.

The model-documentation interface shown in Fig. 6 summarizes the hybrid stacking architectures for the PHQ-9 and GAD-7 severity classifications, as well as the data characteristics and evaluation results needed to interpret these models. The architecture descriptions detail the makeup of the models and how the constituent learners and the final predictions relate to one another. The information about the training data situates these architectures in the context of the survey available, and the severity categories defined in the questionnaire explain the meaning of the labels predicted. Providing this information together with the test metrics that are held out provides users with the chance to connect reported performance with the classification tasks and the data used during model development. The interface also differentiates between the models used in the evaluation phase and the deployment phase, indicating that the results reported are from various phases of the modeling process. This distinction helps in the interpretation of the performance differences observed after refitting and prevents treating the two model versions as interchangeable. Special consideration is paid to the selection of macro-F1 as the main model selection measure since the severity categories are not equally represented. Accuracy is an overall measure summarizing the percentage of records correctly classified and can be misleading if predictions tend to cluster at the more frequent categories. However, Macro-F1 computes the F1 score for each severity class and averages them equally. Performance in less common categories thus has the same weight as performance in more common categories, although estimates on the less common categories are less certain. The documentation offers a foundation for reading accuracy and macro-F1 in conjunction with class representation, and not individually. These explanations will enable the readers to understand what the results reported describe and the range in which the model comparisons should be viewed.

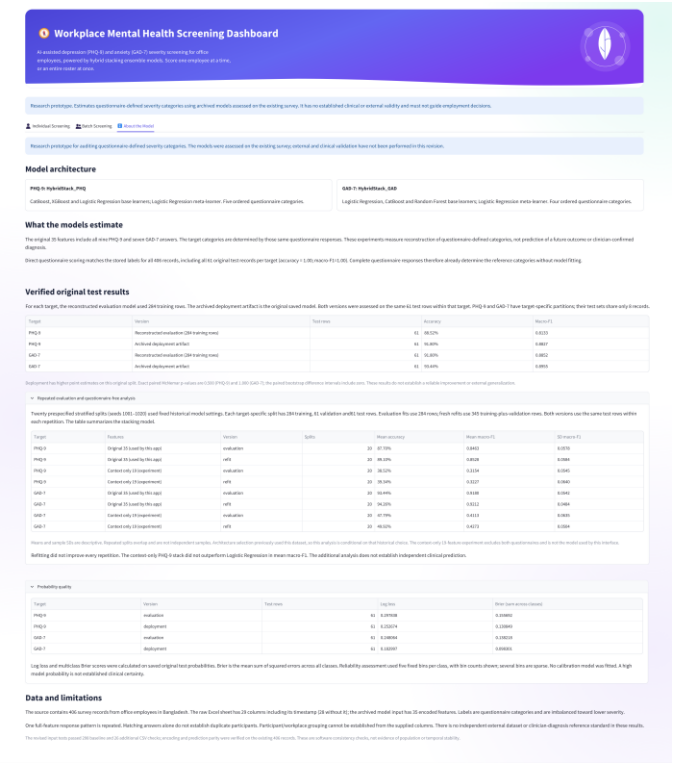

Fig. 6. Model architecture, performance metrics, and training data overview.

V. Discussion

The results point to the need to differentiate implementation consistency, reproduction of questionnaire scores, and predictive validity of the questionnaire beyond the context of deployment. Observed accuracy and macro-F1 were higher for archived deployment-stage models compared to reconstructed evaluation-stage models for both the PHQ-9 and GAD-7. Neither exact McNemar comparison was statistically significant, and all paired bootstrap intervals for the differences included zero. The extra re-evaluations also revealed slight and variable improvements after refitting, which cannot be taken for granted even when a model is fitted with extra development records.

One key finding of the updated analysis is the relationship between the predictors and the severity targets. The original set of predictors included all nine PHQ-9 items and seven GAD-7 items, and the target categories were derived from the same questionnaire items. Model-free direct scoring baseline

reproduced all 406 labels for each target. Thus, the high original-feature classification scores largely reflect the models' capacity to reconstruct categories defined in the questionnaires. If all 16 questionnaire items were left out, the HybridStack performance dropped substantially when using context-only. This finding provides further evidence for the difference between replication of questionnaire scores and prediction of independent clinical diagnoses or future mental-health outcomes. Direct scoring was in perfect agreement with the supplied labels and outperformed the machine-learning classifiers.

The repeated comparison also showed that the performance of the stacking ensembles was dependent upon the predictors that were available and on the evaluation measure. HybridStack achieved high mean accuracy and macro-F1 over the repeated partitions for the original 35 features. However, simpler models were competitive and sometimes better in context-only analysis. Logistic Regression had a higher mean PHQ-9 macro-F1 score, and CatBoost had a higher mean GAD-7 accuracy. These findings aren't necessarily proof that one type of family is best. Instead, they show that the choice of architecture should be based on the target definition, availability of the inputs, class imbalance, and the evaluation metric.

The probability-quality assessment gave more information than accuracy and macro-F1. The deployment models that were archived obtained lower observed log loss and multiclass Brier scores than the reconstructed evaluation-stage models on the original held-out partitions. Re-evaluation was carried out, and similar descriptive comparisons were made. The small class sizes and sparse reliability bins, however, prevent conclusions regarding the quality of probability. No calibration model was fitted, and the model-produced probabilities should not be regarded as validated clinical confidence measures. Such probabilities would require additional calibration assessment and evaluation of independently collected data before they could be used for higher-stakes decisions.

Lastly, the implementation audit showed that the archived model inference pipeline and the updated application matched the current records. The updated application fixed the input-validation issues first identified and passed baseline, additional CSV, and documentation/chart tests. These results validate the technical consistency of the implemented research prototype in the investigated input contract. They are not used to set up external robustness, practical trustworthiness, clinical utility, or usability in the workplace. The study's main value, therefore, lies in its clear explanation of the difference between model performance, labels derived from questionnaires, model artifacts refitted to the workplace, and actual application performance, rather than in its demonstration of clinically validated prediction of workplace mental health.

VI. Limitations

There are some limitations in the data set, the construction of the target, the history of model development, and the validation scope. The analysis is based on one cross-sectional survey of 406 office workers in the banking and information technology (IT) sectors of Bangladesh. The sample is relatively small, has some demographic imbalance, and does not provide adequate information on participants or workplace to determine independence or clustering beyond what is provided in the records. The PHQ-9 and GAD-7 severity labels were directly obtained from questionnaire responses, which were also part of the original model predictors. Direct scoring was able to replicate the provided labels completely, but this agreement is not necessarily an agreement that the classifications are the same as those independently confirmed clinically. The context-only analysis used the same targets that were obtained from the questionnaire, which were contemporary to the study, and thus did not provide any prospective predictions of future mental-health conditions.

The original deployment comparison was based on a single held-out test partition with 61 records per target, and a small number of discordant correctness results (two for PHQ-9 and three for GAD-7). For both PHQ-9 and GAD-7, the last partitions contained only three PHQ-9 Severe cases and five GAD-7 Severe cases. As a result, the original comparisons were only partially informative for determining model differences or for estimating the performance of the rare classes. There were 20 more stratified partitions that were tested, but they used the same 406 records and were not independent validation cohorts. Moreover, the architectures previously selected were identified based on this data set, and a complete re-evaluation of model selection in history was not done. The repeated results should thus not be considered as separate confirmation of the superiority of the models, but as internal sensitivity analyses.

The number of samples was small, as was the number of samples in each severity class and the reliability bins, which limited the probability assessment. No model calibration for probability was fit. The assessment therefore characterizes observed probability-quality measures and does not provide calibrated confidence. The revised application has gone through documented preprocessing, input validation, and documentation/chart tests and passed, but these tests do not imply robustness to arbitrary unfamiliar inputs or changes in the underlying population of the workplace. Manual interface observations were only performed for selected workflows, and formal usability testing with employees, clinicians, or HR practitioners was not carried out. The interface is still only available in English, too.

Most significantly, external and temporal validation were not completed as there was no comparable independently collected workplace data available for the present revision. Internal validation (through repeated divisions of the original survey) and synthetic application tests are not a replacement for external validation. The generalizability to the population evaluated in the study, stability in future changes in the distribution of scores, the performance of subgroups, and independent clinical validity are therefore not established. The application is still a non-diagnostic research prototype and should not be used in making individual decisions regarding employment, hiring, disciplinary, or other consequential decisions.

VII. Conclusion and Future Work

In this study, 406 Bangladeshi office workers were surveyed to assess the performance of the models, their severity classification based on the questionnaire, and the consistency of their implementation when using hybrid stacking ensembles. The original target-specific held-out comparisons showed that

the observed accuracy and macro-F1 of archived deployment-stage models were higher than the corresponding evaluation-stage values, but the difference did not demonstrate statistically significant accuracy improvement. A deterministic relationship between the questionnaire inputs and the resulting severity categories was confirmed by the ability of a direct questionnaire-scoring baseline to reproduce all 406 of the severity labels supplied for each target. Further replications on 20 stratified partitions and six model families revealed that performance as observed varied with feature availability and that the results of refitting did not follow a consistent pattern. The 19-feature context-only analysis showed substantially lower performance than the 35-feature original model and indicated that stacking was not always better than simpler models.

Probability-quality assessment and extended application testing provided further evidence for archived model outputs and behavior of application implementation. The updated version maintained the consistency of features, predictions, and probability vectors in the available records and passed the documented validation checks. These findings, however, do not provide evidence of calibrated clinical confidence, independent validity, temporal stability, or generalizability to new workplaces. Future work should focus on external evaluation based on data independently collected, prospective outcomes, assessment of demographic sub-groups, formal probability calibration and validation, and structured usability studies with appropriate practitioners and end users, and broader assessment of distributional robustness. The current results suggest that the use of the questionnaire-derived machine learning classifications should be interpreted with care and should not be used as clinically validated screening decisions in the workplace.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the Department of Computer Science, American International University-Bangladesh (AIUB), for providing institutional support and research facilities that made this work possible.

DECLARATIONS

- Conflict of interest: The authors confirm that they have no known competing financial interests or personal relationships that could influence the work described in this study.
- Data availability: The revised application source code, preprocessing pipeline, and deployment models are publicly available at: https://github.com/akib1437/workplace-mental-health-screening-revised. A live demonstration is available at https://workplace-mental-health-screening.streamlit.app/.
- Funding: This work was not supported by any grants from public or private funding agencies.

REFERENCES

[1] R. Rugulies, B. Aust, B. A. Greiner, E. Arensman, N. Kawakami, A. D. LaMontagne, and I. E. H. Madsen, "Work-related causes of mental health conditions and interventions for their improvement in workplaces," The Lancet, vol. 402, no. 10410, pp. 1368–1381, Oct. 2023, doi: 10.1016/S0140-6736(23)00869-3.

[2] A. Khalid and J. Syed, "Mental health and well-being at work: a systematic review of literature and directions for future research," Human Resource Management Review, vol. 34, no. 1, p. 100998, Mar. 2024, doi: 10.1016/j.hrmr.2023.100998.

[3] Organisation for Economic Co-operation and Development (OECD), "Promoting health and well-being at work: policy and practices," OECD Health Policy Studies, OECD Publishing, Paris, France, 2022, doi: 10.1787/e179b2a5-en.

[4] B. Aust et al., "How effective are organizational-level interventions in improving the psychosocial work environment, health, and retention of workers? A systematic overview of systematic reviews," Scand. J. Work Environ. Health, vol. 49, no. 5, pp. 315–329, Jul. 2023, doi: 10.5271/sjweh.4097.

[5] A. Waddell, B. Kunstler, A. Lennox, L. Pattuwage, E. A. C. Grundy, D. Tsering, P. Olivier, and P. Bragge, "How effective are interventions in optimizing workplace mental health and well-being? A scoping review of reviews and evidence map," Scand. J. Work Environ. Health, vol. 49, no. 4, pp. 235–248, May 2023, doi: 10.5271/sjweh.4087.

[6] S. Chowdhary, A. Kawakami, M. L. Gray, J. Suh, A. Olteanu, and K. Saha, "Can workers meaningfully consent to workplace wellbeing technologies?" in Proc. 2023 ACM Conf. Fairness, Accountability, and Transparency (FAccT), Chicago, IL, USA, 2023, pp. 569–582, doi: 10.1145/3593013.3594023.

[7] A. B. Bakker and E. Demerouti, "The Job Demands-Resources model: state of the art," J. Manag. Psychol., vol. 22, no. 3, pp. 309–328, Apr. 2007, doi: 10.1108/02683940710733115.

[8] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, "The PHQ-9: validity of a brief depression severity measure," J. Gen. Intern. Med., vol. 16, no. 9, pp. 606–613, Sep. 2001, doi: 10.1046/j.1525-1497.2001.016009606.x.

[9] R. L. Spitzer, K. Kroenke, J. B. W. Williams, and B. Löwe, "A brief measure for assessing generalized anxiety disorder: the GAD-7," Arch. Intern. Med., vol. 166, no. 10, pp. 1092–1097, May 2006, doi: 10.1001/archinte.166.10.1092.

[10] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," BMC Bioinformatics, vol. 7, art. no. 91, Feb. 2006, doi: 10.1186/1471-2105-7-91.

[11] G. C. Cawley and N. L. C. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," J. Mach. Learn. Res., vol. 11, pp. 2079–2107, 2010.

[12] G. S. Collins et al., "TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods," BMJ, vol. 385, e078378, Apr. 2024, doi: 10.1136/bmj-2023-078378.

[13] K. G. M. Moons et al., "PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods," BMJ, vol. 388, e082505, Mar. 2025, doi: 10.1136/bmj-2024-082505.

[14] P. Varje, A. Väänänen, O. Haavisto, I. Kivimäki, S. Taimela, and T. Kalliomäki-Levanto, "Machine learning in the analysis of mental health at work: a scoping review," Journal of Occupational Health, vol. 68, no. 1, p. uiag014, 2026, doi: 10.1093/joccuh/uiag014.

[15] A. Thieme, D. Belgrave, and G. Doherty, "Machine learning in mental health: a systematic review of the HCI literature to support the development of effective and implementable ML systems," ACM Trans. Comput.-Hum. Interact., vol. 27, no. 5, pp. 1–53, Oct. 2020, doi: 10.1145/3398069.

[16] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, and D. Dennison, "Hidden technical debt in machine learning systems," in Advances in Neural Information Processing Systems 28 (NeurIPS 2015), 2015, pp. 2503–2511.

[17] D. H. Wolpert, "Stacked generalization," Neural Networks, vol. 5, no. 2, pp. 241–259, 1992, doi: 10.1016/S0893-6080(05)80023-1.

[18] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," in Advances in Neural Information Processing Systems 31 (NeurIPS 2018), 2018, pp. 6638–6648.

[19] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[20] F. Pedregosa et al., "Scikit-learn: machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[21] National Institute of Standards and Technology (NIST), "Artificial intelligence risk management framework (AI RMF 1.0)," NIST AI 100-1, 2023, doi: 10.6028/NIST.AI.100-1.

[22] World Health Organization, "WHO guidelines on mental health at work," Geneva, Switzerland: World Health Organization, 2022, ISBN: 978-92-4-005305-2.

[23] R. Buga, C. Gh. Buzea, M. Agop, L. Ochiuz, D. Vasincu, O. Popa, D. I. Rusu, I. Știrban, and L. Eva, "Streamlit application and deep learning model for brain metastasis monitoring after Gamma Knife treatment," Biomedicines, vol. 13, no. 2, p. 423, Feb. 2025, doi: 10.3390/biomedicines13020423.

[24] D. K. Assumang, E. Umah, M. M. Vincent, B. I. Abolade, and B. E. Shasere, "A data-driven dashboard framework for employee mental health and organisational decision-making," Journal of Scientific Research and Reports, vol. 31, no. 7, pp. 979–1000, Jul. 2025, doi: 10.9734/jsrr/2025/v31i73315.

[25] R. Saunders et al., "Measurement invariance of the PHQ-9 and GAD-7 across males and females seeking treatment for common mental health disorders," BMC Psychiatry, vol. 23, art. no. 298, Apr. 2023, doi: 10.1186/s12888-023-04804-x.

[26] H. Delamain et al., "Measurement invariance and differential item functioning of the PHQ-9 and GAD-7 between working-age and older adults seeking treatment for common mental disorders," J. Affect. Disord., vol. 347, pp. 15–22, Feb. 2024, doi: 10.1016/j.jad.2023.11.048.

[27] "Instructions for Patient Health Questionnaire (PHQ) and GAD-7 measures," instruction manual, n.d. [Online]. Available: https://www.uab.edu/medicine/pcp-sci/images/SCIMS/PHQ-9_Instruction_Manual.pdf. [Accessed: Oct. 23, 2026].

[28] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," Psychometrika, vol. 12, no. 2, pp. 153–157, Jun. 1947, doi: 10.1007/BF02295996.

[29] B. Efron and R. J. Tibshirani, An Introduction to the Bootstrap. New York, NY, USA: Chapman & Hall, 1993.

[30] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," Neural Comput., vol. 10, no. 7, pp. 1895–1923, Oct. 1998, doi: 10.1162/089976698300017197.

[31] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," Journal of the American Statistical Association, vol. 102, no. 477, pp. 359–378, 2007, doi: 10.1198/016214506000001437.