Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

A Leakage-Controlled Hierarchical Evaluation Framework for Breast Cancer Variant Pathogenicity Prediction with Locked-Test Validation

Dilip Sisodia[1,2] , Virendra Kumar Sharma[3]
[1] Research Scholar, Department of Computer Science & Engineering, Bhagwant University, Ajmer, Rajasthan, India
[2] Assistant Professor, Department of Computer Science & Engineering, Engineering College, Ajmer, Rajasthan, India
[3] Vice Chancellor, Bhagwant University, Ajmer, Rajasthan, India

Abstract*—Reliable evaluation of genomic classifiers requires more than a random separation of variant rows, particularly when several observations originate from the same patient. In this study, we evaluated breast-cancer variant pathogenicity prediction using 4,999 annotated variants obtained from 75 sequencing samples representing 50 patient groups. Exact ClinVar matching provided 106 pathogenic/likely-pathogenic and 4,893 benign/likely-benign observations, while clinical significance itself was not used as a predictor. The 26 input predictors produced 71 encoded variables. TRAIN, VALIDATION, and the locked TEST were formed with no overlap in patient groups or exact alleles. Because the positive class accounted for only 2.12% of the dataset, Average Precision (AP) was specified as the principal ranking measure. Fourteen classical machine-learning models and 13 deep architectures were evaluated. Deep-model family selection used repeated seeds, and both the final model and operating threshold were fixed from VALIDATION evidence before TEST access. HistGradientBoosting was ultimately retained. On the locked TEST, AP was 0.6105 (95% CI 0.3531-0.8114), ROC-AUC was 0.9730 (0.9449-0.9977), and MCC was 0.6627 (0.3961-0.7488). Subsequent analyses used the frozen procedure to examine uncertainty, calibration, patient dependence, annotation dependence, selection stability, unseen-gene performance, and deep-model seed variability. TEST data were excluded from XAI, and independent audits checked the declared evaluation boundaries. Although discrimination was strong, the small positive class produced wide uncertainty, probability estimates showed systematic overprediction, and performance declined markedly when annotation-rich information was removed. The contribution of this study is thus the evaluation framework and its auditable separation of development from final testing, rather than the introduction of another classifier.

Keywords—breast cancer genomics; variant pathogenicity; data leakage; Average Precision; locked test set; explainable machine learning

I. INTRODUCTION

Next-generation sequencing produces many variant observations, but those observations are not necessarily independent. In our data, multiple variants could originate from the same patient and therefore share biological background, preprocessing history, and annotation characteristics. The same allele could also occur in more than one sample. For this reason, a conventional row-level split was not appropriate for the question addressed here. We wanted the final estimate to represent performance on patients not encountered during development. Patient groups and exact alleles were consequently kept separate across partitions. This design follows the broader principle that the evaluation split should reflect the intended unit of generalization and that final-test information should not influence development decisions [18]. Class imbalance was another practical constraint: among 4,999 observations, only 106 (2.12%) belonged to the pathogenic/likely-pathogenic class. Under this imbalance, accuracy can favor the majority class and ROC-AUC may remain high even when precision among positive predictions is modest. AP was therefore chosen before model comparison [17]. We also considered a single TEST point estimate insufficient because the locked TEST contains only 11 positives from seven patient groups. This motivated patient-group bootstrap intervals and a separate post-lock calibration analysis rather than further tuning after TEST access. The ClinVar-derived target introduces a separate interpretive concern. Clinical significance was excluded from the predictors, but consequence annotations, SIFT/PolyPhen scores, dbSNP evidence, and related variables can still overlap conceptually with curated pathogenicity knowledge. We therefore do not describe the task as de novo biological pathogenicity prediction. After model locking, feature-family ablation and unseen-gene analysis were used to determine how strongly performance depended on annotation-rich evidence and gene context. Earlier work from this research program addressed automated NGS processing [1], literature synthesis [2], a separate balanced TCGA-BRCA multi-omics experiment [3], and an explainable genomic-AI architecture [4]. The present study asks a different question: can the evaluation pathway itself be made auditable? Development decisions were confined to TRAIN and VALIDATION, deep candidates were checked across seeds, and the final model and threshold were fixed before TEST was opened. Afterward, uncertainty and calibration used frozen evidence, while sensitivity analyses and XAI remained outside TEST. This structure lets us evaluate not only performance, but whether the reported estimate was obtained without TEST feedback.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

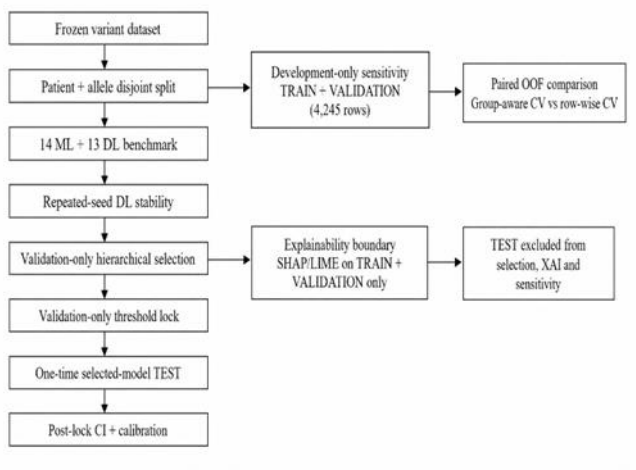

Fig. 1. Leakage-controlled hierarchical evaluation framework. The primary pathway uses patient- and exact-allele-disjoint partitions, validation-only hierarchical model selection and threshold locking, one-time selected-model TEST evaluation, post-lock uncertainty/calibration, development-only sensitivity, and a TEST-isolated explainability boundary.

II. Related Work And Background

ree-based ensemble methods remain useful reference models for structured biomedical datasets. Random Forest [7], gradient boosting [8], XGBoost [9], LightGBM [10], CatBoost [11], and histogram-based boosting differ in implementation but all provide practical ways of modelling nonlinear relationships in tabular predictors. Deep tabular modelling has meanwhile expanded beyond conventional multilayer perceptrons to architectures such as Transformer-based models [12], FT-Transformer [13], TabNet [14], autoencoder classifiers, and gated residual networks. Their added flexibility also introduces another source of variation: two training runs of the same architecture need not produce the same ranking. We therefore considered performance across several seeds rather than selecting a deep-model family from a single favourable run. Model explanation raises a different issue. SHAP [15] and LIME [16] can help identify which inputs influence a fitted model, but their outputs should not be interpreted as evidence of biological causation. There is also an evaluation concern: examining TEST explanations while development choices are still open can indirectly feed TEST information back into the analysis. We therefore treated explainability as a separate evidence layer and kept TEST observations outside the SHAP/LIME workflow. The methodological gap addressed here is therefore not merely which model wins; it is the lack of an end-to-end contract that simultaneously controls biological dependency, rare-class metric choice, stochastic model-selection variability, threshold tuning, final-test multiplicity, uncertainty, calibration, annotation semantics, generalization claims, and explainability leakage.

III. MATERIALS AND METHODS

A. *Dataset, target, predictors and partitions*

The frozen predictive dataset contained 4,999 variant observations from 75 breast-cancer sequencing samples resolved to 50 patient groups. Exact ClinVar clinical-significance matching with minimum two-star review status defined pathogenic/likely-pathogenic positives and benign/likely-benign negatives [5]. Clinical significance, patient/group identifiers, chromosome-position-allele identifiers, canonical key, gene symbol/HGNC identifiers, and split labels were excluded from prediction. Twenty-six original predictors covered variant/consequence identity, transcript/feature context, SnpEff impact and annotation burden [19], VEP SIFT/PolyPhen scores [20], sequence-change flags, consequence flags, dbSNP presence, and score-missingness indicators; learned preprocessing expanded them to 71 encoded variables. Upstream QC, alignment, calling, and annotation followed the previously frozen NGS workflow [1]. Partitioning was performed at the patient-group level: 35 groups (3,406 rows) to TRAIN, eight groups (839 rows) to VALIDATION, and seven groups (754 rows) to TEST. Patient-group overlap and exact canonical-allele overlap were both zero across partitions. TRAIN supported grouped cross-validation and fitting; VALIDATION supported candidate comparison and final threshold selection; TEST remained sealed until the model identity and threshold were frozen.

TABLE I. PATIENT-GROUPED DATASET PARTITIONS

Partition	Rows	Groups	Pos	Neg	Prev.
TRAIN	3,406	35	77	3329	2.26%
VALIDATION	839	8	18	821	2.15%
TEST	754	7	11	743	1.46%

TABLE II. FROZEN PREDICTOR FAMILIES

Predictor family	Original predictors
Variant/consequence	Variant_Type; Worst_Impact; SnpEff_variant_type; SnpEff_selected_impact
Transcript/context	SnpEff_selected_transcript_biotype; SnpEff_selected_feature_type; NCBI_Gene_Type
Annotation scores	SnpEff_annotation_count; vep_sift_score; vep_polyphen_score
Sequence flags	SnpEff_is_snv; SnpEff_is_indel; SnpEff_is_transition; SnpEff_is_transversion
Impact/consequence flags	High/moderate/low/modifier impact; protein-coding; loss-of-function; splice; missense; synonymous
External/missingness	dbSNP_Present; vep_sift_score_missing; vep_polyphen_score_missing

B. *Candidate benchmark, metrics and deep-learning stability*

Fourteen classical candidates were evaluated: Dummy, Logistic Regression, SVC, Decision Tree, Random Forest, Gaussian Naive Bayes, Extra Trees, HistGradientBoosting, XGBoost, LightGBM, CatBoost, k-Nearest Neighbors, Gradient Boosting, and AdaBoost. Five-fold grouped development evaluation used patient-group identity. Candidate class metrics used a common descriptive threshold of 0.5; candidate-specific threshold optimization was prohibited during ranking. Thirteen deep architectures were evaluated: MLP, Deep MLP, Residual MLP, Wide & Deep, Self-Normalizing Network, Gated Residual Network, Deep Cross Network, Autoencoder and Denoising Autoencoder classifiers, TabTransformer, FT-Transformer, Transformer Encoder baseline, and TabNet. Pre-lock family selection used seeds 42, 43, and 44; the deep champion was chosen by mean held-out validation AP, with AP SD, inner-CV AP, Brier score, and model identity as stability/context evidence. The 27-model visualization uses primary-seed deep predictions, while the family decision uses repeated-seed means. A later 10-seed analysis of the frozen top five architectures was sensitivity-only and could not reopen selection. AP was primary because of 2.12% prevalence; ROC-AUC, Brier score, MCC, balanced accuracy, sensitivity, specificity, precision, F1, and ECE10 were secondary. Calibration-in-the-large and logistic calibration intercept/slope were reserved for post-lock TEST description. MCC used all four confusion-matrix cells, and the Brier score summarized mean squared probability error.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

C. *Validation-only hierarchy, TEST boundary and explainability*

The highest-ranked classical model and repeated-seed deep champion were first identified within their families without TEST. Those finalists were then compared using VALIDATION evidence only. After model identity was fixed, a validation-only threshold search locked 0.9769682788 (validation MCC 0.7377, sensitivity 0.7222, specificity 0.9951, precision 0.7647). TEST was then opened once for the selected model only. No 27-model TEST leaderboard, post-TEST candidate replacement, refit, or threshold re-optimization was allowed. SHAP and LIME were restricted to TRAIN and VALIDATION: 3,406 and 839 rows, respectively, with 71 encoded features mapped back to 26 original predictors. Global SHAP stability used train-validation rank correlation and top-k overlap; 15 validation cases spanning true-positive, false-positive, false-negative, threshold-boundary, and low-probability cases supported local SHAP/LIME comparison. TEST rows, labels, predictions, SHAP values, and LIME explanations were not read or generated by XAI.

D. *Post-lock uncertainty, calibration, leakage sensitivity and robustness*

Patient-group bootstrap uncertainty used 2,000 requested replicates on already frozen TEST predictions, resampling patient_group_id so each group moved together. Percentile 95% CIs were reported for AP, ROC-AUC, Brier, MCC, sensitivity, specificity, precision, and F1; undefined class-dependent metrics were omitted only for the affected replicate. Calibration used frozen TEST probabilities only: the original 10-bin table was reproduced, then Brier, ECE10, calibration-in-the-large, and joint logistic intercept/slope were estimated with patient-group bootstrap (seed 4610; 1,687 estimable logistic replicates). No recalibration mapping or threshold change was permitted. Development-only leakage sensitivity used the combined TRAIN+VALIDATION population (4,245 rows; 43 groups). An initial 50 repeated row-level holdout analysis was retained as supplementary sensitivity evidence. The stronger paired experiment compared five-fold StratifiedGroupKFold with conventional StratifiedKFold on exactly the same rows across 10 repeats (base seed 5500), yielding one OOF probability per row/method/repeat. Patient and exact-allele overlap were measured fold-wise; paired AP difference was primary. Five post-lock robustness analyses were also development-only and could not change the frozen procedure: (1) 20x5 grouped OOF semantic-dependence ablation comparing FULL_26, NO_IN_SILICO_22, LOW_SEMANTIC_10, and BASIC_SEQUENCE_5; (2) 10x5 grouped CV stability of all 14 frozen classical candidates; (3) portability of the locked threshold against repeat-specific MCC-optimal thresholds from 20 FULL_26 OOF repeats; (4) a patient-disjoint unseen-gene subset analysis after strict simultaneous patient+gene disjoint splitting proved infeasible because 43 groups and 2,041 genes formed one connected component; and (5) 10-seed evaluation (42-51) of the frozen top five deep architectures, reusing seeds 42-44 and fitting only 45-51.

E. *Reproducibility safeguards*

Frozen probability checks, SHAP additivity checks, encoded-to-original attribution conservation, model-refit checks, hash pins, and TEST-access counters were retained as machine-readable artifacts. Independent read-only audits verified source identities, split contracts, metric reconstruction, output hashes, and declared TEST isolation. The publication-layer package records program/configuration/marker hashes and an evidence map; these safeguards establish adherence to the evaluation contract but do not substitute for held-out predictive evidence.

IV. RESULTS

A. *Validation results and deep-model family selection*

The validation cohort contained 839 observations, including 18 positives. Among the classical models, HistGradientBoosting produced the highest validation AP (0.6852), with LightGBM (0.6675) and XGBoost (0.6242) following it. The separation between these AP values provided the initial classical-model ranking. Several models had strong ROC-AUC but materially lower AP; Logistic Regression, for example, reached ROC-AUC 0.9729 but AP 0.5789, illustrating why AP rather than ROC-AUC or accuracy governed ranking. On the primary deep seed, Gated Residual Network had AP 0.7067 and TabNet 0.6969. Across seeds 42/43/44, however, Transformer Encoder achieved the highest mean AP 0.658599 with SD 0.011267, narrowly exceeding Gated Residual Network mean 0.658536 but with far lower variability (SD 0.060313). Thus the pre-specified stability rule changed the deep-family winner.

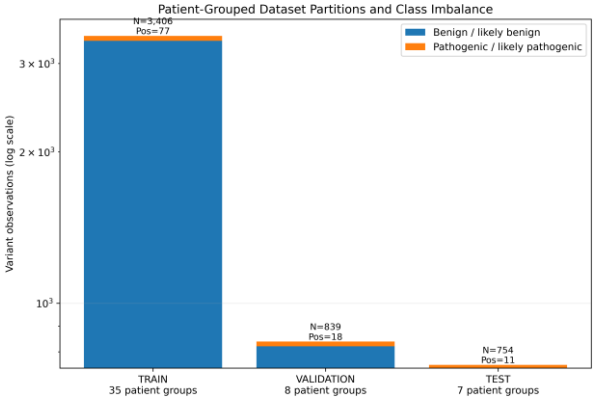

Fig. 2. Patient-grouped dataset partitions and severe class imbalance. The final TEST contained only 11 positive observations, reinforcing AP-focused evaluation and cautious interpretation of sensitivity.

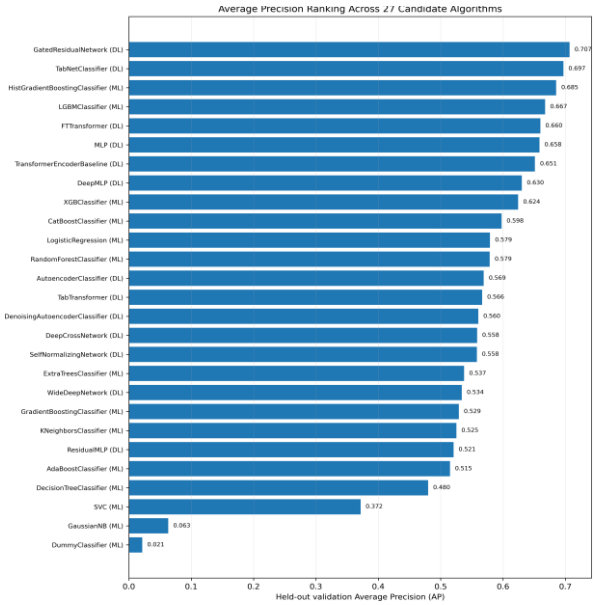

Fig. 3. Held-out validation Average Precision for all 27 candidate algorithms. Deep-learning values are primary-seed direct predictions for cross-model visualization; repeated-seed means governed deep-family selection.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

TABLE III. CLASSICAL VALIDATION METRICS – RANKING / PROBABILITY METRICS

Model	AP	ROC-AUC	Brier
HGB	0.6852	0.9545	0.0176
LGBM	0.6675	0.9435	0.0165
XGB	0.6242	0.9586	0.0188
CatBoost	0.5977	0.9406	0.0191
LR	0.5789	0.9729	0.0700
RF	0.5785	0.9524	0.0129
ET	0.5375	0.8693	0.0168
GB	0.5291	0.9530	0.0120
KNN	0.5251	0.8759	0.0124
AdaBoost	0.5149	0.9340	0.0526
DT	0.4799	0.8561	0.0195
SVC	0.3717	0.9660	0.0158
GNB	0.0632	0.8242	0.2908
Dummy	0.0215	0.5000	0.0210

TABLE IV. CLASSICAL VALIDATION METRICS - THRESHOLD 0.5 CLASS

Model	MCC	Sens.	Spec.
HGB	0.6514	0.7778	0.9866
LGBM	0.6381	0.7778	0.9854
XGB	0.5790	0.7222	0.9829
CatBoost	0.5679	0.7222	0.9817
LR	0.3750	0.9444	0.8977
RF	0.6026	0.6111	0.9915
ET	0.5935	0.6667	0.9878
GB	0.6594	0.6667	0.9927
KNN	0.6629	0.6111	0.9951
AdaBoost	-0.0051	0.0000	0.9988
DT	0.6030	0.7222	0.9854
SVC	-0.0072	0.0000	0.9976
GNB	0.2032	0.9444	0.7040
Dummy	0.0000	0.0000	1.0000

Abbreviations: HGB=HistGradientBoosting; LR=Logistic Regression; RF=Random Forest; ET=Extra Trees; GB=Gradient Boosting; KNN=k-nearest neighbors; DT=Decision Tree; GNB=Gaussian Naive Bayes.

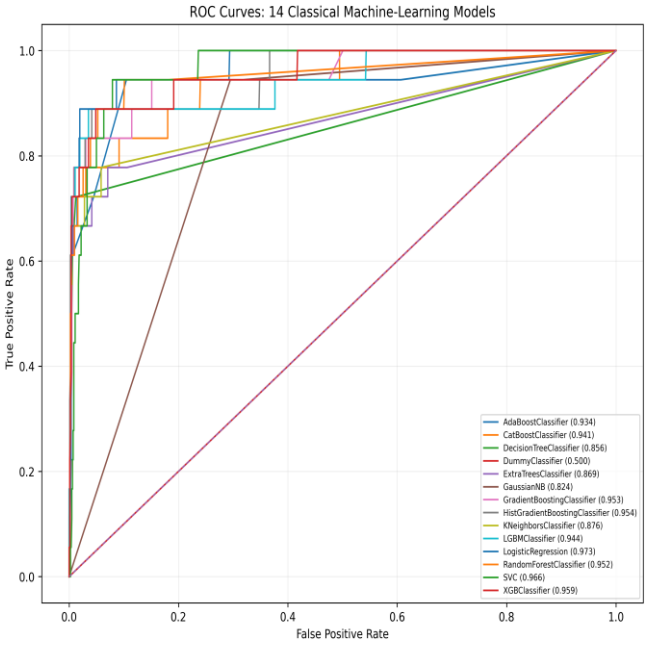

Fig. 4. ROC curves for the 14 classical machine-learning candidates on the common held-out validation cohort.

TABLE V. DEEP VALIDATION METRICS - PRIMARY SEED, RANKING/PROBABILITY METRICS

Architecture	AP	ROC-AUC	Brier
GRN	0.7067	0.9758	0.0581
TabNet	0.6969	0.9718	0.0595
FT-Tr.	0.6599	0.9761	0.0363
MLP	0.6584	0.9751	0.0691
Tr. Enc.	0.6512	0.9538	0.0827
Deep MLP	0.6301	0.9505	0.1105
AE	0.5690	0.9745	0.0552
TabTr.	0.5665	0.9569	0.0562
DAE	0.5604	0.9736	0.0623
DCN	0.5584	0.9707	0.0930
SNN	0.5580	0.9752	0.0838
Wide&Deep	0.5338	0.9651	0.0528
Res. MLP	0.5206	0.9457	0.0583

TABLE VI. DEEP VALIDATION METRICS - PRIMARY SEED, THRESHOLD 0.5 CLASS METRICS

Architecture	MCC	Sens.	Spec.
GRN	0.4222	0.9444	0.9208
TabNet	0.3918	0.8889	0.9184
FT-Tr.	0.5078	0.8889	0.9562
MLP	0.4574	0.9444	0.9342
Tr. Enc.	0.3626	0.9444	0.8904
Deep MLP	0.3884	0.9444	0.9050
AE	0.4165	0.9444	0.9184
TabTr.	0.4373	0.9444	0.9269
DAE	0.3981	0.9444	0.9099
DCN	0.3061	0.9444	0.8477
SNN	0.3458	0.9444	0.8794
Wide&Deep	0.3812	0.7778	0.9354
Res. MLP	0.4181	0.8889	0.9294

Abbreviations: GRN=Gated Residual Network; Tr. Enc.=Transformer Encoder; FT-Tr.=FT-Transformer; AE=Autoencoder; DAE=Denoising Autoencoder; DCN=Deep Cross Network; SNN=Self-Normalizing Network.

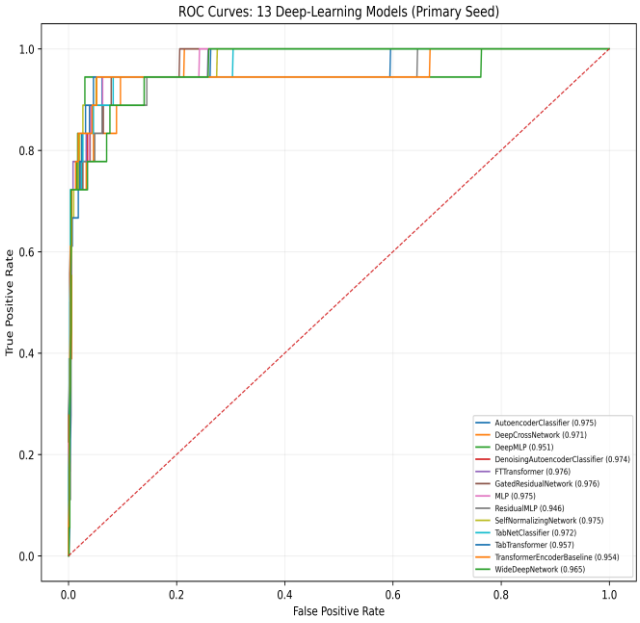

Fig. 5. ROC curves for 13 deep-learning candidates on the primary-seed validation snapshot.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

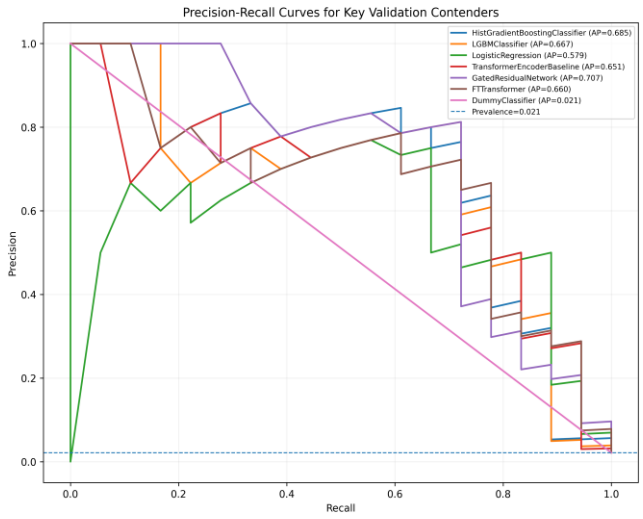

Fig. 6. Precision-recall curves for representative top contenders and references. The horizontal prevalence line emphasizes the low positive base rate.

TABLE VII. INITIAL THREE-SEED DEEP-LEARNING STABILITY - AP STABILITY

Rank	Architecture	Mean AP	AP SD
1	Tr. Encoder	0.658599	0.011267
2	GRN	0.658536	0.060313
3	Deep MLP	0.646606	0.015442
4	TabNet	0.643573	0.060400
5	FT-Tr.	0.632519	0.047309
6	MLP	0.590738	0.058798
7	DCN	0.582158	0.025473
8	TabTr.	0.581520	0.026621
9	AE	0.577459	0.020857
10	DAE	0.571213	0.013744
11	SNN	0.562336	0.003826
12	Wide&Deep	0.556722	0.043980
13	Res. MLP	0.548924	0.029258

TABLE VIII. INITIAL THREE-SEED DEEP-LEARNING STABILITY - SECONDARY METRICS

Architecture	Mean ROC-AUC	Mean Brier
Tr. Encoder	0.963042	0.074975
GRN	0.974444	0.074554
Deep MLP	0.966764	0.091357
TabNet	0.959749	0.057135
FT-Tr.	0.973587	0.045846
MLP	0.969403	0.071804
DCN	0.972718	0.074494
TabTr.	0.959501	0.055734
AE	0.973282	0.053149
DAE	0.972718	0.062727
SNN	0.968726	0.086505
Wide&Deep	0.967170	0.047332
Res. MLP	0.933493	0.054248

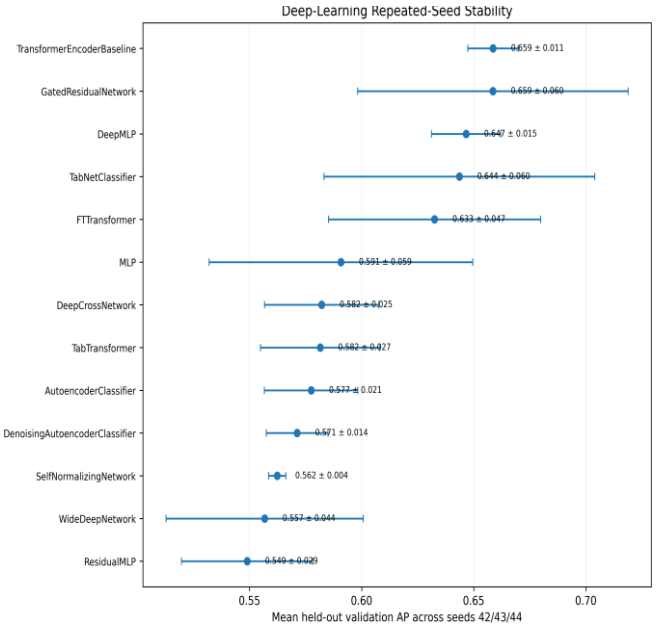

Fig. 7. Initial mean validation AP +/- standard deviation across seeds 42/43/44. The Transformer Encoder combined the highest mean AP with substantially lower variability than the Gated Residual Network; this original three-seed rule governed the frozen deep-family selection.

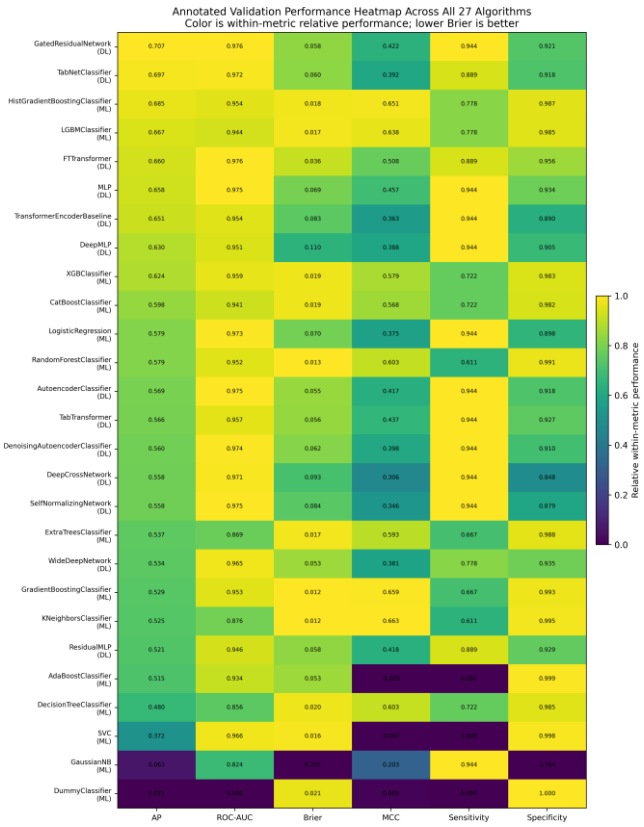

Fig. 8. Annotated validation performance heatmap for all 27 algorithms. Every cell contains the observed metric value; lower Brier is better

B. Hierarchical selection, threshold locking and one-time TEST

The classical finalist was HistGradientBoosting (validation AP 0.68523; Brier 0.01765), while the repeated-seed deep finalist was Transformer Encoder (mean AP 0.658599; SD 0.011267). Validation-only hierarchical comparison therefore

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

retained HistGradientBoosting. The locked threshold was 0.976968279. On the one-time TEST of 754 rows from seven patient groups (11 positives, 743 negatives), AP was 0.61048, ROC-AUC 0.97296, Brier 0.01400, ECE10 0.01959, and MCC 0.66274; sensitivity was 0.63636 and specificity 0.99596, with TP=7, TN=740, FP=3, FN=4. No model or threshold was changed afterward.

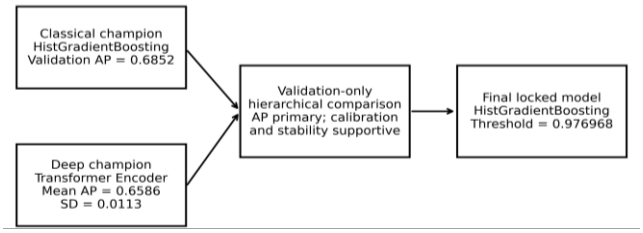

Fig. 9. Hierarchical final selection. The classical and deep-family champions were compared using VALIDATION evidence only, after which the final model and operating threshold were frozen.

TABLE IX. HIERARCHICAL SELECTION, THRESHOLD LOCKING AND LOCKED TEST SUMMARY

Evidence layer	Model/value	Supporting result
Classical champion	HGB	AP=0.68523; Brier=0.01765
Deep champion	Transformer Encoder	Mean AP=0.658599; SD=0.011267
Locked threshold	0.976968279	Validation MCC=0.7377; Sens=0.7222; Spec=0.9951
Final TEST cohort	754 rows / 7 groups	11 positive; 743 negative
Final TEST discrimination	AP=0.61048	ROC-AUC=0.97296
Final TEST calibration	Brier=0.01400	ECE10=0.01959
Final TEST operating point	MCC=0.66274	Sensitivity=0.63636; Specificity=0.99596
Final TEST confusion	TP=7, TN=740	FP=3, FN=4

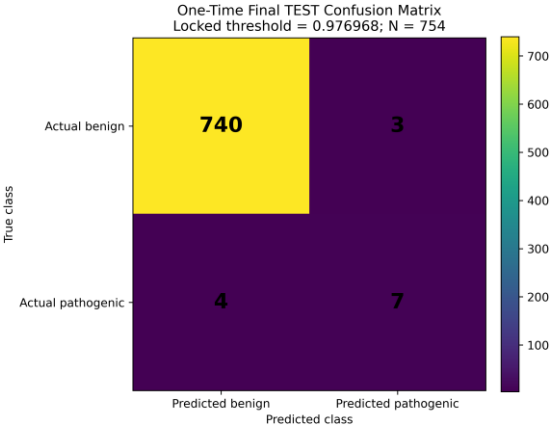

Fig. 10. Final TEST confusion matrix for the selected HistGradientBoosting model at the pre-locked threshold.

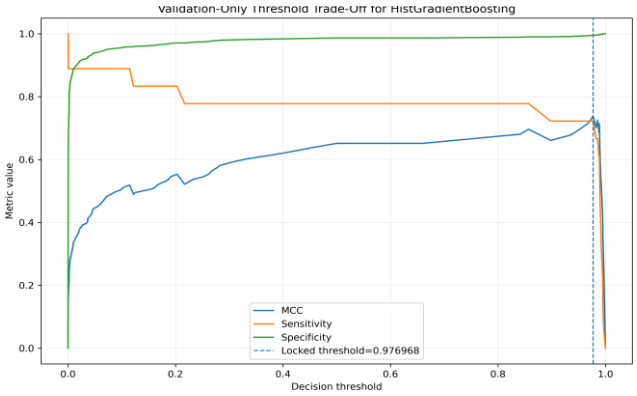

Fig. 11. Validation-only threshold trade-off for the selected model. The dashed line marks the threshold locked before final TEST access.

The validation threshold trade-off showed that the chosen operating point was intentionally conservative and high-specificity. This threshold is an evaluation operating point, not a calibrated clinical probability cut-point. The validation-only hierarchy is shown in Fig. 9, while the final TEST confusion matrix and validation-only threshold trade-off are shown separately in Figs. 10 and 11.

C. Explainability, feature structure and XAI quality control

Global SHAP rankings were highly stable: train-validation Spearman rho was 0.998616 and top-10 Jaccard 1.0. The leading original predictors were SnpEff_annotation_count, synonymous indicator, Worst_Impact, selected impact, PolyPhen, SIFT, SIFT missingness, splice indicator, transition/transversion flags, dbSNP presence, and other annotation/context variables. Correlation and standardized class-distribution checks were used to contextualize—not causally validate—these learned associations. SHAP additivity error was effectively numerical zero (TRAIN 1.865e-14; VALIDATION 2.132e-14), and frozen prediction reproduction error was 0. Local SHAP-LIME agreement varied by case type rather than being uniformly high: mean rank Spearman ranged from 0.228 for false positives to 0.730 for false negatives, and mean top-5 Jaccard ranged from 0.143 to 0.698. The variation across case types led us to use SHAP and LIME as complementary descriptions of individual model decisions, rather than regarding either method as a definitive explanation. XAI read no TEST rows or labels and performed no TEST inference, SHAP, or LIME.

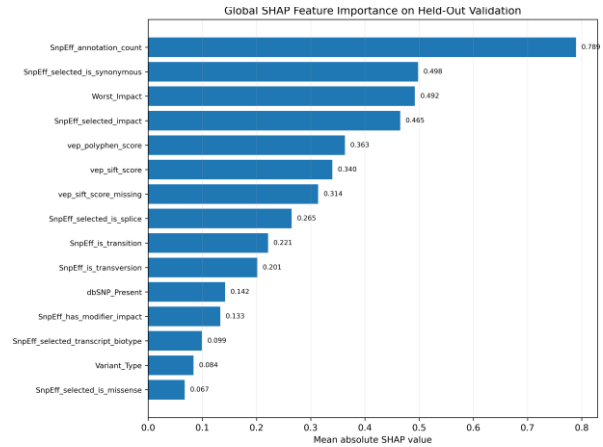

Fig. 12. Global SHAP importance for the top 15 original predictors on held-out VALIDATION.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

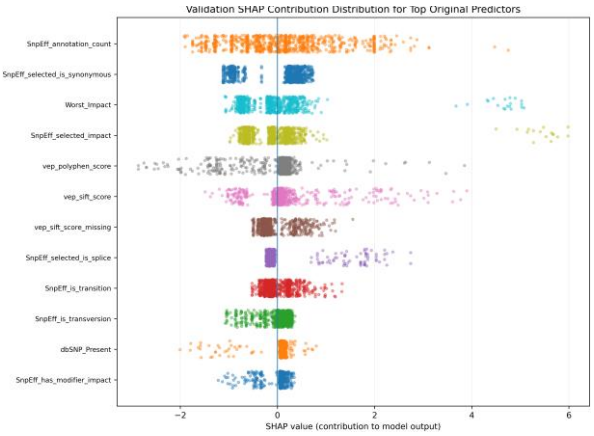

Fig. 13. Distribution of VALIDATION SHAP contributions for the top original predictors. Feature-wise point colors are categorical display aids only.

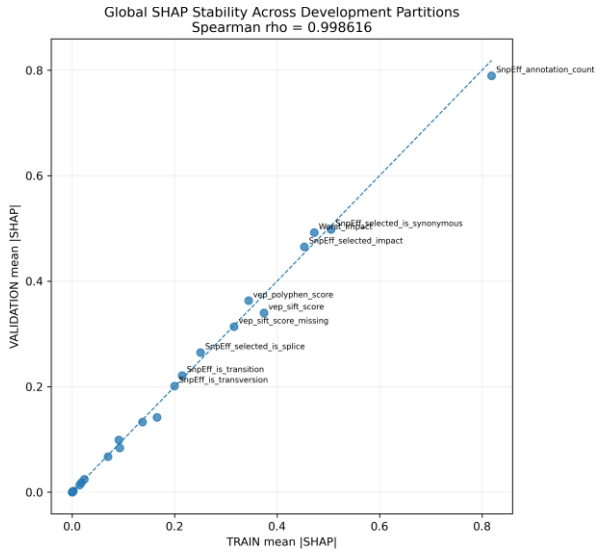

Fig. 14. Global SHAP stability between TRAIN and VALIDATION. The near-diagonal pattern corresponds to Spearman rho = 0.998616.

TABLE X. GLOBAL SHAP RANKING STABILITY ACROSS TRAIN AND VALIDATION

Val/Train rank	Original predictor	TRAIN |SHAP|	VALID. |SHAP|
1/1	SnpEff_annotation_count	0.8185	0.7894
2/2	SnpEff_selected_is_synonymous	0.5054	0.4981
3/3	Worst_Impact	0.4725	0.4921
4/4	SnpEff_selected_impact	0.4531	0.4651
5/6	vep_polyphen_score	0.3444	0.3631
6/5	vep_sift_score	0.3744	0.3399
7/7	vep_sift_score_missing	0.3160	0.3137
8/8	SnpEff_selected_is_splice	0.2506	0.2648
9/9	SnpEff_is_transition	0.2146	0.2212
10/10	SnpEff_is_transversion	0.1996	0.2012
11/11	dbSNP_Present	0.1656	0.1420
12/12	SnpEff_has_modifier_impact	0.1373	0.1331
13/14	SnpEff_selected_transcript_biotype	0.0913	0.0991
14/13	Variant_Type	0.0930	0.0837
15/15	SnpEff_selected_is_missense	0.0701	0.0675

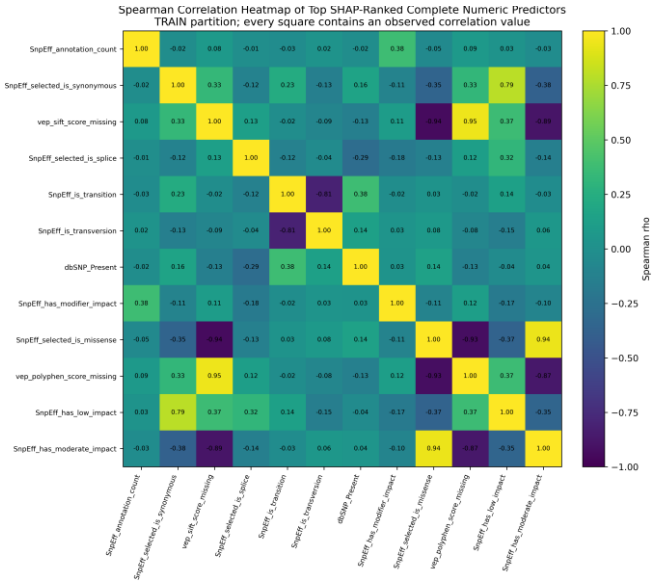

Fig. 15. Spearman correlation heatmap of top SHAP-ranked complete numeric predictors in TRAIN. Every square contains a numeric correlation value.

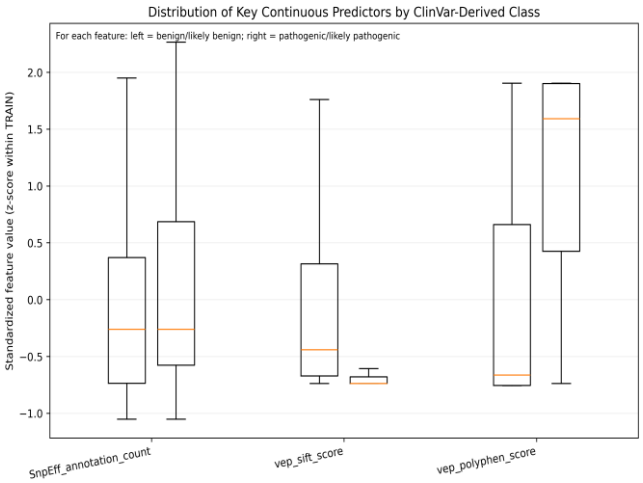

Fig. 16. Standardized distributions of key continuous predictors by ClinVar-derived class in TRAIN. Left/right boxes within each feature correspond to benign and pathogenic classes, respectively.

TABLE XI. EXPLAINABILITY REPRODUCIBILITY AND TEST-ISOLATION SAFEGUARDS

QC / safeguard	Observed value
Selected model	HistGradientBoosting (HGB)
Locked threshold	0.976968279
TRAIN / VALIDATION explained	3,406 / 839
Encoded / original predictors	71 / 26
Frozen probability max abs error	0.000e+00
TRAIN SHAP additivity max abs error	1.865e-14
VALIDATION SHAP additivity max abs error	2.132e-14
Train-validation SHAP Spearman	0.998616
Top-10 Jaccard	1.000000
TEST rows/labels read in XAI	0
TEST inference in XAI	0
TEST SHAP / LIME	0 / 0

TABLE XII. LOCAL SHAP-LIME AGREEMENT ACROSS SELECTED VALIDATION CASE TYPES

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

Selection reason	N	Mean rank Spearman	Mean top-5 Jaccard
False Negative	3	0.730	0.698
False Positive	3	0.228	0.369
Low Probability	3	0.634	0.143
Threshold Boundary	3	0.444	0.528
True Positive	3	0.450	0.508

D. Locked TEST uncertainty and probability calibration

Patient-group bootstrap made the small positive-class sample size explicit. TEST AP 0.6105 had 95% CI 0.3531-0.8114; ROC-AUC 0.9730 had 0.9449-0.9977; MCC 0.6627 had 0.3961-0.7488; sensitivity 0.6364 had 0.2857-0.8750; specificity 0.9960 had 0.9915-0.9990; precision 0.7000 had 0.3333-0.9000; F1 0.6667 had 0.3636-0.7500; and Brier 0.0140 had 0.0071-0.0218. AP, ROC-AUC, MCC, and sensitivity had 1,994 valid bootstrap replicates because six resamples contained only one outcome class; Brier score, specificity, precision, and F1 remained estimable in all 2,000 replicates, as reported in Table XIII. Point estimates remain the primary results; bootstrap medians were not substituted.

Calibration separated discrimination from absolute probability reliability. Brier was 0.013997 and ECE10 0.019591, but calibration-in-the-large was -3.205709 (95% CI -5.408469 to -1.898353) and the joint logistic intercept -2.427804 (-4.109578 to -1.704487), indicating systematic probability overprediction. The slope was 0.616364 (0.535018-1.130981): its point estimate suggests overly extreme predictions, but the interval includes 1. No TEST-based recalibration was performed.

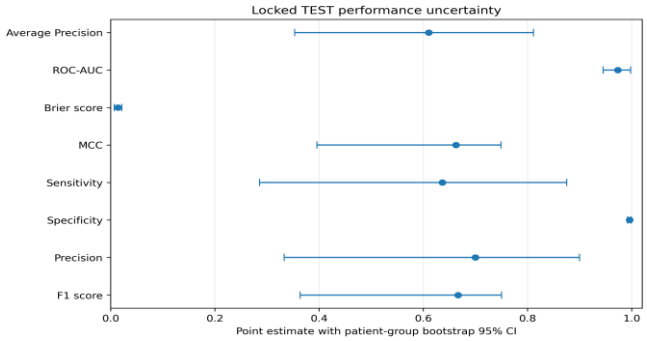

Fig. 17. Patient-group bootstrap 95% confidence intervals for the locked TEST point estimates. Positive-class metrics have visibly wider uncertainty than ROC-AUC and specificity.

TABLE XIII. LOCKED TEST PERFORMANCE WITH PATIENT-GROUP BOOTSTRAP UNCERTAINTY

Metric	Point	95% CI	Valid/requested
Average Precision	0.6105	0.3531-0.8114	1,994 / 2,000
ROC-AUC	0.9730	0.9449-0.9977	1,994 / 2,000
Brier score	0.0140	0.0071-0.0218	2,000 / 2,000
MCC	0.6627	0.3961-0.7488	1,994 / 2,000
Sensitivity	0.6364	0.2857-0.8750	1,994 / 2,000
Specificity	0.9960	0.9915-0.9990	2,000 / 2,000
Precision	0.7000	0.3333-0.9000	2,000 / 2,000
F1 score	0.6667	0.3636-0.7500	2,000 / 2,000

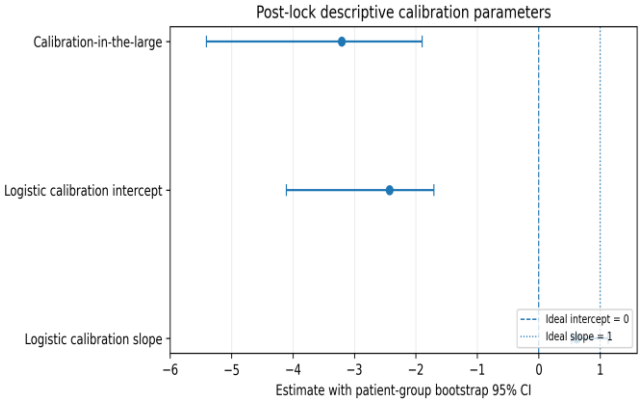

Fig. 18. Post-lock descriptive calibration parameters with patient-group bootstrap 95% confidence intervals. Ideal intercepts are 0 and the ideal slope is 1.

TABLE XIV. POST-LOCK DESCRIPTIVE CALIBRATION - ESTIMATES

Measure	Estimate	Ideal
Brier score	0.013997	0
ECE10	0.019591	0
Calibration-in-the-large	-3.205709	0
Logistic calibration intercept	-2.427804	0
Logistic calibration slope	0.616364	1

TABLE XV. POST-LOCK DESCRIPTIVE CALIBRATION - BOOTSTRAP UNCERTAINTY

Measure	Bootstrap 95% CI	Valid/requested
Brier score	0.007055-0.021815	2,000 / 2,000
ECE10	Not separately reported	-
Calibration-in-the-large	-5.408469 to -1.898353	1,687 / 2,000
Logistic calibration intercept	-4.109578 to -1.704487	1,687 / 2,000
Logistic calibration slope	0.535018-1.130981	1,687 / 2,000

E. Patient-dependence sensitivity and post-lock robustness

In paired OOF evaluation on the same 4,245 development rows, group-aware AP was 0.6048+/-0.0185 and row-wise AP 0.6104+/-0.0199; mean paired delta was only +0.0057 with empirical 2.5th-97.5th percentile -0.0470 to +0.0392. Row-wise CV exposed 100% of evaluation rows to a patient represented in fold-training, while group-aware CV maintained 0% patient overlap; exact-allele overlap remained 0 in both strategies. ROC-AUC, Brier, and MCC differences were similarly small and variable. Thus patient leakage violated the intended independence assumption without producing a predictable optimistic AP shift. The independent audit passed all 474 checks and did not read TEST or perform audit-time model loading/inference/refit. The five post-lock robustness analyses further constrained claims. FULL_26 grouped OOF AP was 0.6153+/-0.0215; removing SIFT/PolyPhen and their missingness indicators gave 0.5779+/-0.0197, whereas LOW_SEMANTIC_10 and BASIC_SEQUENCE_5 collapsed to 0.0570+/-0.0017 and 0.0544+/-0.0017. Hence performance was not solely driven by SIFT/PolyPhen but remained strongly annotation-semantic. Repeated classical ranking placed HistGradientBoosting at 0.5987+/-0.0159 and LightGBM at 0.5975+/-0.0127; each won 5/10 repeats and HGB was top-three 10/10, supporting a stable top-tier—not uniquely superior—selection. Repeat-specific MCC-optimal thresholds averaged 0.9614+/-0.0241 (median 0.9691; empirical 0.9005-0.9802); the frozen 0.976968 lay within this range, 12/20 optima were within +/-0.01, 18/20 within +/-0.05, and mean MCC regret was 0.0171.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

Strict simultaneous patient+gene disjoint development splitting was structurally infeasible because the patient-gene graph was one connected component. Under the nested patient-disjoint alternative, AP was 0.6153 overall, 0.7185 for seen-gene rows, and 0.4891 for unseen-gene rows (unseen-all -0.1262; unseen-seen -0.2295), demonstrating transfer with material attenuation rather than gene-independent performance. Across seeds 42-51, Transformer Encoder retained the highest mean AP 0.6675+/-0.0341, followed by Gated Residual Network 0.6602+/-0.0545; they won 4/10 and 5/10 seeds, respectively. At threshold 0.5, Transformer averaged 89.56% accuracy, 94.44% sensitivity and MCC 0.3709, whereas FT-Transformer averaged 93.49% accuracy and MCC 0.4511, reinforcing why AP—not accuracy—governed model selection.

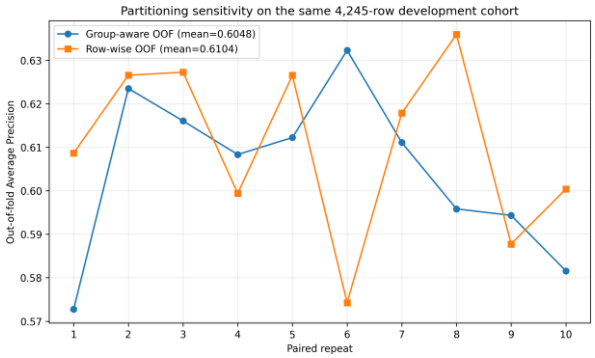

Fig. 19. Paired development-only OOF Average Precision across ten repeats on the same 4,245-row population. Row-wise cross-validation exposed every evaluation row to a patient already represented in fold-training, whereas group-aware cross-validation preserved patient independence.

TABLE XVI. PAIRED OOF SENSITIVITY - METHOD-LEVEL METRICS

Metric	Group-aware OOF	Row-wise OOF
Average Precision	0.6048 +/- 0.0185	0.6104 +/- 0.0199
ROC-AUC	0.9562 +/- 0.0072	0.9498 +/- 0.0097
Brier score	0.02007 +/- 0.00051	0.01979 +/- 0.00094
MCC	0.5909 +/- 0.0139	0.5967 +/- 0.0225
Eval rows with patient in fold-training	0.000	1.000
Eval rows with exact allele in fold-training	0.000	0.000

TABLE XVII. PAIRED OOF SENSITIVITY - PAIRED DIFFERENCES

Metric	Mean delta	Empirical 2.5-97.5%
Average Precision	+0.0057	-0.0470 to +0.0392
ROC-AUC	-0.0064	-0.0306 to +0.0124
Brier score	-0.00027	-0.00211 to +0.00145
MCC	+0.0058	-0.0248 to +0.0541
Eval rows with patient in fold-training	+1.000	1.000 in every row-wise repeat
Eval rows with exact allele in fold-training	0.000	0.000 in both strategies

TABLE XVIII. POST-LOCK DEVELOPMENT-ONLY ROBUSTNESS - ANALYSIS DESIGN

Analysis	Design
Semantic-dependence ablation	20 x 5 grouped OOF; fixed HGB hyperparameters
Classical model-selection stability	10 x 5 grouped CV; 14 frozen candidates
Frozen-threshold stability	20 grouped OOF repeats; no model fit/relock
Unseen-gene generalization	Nested unseen-gene subset within patient-disjoint CV
Expanded DL seed stability	Frozen top five architectures; seeds 42-51

TABLE XIX. POST-LOCK DEVELOPMENT-ONLY ROBUSTNESS - FINDINGS AND INTERPRETATION

Analysis	Key result	Interpretation
Semantic-dependence ablation	AP: FULL_26 0.6153+/-0.0215; NO_IN_SILICO_22 0.5779+/-0.0197; LOW_SEMANTIC_10 0.0570+/-0.0017; BASIC_SEQUENCE_5 0.0544+/-0.0017	Not solely SIFT/PolyPhen-driven, but strongly dependent on broader annotation semantics.
Classical model-selection stability	HGB 0.5987+/-0.0159 and LightGBM 0.5975+/-0.0127; each won 5/10; HGB top-3 10/10	HGB is a stable top-tier choice, not uniquely superior.
Frozen-threshold stability	Optimal threshold 0.9614+/-0.0241; q2.5-q97.5 0.9005-0.9802; frozen=0.976968; mean MCC regret 0.0171	Frozen high-specificity threshold is reasonably portable across development resamples.
Unseen-gene generalization	AP: ALL 0.6153; SEEN 0.7185; UNSEEN 0.4891; unseen-all -0.1262	Transfer persists to unseen-gene variants but is materially attenuated.
Expanded DL seed stability	Transformer 0.6675+/-0.0341; Gated 0.6602+/-0.0545; wins 4/10 vs 5/10	Highest mean AP remains Transformer, but the top-two margin is seed-dependent.

V. DISCUSSION

A. *Evaluation governance is the central contribution*

HistGradientBoosting was selected in this cohort, but the more important result is how that decision was reached. Patient- and allele-level separation defined the intended generalization setting, AP governed ranking under severe imbalance, and repeated seeds reduced dependence on a favorable deep-learning initialization. The final model and operating threshold were then decided on VALIDATION before TEST was opened. Later analyses could describe uncertainty, calibration, robustness, or explainability, but could not reverse those choices. The 27-model benchmark is therefore a test case for the evaluation procedure, not a claim that HistGradientBoosting should dominate on other datasets.

B. *Prevalence-sensitive and stability-aware selection reduced artifacts*

The choice of metric and the decision to examine repeated seeds both affected which models appeared strongest. Gated Residual Network led the deep models in one primary-seed run, but that result did not persist as a clear advantage across seeds. Over the original three seeds, Transformer Encoder had the slightly higher mean AP and substantially lower variability. Extending the comparison to ten seeds left Transformer marginally ahead in mean AP, although Gated Residual Network won more individual runs. A similar pattern appeared among the classical models, where HistGradientBoosting and LightGBM remained very close under repeated grouped evaluation. These results made a single favourable run an inadequate basis for selection; the ranking needed to be interpreted through the prespecified rare-class metric together with repeated evidence.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

C. *Model choice, threshold locking and final estimation were kept separate*

We deliberately handled model choice, threshold selection, and final performance estimation as separate stages. None of the 27 candidates was ranked using TEST. HistGradientBoosting was first selected from development evidence, after which its operating threshold was determined on VALIDATION and fixed. Repeated development resampling later showed that this frozen threshold incurred only a small average MCC regret. When the selected procedure was finally applied to TEST, discrimination remained strong, but uncertainty around positive-class measures was wide because TEST contained only 11 positives. The calibration analysis added a different perspective. Brier score and ECE10 were small, whereas the negative calibration intercepts showed systematic overprediction of absolute probabilities and the estimated slope was below 1. We did not use those TEST results to recalibrate the model. Accordingly, the locked threshold represents the operating point evaluated in this study and should not be interpreted as a clinically calibrated probability cut-off.

D. *Patient grouping determines what is being generalized*

The paired OOF experiment also clarified why patient-level grouping matters even when leakage does not produce an obvious increase in performance. With row-wise cross-validation, every evaluation row belonged to a patient who was already represented in the corresponding training fold. Group-aware cross-validation removed that overlap. Despite this clear difference in dependence, the average AP values were close and the direction of the paired difference changed across repeats. The two procedures were therefore estimating different problems rather than simply giving optimistic and pessimistic versions of the same estimate. Row-wise CV partly measures prediction of additional variants from patients already encountered during training; group-aware CV addresses prediction for previously unseen patients. Exact-allele overlap was zero under both strategies, allowing the comparison to focus more directly on patient dependence. Our reason for using grouped evaluation is therefore the intended generalization target, not an assumption that every form of leakage must increase a performance metric.

E. *Annotation dependence limits the scope of biological interpretation*

The ablation experiments helped determine which claims the model can reasonably support. Removing SIFT and PolyPhen together with their missingness indicators lowered AP, but the reduction was moderate, showing that these two score families alone did not account for the observed performance. The picture changed when much of the broader annotation context was removed. AP fell to 0.0570 for LOW_SEMANTIC_10 and 0.0544 for BASIC_SEQUENCE_5, values close to the rare-class baseline. Thus the classifier depends substantially on information carried by annotation-rich predictors. Generalization across gene context was also incomplete. The unseen-gene subset reached AP 0.4891 compared with 0.6153 overall, indicating that useful transfer remained but with a clear loss in performance. The model is therefore best described as predicting concordance with curated ClinVar-derived labels from annotation-rich evidence, not as gene-independent or knowledge-base-independent pathogenicity inference.

F. *Explainability is stable but non-causal*

Explainability was more stable globally than locally. TRAIN and VALIDATION SHAP rankings were almost identical and the additivity/reproduction checks showed only numerical-level error, but local SHAP-LIME agreement varied by case type. We therefore use these explanations to describe model behavior, not molecular mechanism, treatment benefit, or clinical actionability. Keeping TEST outside XAI also prevents post-hoc interpretation from becoming another source of feedback on the final performance estimate.

G. *Relationship to prior work and limitations*

This study is distinct from our earlier pipeline, review, multi-omics, and architecture publications [1]-[4]. Here the question is narrower: can rare-class genomic prediction be evaluated under a traceable contract in which the unit of generalization, ranking metric, model-selection path, operating threshold, and TEST boundary are explicit? These additional analyses were used to test the boundaries of the evaluation framework, not to create new claims of algorithmic superiority. The study also has several constraints that affect how the results should be read. Although 50 patient groups were available overall, the locked TEST contained only seven groups and 11 positive observations, which is reflected in the wide confidence intervals for positive-class metrics. A second limitation comes from the target itself. The labels were derived from ClinVar, while several predictors contain annotation-rich information related to existing variant knowledge; excluding clinical significance and exact-allele reuse reduces this dependence but does not remove it. We also could not construct a strict development split that was simultaneously patient-disjoint and gene-disjoint because the patient-gene graph formed a single connected component. Consequently, the unseen-gene analysis is a nested patient-disjoint assessment rather than a fully gene-disjoint external validation. The paired OOF experiment uses 43 development groups and 10 repeats, while the earlier row-level holdouts are composition-confounded and remain supplementary. No TEST-based recalibration, clinical trial, prospective deployment, or molecular-causal validation was performed.

VI. CONCLUSION

This study evaluated a leakage-controlled procedure using 4,999 naturally imbalanced breast-cancer variants from 50 patient groups. The predefined selection process retained HistGradientBoosting, which achieved an AP of 0.6105 and ROC-AUC of 0.9730 on the locked TEST. Those figures do not tell the whole story. TEST included few positive observations, producing wide uncertainty, and the frozen probabilities showed systematic overprediction of absolute risk. The development-only analyses also exposed important limits to generalization. Row-wise OOF introduced complete patient dependence but did not consistently increase AP; removing broader annotation information caused a substantial loss of performance; and performance on unseen-gene observations was lower than the overall OOF result. Repeated classical-model evaluation, threshold-resampling analysis, and the expanded deep-model seed experiment nevertheless showed that the selected procedure was broadly stable without demonstrating that one algorithm was uniquely superior. The practical value of the framework lies in the separation of development decisions from final performance estimation: model selection and threshold locking occurred before TEST access, while subsequent robustness and explainability analyses were prevented from feeding back into

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

those choices. The resulting benchmark is reproducible and makes its principal sources of uncertainty and dependence visible. It should be interpreted as a controlled genomic-AI evaluation study, not as a calibrated clinical decision system.

ACKNOWLEDGMENT

The authors gratefully acknowledge the Computation Laboratory of the Department of Mechanical Engineering, Engineering College, Ajmer for providing the high-end computational facility used in this research, procured under the World Bank-supported TEQIP-III project. The authors also acknowledge the Department of Computer Science & Engineering, Bhagwant University, Ajmer, for academic support, and the maintainers of the public genomic databases and open-source computational tools used in this study.

REFERENCES

[1] D. Sisodia, V. K. Sharma, R. Joshi, H. Arya, and T. K. Bhatt, "Developing an automated computational genomics pipeline for breast cancer detection using NGS," Research Plateau Current Trends in Engineering and Technology, vol. 4, pp. 36–45, 2025, presented at the 3rd International Conference on Recent Trends in Materials Science & Devices (ICRTMD-2025).

[2] D. Sisodia and V. K. Sharma, "A comprehensive literature survey on machine learning-based mutation prediction in breast cancer using next-generation sequencing data," International Education and Research Journal (IERJ), vol. 12, no. 04, 2026. doi: 10.5281/zenodo.20021195.

[3] D. Sisodia and V. K. Sharma, "A Hybrid Explainable Multi-Omics Machine Learning Framework For Breast Cancer Mutation Prediction And Clinical Risk Stratification Using Tcga-Brca Data". International Education and Research Journal (IERJ), 12(05), 299–313. https://doi.org/10.5281/zenodo.20484976

[4] D. Sisodia and V. K. Sharma, "Computational formulation of explainable genomic AI for breast cancer risk prediction," in Programme and Abstracts, 6th IEEE-Sponsored International Conference on Emerging Trends in Networks and Computer Communications (ETNCC 2026), Windhoek, Namibia, Aug. 4–6, 2026, Proceedings publication pending, Online Abstract Available: https://etncc.nust.na/sites/default/files/2026-08/ETNCC-2026-Progrm-4-6-August-2026-V5_0.pdf , Paper ID 516, p. 95.

[5] M. J. Landrum et al., "ClinVar: public archive of interpretations of clinically relevant variants," Nucleic Acids Research, vol. 44, no. D1, pp. D862–D868, 2016, doi: 10.1093/nar/gkv1222.

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

[7] L. Breiman, "Random forests," Machine Learning, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[8] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001, doi: 10.1214/aos/1013203451.

[9] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[10] G. Ke et al., "LightGBM: a highly efficient gradient boosting decision tree," Advances in Neural Information Processing Systems, vol. 30, pp. 3146–3154, 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

[11] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," Advances in Neural Information Processing Systems, vol. 31, pp. 6638–6648, 2018. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2018/file/14491b756b3a51daac41c24863285549-Paper.pdf

[12] A. Vaswani et al., "Attention is all you need," Advances in Neural Information Processing Systems, vol. 30, pp. 5998–6008, 2017.

[13] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," Advances in Neural Information Processing Systems, vol. 34, pp. 18932–18943, 2021. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[14] S. O. Arik and T. Pfister, "TabNet: attentive interpretable tabular learning," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 8, pp. 6679–6687, 2021, doi: 10.1609/aaai.v35i8.16826.

[15] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," Advances in Neural Information Processing Systems, vol. 30, pp. 4765–4774, 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[16] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.

[17] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," PLoS ONE, vol. 10, no. 3, e0118432, 2015, doi: 10.1371/journal.pone.0118432.

[18] 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, 2024, doi: 10.1136/bmj-2023-078378.

[19] P. Cingolani et al., "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff," Fly, vol. 6, no. 2, pp. 80–92, 2012, doi: 10.4161/fly.19695.

[20] W. McLaren et al., "The Ensembl Variant Effect Predictor," Genome Biology, vol. 17, art. 122, 2016, doi: 10.1186/s13059-016-0974-4.