IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15604

QUANTIFYING EXPLANATION DRIFT UNDER MODEL COMPRESSION IN CLINICAL RISK PREDICTION

Stow May Tamara[1], Maudlyn Ireju Victor-Ikoh[2]

Department of Computer Science and Informatics, Federal University Otuoke, Bayelsa State, Nigeria[1,2]

Abstract: Predictive models on tabular clinical data increasingly use SHAP and LIME explanations, while compression is routine. This study quantifies how compression affects post hoc explanations on five clinical benchmarks. A multilayer perceptron was trained, then subjected to L1 pruning at four sparsities (30 to 90 percent) and quantization to four bit widths (8 to 2 bits), yielding nine variants. SHAP, LIME, and permutation importance were applied to each variant and compared to the full model. The transparency cost of compression is compression-type-dependent: heavy pruning generally degrades both accuracy and explanations together, so accuracy alone catches the problem; heavy quantization more often preserves accuracy while degrading explanations, so accuracy alone misses the problem. On two of five datasets at 2 bit quantization, AUC retention exceeds 0.92 while SHAP rank correlation against the full model falls below 0.70. Explanation fidelity should be reported alongside accuracy specifically when quantization is used.

Keywords: Explainable Artificial Intelligence, SHAP, LIME, Model Compression, Clinical Risk Prediction

1. INTRODUCTION

Artificial intelligence and machine learning are applied to a widening range of clinical decision making tasks, from diagnostic imaging [1] to risk stratification and treatment selection [2, 3], with tabular clinical models matching or surpassing traditional risk scores in tasks such as pneumonia risk stratification and hospital readmission prediction [4]. This adoption has motivated calls for greater transparency, since clinicians, patients, and regulators reasonably expect to understand the basis for decisions affecting health outcomes [5], and data protection frameworks have been interpreted as supporting a right to meaningful information about automated decisions [6]. The field of explainable artificial intelligence has developed two dominant post hoc methods for tabular data: Local Interpretable Model-agnostic Explanations (LIME) [7] and SHapley Additive exPlanations (SHAP) [8, 9], both generating per-instance attribution vectors indicating which features pushed a prediction toward or away from the positive class.

In parallel, deployment has motivated a second body of work on model compression. Modern neural networks are routinely subjected to pruning and quantization to reduce memory footprint, energy consumption, and inference latency [10, 11], and the Green artificial intelligence literature argues that efficiency should sit alongside predictive accuracy as a first class evaluation criterion [12, 13, 14]. Compression is typically evaluated by preservation of headline accuracy metrics; if these are preserved within an acceptable tolerance, the compressed model is treated as an adequate replacement for the full model.

This standard evaluation protocol embeds a tacit assumption: that two models which agree on predictions will also agree on explanations. The explanation produced for an individual prediction, however, depends on decision boundary geometry, weight magnitudes, and information flow through the layers, precisely the properties that pruning and quantization alter. There is no theoretical guarantee that a compressed model whose accuracy is preserved will also produce the same explanations as the full model. The present study treats this as an empirical question, characterising how magnitude based L1 pruning and uniform affine post training quantization affect SHAP, LIME, and permutation importance on five binary clinical benchmark datasets: Wisconsin Breast Cancer, Pima Indians Diabetes, Heart (Statlog), Breast Cancer Recurrence, and CSI PECARN. The specific objectives are to: (i) train a baseline multilayer perceptron on each dataset; (ii) apply compression operators at multiple intensities to produce variants whose accuracy approximates the baseline; (iii) generate explanations for each variant and quantify drift using established agreement metrics; (iv) compute an explainer noise floor bounding the drift attributable to explainer stochasticity rather than model change; and (v) report explanation drift alongside accuracy retention to make the transparency cost of compression visible.

The contribution is to make explicit and quantify the gap between accuracy preservation and explanation preservation in compressed models on clinical benchmarks, to provide a small set of agreement metrics for that purpose, and to argue that explanation fidelity should be reported alongside predictive performance whenever compressed models are evaluated for clinical deployment. The remainder of the paper is organised as follows. Section 2 reviews related work and identifies

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15604

the research gap. Section 3 sets out the methodology. Section 4 presents results across nine compressed variants on each dataset. Section 5 discusses the findings. Section 6 concludes with recommendations.

2. Related Works

2.1. Post Hoc Explanation Methods in Clinical Risk Prediction

The two most widely adopted post hoc explanation methods for tabular clinical data are LIME, which approximates a model locally using a sparse interpretable surrogate [7], and SHAP, which applies cooperative game theory to derive axiomatically grounded feature attributions [8] and extends in polynomial time to tree based models [9]. Alternative families include Integrated Gradients [15] and Grad-CAM [16] for differentiable models, and counterfactual explanations that identify minimal input changes altering a prediction [17]. A predictive, descriptive, relevant framework organises this field [18], with a recent comprehensive review covering both feature attribution and model specific transparency [19]. In the medical domain, method choice is often dictated by familiarity rather than formal comparison [20], and explanations must support causability, the property that physicians can verify against their own causal reasoning [21]. A systematic review of explainable artificial intelligence in clinical decision support documents a marked gap between the rapid uptake of these methods and the empirical evaluation of their fidelity [5].

2.2. Reliability and Limitations of Post Hoc Explanations

A growing body of work has questioned the reliability of post hoc explanations. The interpretation of neural networks can be fragile, in the sense that two perceptually indistinguishable inputs may yield very different attribution maps [22]. LIME and SHAP can both be adversarially fooled into producing innocuous explanations for biased classifiers [23]. Several saliency based explanation methods produce outputs that are largely invariant to model parameters and to training labels, raising fundamental concerns about whether such methods reflect model behaviour at all [24]. A recent impossibility result sharpens these concerns by proving that, for moderately rich model classes that include neural networks, complete and linear attribution methods such as SHAP and Integrated Gradients can fail to improve on random guessing for inferring local model behaviour, identifying spurious features, and supporting algorithmic recourse [25]. Broader conceptual concerns have been raised that interpretability is an underspecified construct [26] and that high stakes domains may be better served by inherently interpretable models than by post hoc explanations of opaque ones [27]. A survey of evaluation methods proposed for interpretability observes that consensus on what constitutes a faithful explanation has not yet been reached [28], and a systematic review of more than 300 papers identifies twelve quantifiable properties of explanations of which fidelity and stability are central to the present study [29]. A taxonomy of application grounded, human grounded, and functionally grounded evaluation frames the agreement based assessment used in the present work [30], and a recent meta survey documents the open challenge of quantifying explanation reliability under real world deployment conditions [31]. A study of industrial deployments documents a gap between the stated goal of transparency for end users and actual usage by engineers for debugging [32]. These contributions establish that fidelity, robustness, and reliability of explanations cannot be assumed, and must be tested empirically, particularly when the underlying model has been modified through processes such as compression. The disagreement between explanation methods applied to the same model is itself a recognised phenomenon that complicates the practical use of post hoc explanations in deployment [33, 34, 35], and has been systematically quantified for SHAP and LIME across tabular classifiers, with Kendall rank correlations reported as low as 0.006 on tree based ensembles and patterns of disagreement that vary systematically with model complexity [36].

2.3. Model Compression for Efficient Deployment

The deployment of deep neural networks under resource constraints has motivated an extensive literature on model compression. A comprehensive survey organises the field into four broad families, namely parameter pruning and sharing, low rank factorisation, transferred or compact convolutional filters, and knowledge distillation [10]. A more recent survey updates this landscape with focus on pruning and quantization, the two compression operators studied in the present work, and documents the trade-offs between sparsity, bit width, and accuracy across modern architectures [11]. Within pruning, magnitude based methods remain a standard baseline because of their simplicity and competitive performance, and they have been extended to channel level pruning for convolutional architectures [37]. On the quantization side, an integer arithmetic only quantization scheme is now widely deployed on mobile and embedded hardware [38]. The motivation for compression has been reinforced by the Green artificial intelligence literature, which has quantified the financial and environmental cost of training and serving large models [12, 13], and proposes that efficiency be reported alongside accuracy as a first class evaluation criterion. Recent work extends this position with detailed accounting of training emissions and best practices for emission reduction [14]. Most of these contributions evaluate compression outcomes purely in terms of accuracy preservation; little attention has been paid to how the explanations of compressed models compare to those of their uncompressed counterparts.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15604

2.4. Performance Evaluation in Imbalanced Clinical Classification

Beyond the area under the receiver operating characteristic curve, clinical data frequently exhibits class imbalance, motivating the use of the area under the precision recall curve [39] and the Matthews correlation coefficient [40] as supporting metrics, both of which are reported in the present study.

2.5. Research Gap

Across the four bodies of literature reviewed above, three findings stand out. First, post hoc explanation methods are now embedded in clinical machine learning workflows [5, 20], and their outputs are read by clinicians, regulators, and patients [21]. Second, the reliability of such explanations under perturbation is known to be limited [22, 23] and is now formally questioned [25]. Third, compression of neural networks for efficient deployment is now routine, and is evaluated almost exclusively in terms of accuracy retention [10, 13, 14]. What is conspicuously missing is a systematic empirical characterisation of how compression operators affect the post hoc explanations that clinical users actually see. The present study addresses this gap on five binary clinical benchmarks. The hypothesis to be tested is that accuracy preservation, the standard signal of compression success, is decoupled from explanation preservation, and that the magnitude of this decoupling is large enough to be relevant for clinical deployment. The empirical results presented in Section 4 partially support this hypothesis: the decoupling is consistently observed under aggressive quantization but, on most datasets, not under aggressive pruning, where accuracy and explanation fidelity tend to decline together.

3. METHODOLOGY

This section sets out the experimental design, including the datasets, model architecture, compression procedures, explanation methods, agreement metrics, and the explainer noise floor used to contextualise the drift measurements.

3.1. Datasets

Five binary tabular clinical classification benchmarks were used. The Wisconsin Diagnostic Breast Cancer dataset consists of 569 instances and 30 numerical features derived from digitised images of fine needle aspirates of breast masses, with a binary target distinguishing malignant from benign cases. The Pima Indians Diabetes dataset consists of 768 instances and 8 physiological and demographic features for female patients of Pima heritage at least 21 years of age, with a binary target indicating diabetes diagnosis within a defined follow-up period. The Heart (Statlog) dataset consists of 270 instances and 15 features derived from the UCI Heart Disease database family, with a binary target indicating cardiac disease presence; features in the public mirror used here are anonymised, which does not affect the explanation drift analysis. The Breast Cancer Recurrence dataset (UCI 1988) consists of 277 instances and 17 categorical features such as tumour size, node involvement, and irradiation history, with a binary recurrence target; this dataset is methodologically distinct from the Wisconsin dataset above, with a different population, feature set, and prediction task. The CSI PECARN dataset consists of 3313 paediatric trauma cases and 39 clinical decision rule features, with a binary target indicating cervical spine injury. Each dataset was partitioned with a fixed random seed of 42 into 70 percent training and 30 percent test subsets using stratified sampling. All features were standardised to zero mean and unit variance using statistics computed on the training subset only. Pima physiological zeros in glucose, blood pressure, skin thickness, insulin, and body mass index were treated as missing values and imputed with the column median, following standard practice for that dataset.

3.2. Model Architecture

A multilayer perceptron with two hidden layers of 64 and 32 units, with ReLU activation between layers and a two-dimensional logit output, was used as the baseline model on all five datasets. Class probabilities are obtained by applying the softmax function to the logits, which is equivalent to the sigmoid for binary classification when probabilities of the two classes sum to one. Training used the Adam optimiser with a learning rate of 1e-3, the cross-entropy loss with weight decay of 1e-4, and 300 epochs of full-batch gradient descent. The choice of a multilayer perceptron of modest depth and width was deliberate: the aim of the study is to characterise the behaviour of routine compression operators on a model that is representative of the small to medium scale models actually used on tabular clinical data, rather than to push state of the art predictive performance. The complete pipeline, from data preprocessing through model training, compression, explanation generation, and evaluation, is illustrated in Figure 1. For an input x of dimension d, the forward computation is given by:

$$h_1 = \mathrm{ReLU}(W_1 x + b_1) \quad (1)$$

$$h_2 = \mathrm{ReLU}(W_2 h_1 + b_2) \quad (2)$$

$$z = W_3 h_2 + b_3, \; p = \mathrm{softmax}(z) \quad (3)$$

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15604

where $W_1 \in \mathbb{R}^{64xd}$, $W_2 \in \mathbb{R}^{32x64}$, $W_3 \in \mathbb{R}^{2x32}$, and softmax produces a two-element probability vector. The training objective is cross-entropy with L2 weight regularisation:

$$\mathscr{L} = -(1/N) \Sigma_i \log p_{i,\gamma i} + \lambda \|\theta\|^2_2 \quad (4)$$

where $p_{i,\gamma i}$ is the predicted probability of the true class for instance i, θ denotes all trainable parameters, N is the number of training instances, and the weight decay coefficient λ equals 10^{-4}.

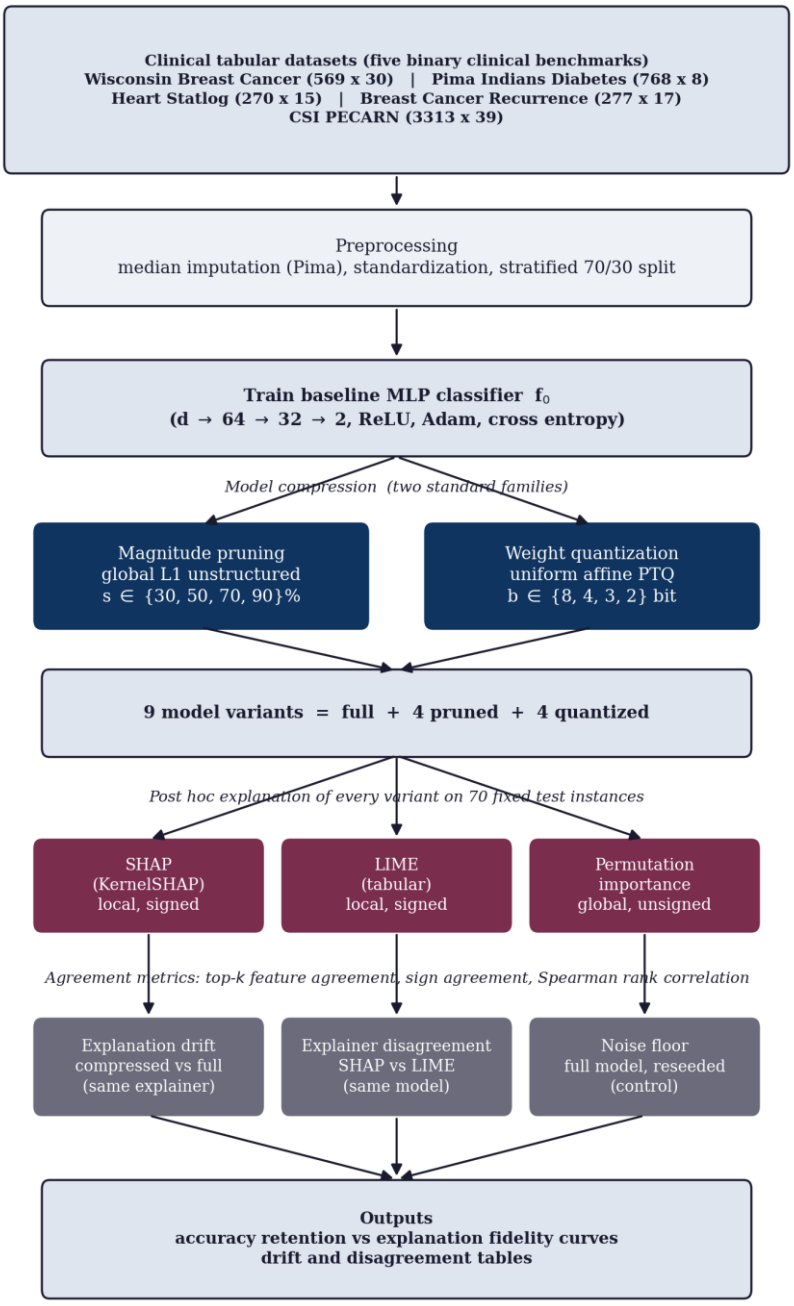


Figure 1. Methodology pipeline from data preprocessing through compression, post hoc explanation, and evaluation, showing the nine compressed variants produced per dataset and the agreement metrics used to quantify explanation drift.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15604

3.3. Compression Operators

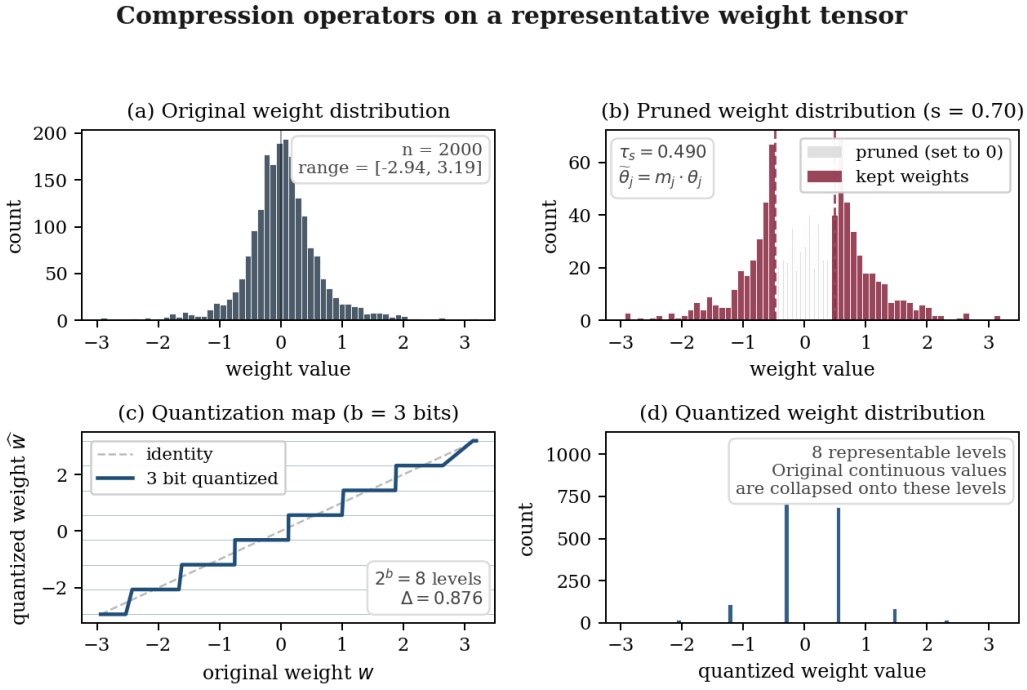
Two families of compression operators were applied independently. Magnitude based L1 unstructured pruning, applied globally across all linear layers, sets to zero those weights whose absolute value falls below a magnitude threshold, retaining only the largest magnitude weights. Pruning was applied at four intensities, namely 30, 50, 70, and 90 percent global sparsity. For a weight tensor W, the pruning mask M is defined by:

$$M_{ij} = 1 \text{ if } |W_{ij}| > \tau \text{ else } 0 \quad (5)$$

where τ is the magnitude threshold chosen to achieve the target sparsity. The pruned weight tensor is $W \odot M$, with $\odot$ the elementwise product. No post-pruning fine-tuning was performed; the compressed variants were evaluated directly so that the compression operator is the only source of variation between the full model and each compressed variant. Uniform affine post training quantization was applied at four bit widths, namely 8, 4, 3, and 2 bits, mapping floating point weights onto a discrete grid:

$$Q(w) = \text{round}((w - w\_min) / s) \cdot s + w\_min \quad (6)$$

where the scale factor $s = (w\_max - w\_min) / (2^B - 1)$ and B is the bit width, so that 2^B equally spaced levels are used between the per-tensor minimum and maximum weight values. Each compression intensity was applied independently to the trained full model, yielding nine model variants per dataset: the full model, four pruned variants (30, 50, 70, 90 percent), and four quantized variants (8, 4, 3, 2 bits). The intent is to span the practical range from minimal to aggressive compression, in line with what is reported in the deployment literature. The behaviour of each family on a representative weight tensor is illustrated in Figure 2.



Pruning enforces sparsity by zeroing low magnitude weights below an absolute threshold; quantization enforces precision reduction by collapsing all continuous values onto a small discrete grid. The two operators alter the weight tensor in qualitatively different ways.

Figure 2. The two complementary compression operators applied to a representative weight tensor. Panel (a) shows the original weight distribution; panel (b) shows the same distribution after L1 magnitude pruning at 70 percent sparsity, with pruned weights shown in grey and surviving weights in red; panel (c) shows the staircase mapping from continuous to quantized weights at 3 bit precision; panel (d) shows the resulting quantized weight distribution collapsed onto eight discrete levels. The remaining pruning sparsities (30, 50, 90 percent) and bit widths (8, 4, 2 bits) used in the experiments follow the same procedure at different intensities.

3.4. Explanation Methods

Three post hoc explanation methods were applied to each model variant. SHAP, as implemented by the shap library version 0.51, was used in KernelSHAP mode with a k-means background of 25 samples drawn from the training set and

DOI: 10.17148/IJARCCE.2026.15604

the library default l1 regularisation setting num_features(10), which selects up to ten features with non-zero attribution per instance via debiased lasso. The Shapley value for feature i on instance x is given by:

$$\varphi_i(x) = \Sigma_s \left[|S|!(d-|S|-1)!/d! \right] \cdot \left[f(x_{S \cup \{i\}}) - f(x_S) \right] \quad (7)$$

where S ranges over subsets of features not containing i, d is the total feature dimensionality, and f(x_S) is the model expectation conditional on the feature subset S. LIME, as implemented in the lime library, was used in tabular mode with the default sampling configuration, the exponential kernel, and automatic discretisation of continuous features. LIME solves the following local objective for an instance x with neighbourhood weights π_x:

$$\xi(x) = \text{argmin}_{m \in G} \; \mathcal{L}(f, m, \pi_x) + \Omega(m) \quad (8)$$

where G is the class of linear models, $\mathcal{L}$ measures fidelity to the model f under sample weighting π_x, and Ω penalises model complexity. Permutation importance was used as a complementary global feature attribution method, defined here as the average drop in AUC after random shuffling of one feature column at a time, over ten repeated shuffles:

$$PI_i = AUC(f, X, y) - (1/R) \Sigma_r AUC(f, X^{(\text{perm } i, r)}, y) \quad (9)$$

with R = 10 repeats per feature. For every model variant, explanations were generated on 70 instances sampled with a fixed seed from the test set on each dataset, yielding 630 explanation rows per dataset across the nine variants.

3.5. Agreement Metrics

Explanation drift and disagreement were measured using agreement metrics applied to pairs of attribution vectors, in each case comparing the attributions from a compressed variant against those from the full model on the same instance. Feature agreement at k measures the fraction of features appearing in the top k of both attribution vectors:

$$FA_k(\varphi^A, \varphi^B) = |\text{topk}(\varphi^A) \cap \text{topk}(\varphi^B)| / k \quad (10)$$

with k = 5. Sign agreement is computed as the count of features that appear in $\text{topk}(\varphi^A) \cap \text{topk}(\varphi^B)$ and have matching signs in both attributions, divided by k. Rank correlation is the Spearman correlation between the absolute attribution vectors over all d features:

$$\rho = 1 - (6 \Sigma d_i^2) / \left[d(d^2 - 1) \right] \quad (11)$$

where d_i is the difference in feature i's rank between the two absolute attribution vectors. For the drift analysis reported in this paper the two principal metrics are feature agreement at k = 5 and rank correlation; sign agreement is additionally reported when comparing SHAP and LIME attributions on the same model. All metrics were averaged over the 70 instances per dataset per variant.

3.6. Explainer Noise Floor

A potential confound in explanation drift measurements is the stochasticity intrinsic to the explainers themselves. KernelSHAP samples coalitions and LIME samples perturbations; neither produces deterministic outputs without seeded sampling. To bound this contribution, SHAP and LIME were each applied twice to the full model on each dataset with different random seeds, and the rank correlation between the two runs was computed. This rank correlation provides a noise floor for each explainer on each dataset: when the rank correlation between the compressed and the full model falls substantially below this noise floor, the difference cannot be attributed to explainer stochasticity alone and is taken to reflect genuine model change. Permutation importance with a fixed random seed is deterministic and therefore does not require a separate noise floor.

3.7. Reproducibility

All experiments used a single fixed seed of 42. The code is provided in an accompanying repository alongside the manuscript. Compression operations, training, and explanation generation were performed on a single CPU machine.

4. RESULTS

4.1. Predictive Performance Across Variants

Table 1 reports predictive performance metrics across all nine variants on each of the five datasets. Full-model AUC ranges from 0.690 on Breast Cancer Recurrence (the most challenging dataset) to 0.997 on Wisconsin Breast Cancer. AUC retention under compression varies considerably across datasets and compression operators: on Wisconsin Breast Cancer it remains at or above 0.947 across all nine variants, while on CSI PECARN it drops as low as 0.613 under 90 percent pruning. The Matthews correlation coefficient and area under the precision-recall curve follow similar trends, with mild compression generally preserving performance and aggressive compression producing dataset-dependent degradation. A notable observation is that aggressive quantization in several cases yields AUC retention values slightly

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15604

above 1.0 (Heart at 2 bit reaches 1.040; Pima at 2 bit reaches 1.022); this reflects the small magnitude of changes interacting with metric variance on small test sets rather than a genuine improvement in classification quality.

Table 1. Predictive performance metrics across nine model variants on five datasets.

Variant	Dataset	AUC	AUPRC	MCC	F1	AUC ret
full	BC	0.997	0.998	0.951	0.981	1.000
prune_30	BC	0.998	0.999	0.951	0.981	1.000
prune_50	BC	0.997	0.998	0.951	0.981	1.000
prune_70	BC	0.997	0.998	0.925	0.972	1.000
prune_90	BC	0.944	0.947	0.814	0.933	0.947
quant_8bit	BC	0.997	0.998	0.951	0.981	1.000
quant_4bit	BC	0.998	0.999	0.951	0.981	1.000
quant_3bit	BC	0.997	0.998	0.951	0.981	1.000
quant_2bit	BC	0.997	0.998	0.939	0.976	1.000
full	Pima	0.793	0.664	0.409	0.607	1.000
prune_30	Pima	0.789	0.649	0.388	0.600	0.995
prune_50	Pima	0.769	0.631	0.367	0.607	0.969
prune_70	Pima	0.650	0.492	0.227	0.551	0.820
prune_90	Pima	0.631	0.519	0.211	0.554	0.796
quant_8bit	Pima	0.794	0.665	0.409	0.607	1.001
quant_4bit	Pima	0.799	0.656	0.420	0.615	1.008
quant_3bit	Pima	0.793	0.668	0.347	0.548	1.000
quant_2bit	Pima	0.810	0.680	0.373	0.593	1.022
full	Heart	0.867	0.839	0.564	0.769	1.000
prune_30	Heart	0.861	0.840	0.571	0.775	0.992
prune_50	Heart	0.852	0.828	0.529	0.756	0.982
prune_70	Heart	0.796	0.765	0.514	0.744	0.917
prune_90	Heart	0.685	0.650	0.292	0.517	0.789
quant_8bit	Heart	0.867	0.839	0.564	0.769	1.000
quant_4bit	Heart	0.860	0.836	0.571	0.775	0.992
quant_3bit	Heart	0.873	0.844	0.635	0.805	1.006
quant_2bit	Heart	0.901	0.873	0.621	0.800	1.040
full	BCR	0.690	0.536	0.258	0.500	1.000
prune_30	BCR	0.696	0.544	0.294	0.526	1.009
prune_50	BCR	0.692	0.543	0.209	0.484	1.003
prune_70	BCR	0.681	0.471	0.173	0.459	0.987
prune_90	BCR	0.555	0.425	-0.126	0.391	0.804
quant_8bit	BCR	0.692	0.538	0.258	0.500	1.003
quant_4bit	BCR	0.705	0.552	0.259	0.508	1.023
quant_3bit	BCR	0.685	0.550	0.242	0.500	0.994
quant_2bit	BCR	0.640	0.492	0.145	0.383	0.928
full	CSI	0.772	0.414	0.308	0.409	1.000
prune_30	CSI	0.763	0.400	0.293	0.400	0.987
prune_50	CSI	0.720	0.342	0.209	0.356	0.932
prune_70	CSI	0.629	0.269	0.119	0.312	0.815
prune_90	CSI	0.474	0.149	0.019	0.281	0.613
quant_8bit	CSI	0.773	0.416	0.310	0.411	1.001
quant_4bit	CSI	0.773	0.425	0.331	0.424	1.001
quant_3bit	CSI	0.729	0.378	0.258	0.359	0.943
quant_2bit	CSI	0.679	0.341	0.252	0.323	0.878

4.2. Explanation Drift and Inter-Explainer Disagreement

Table 2 reports SHAP and LIME rank correlations between each compressed variant and the full model, and SHAP versus LIME rank correlation on the same model, for all nine variants on all five datasets. Values closer to 1.0 indicate

DOI: 10.17148/IJARCCE.2026.15604

stronger preservation of the explanation; values closer to 0 indicate severe drift. The contrast between accuracy preservation and explanation preservation is illustrated across all datasets in Figure 3.

Table 2. Rank correlation of SHAP and LIME between each compressed variant and the full model, and SHAP versus LIME rank correlation on the same model, across nine variants on five datasets.

Variant	Dataset	SHAP rank	LIME rank	PERM rank	SHAP–LIME
full	BC	1.000	1.000	1.000	0.564
prune_30	BC	0.893	0.993	0.854	0.569
prune_50	BC	0.884	0.978	0.945	0.574
prune_70	BC	0.859	0.947	0.793	0.598
prune_90	BC	0.573	0.558	0.444	0.618
quant_8bit	BC	0.909	1.000	0.998	0.571
quant_4bit	BC	0.896	0.992	0.876	0.573
quant_3bit	BC	0.891	0.977	0.926	0.578
quant_2bit	BC	0.841	0.927	0.736	0.573
full	Pima	1.000	1.000	1.000	0.447
prune_30	Pima	0.930	0.944	1.000	0.445
prune_50	Pima	0.856	0.896	0.929	0.436
prune_70	Pima	0.598	0.609	0.905	0.387
prune_90	Pima	0.437	0.318	0.357	0.460
quant_8bit	Pima	0.997	0.996	1.000	0.457
quant_4bit	Pima	0.936	0.917	0.976	0.483
quant_3bit	Pima	0.815	0.779	0.738	0.442
quant_2bit	Pima	0.685	0.669	0.714	0.566
full	Heart	1.000	1.000	1.000	0.173
prune_30	Heart	0.954	0.973	0.971	0.175
prune_50	Heart	0.897	0.818	0.804	0.239
prune_70	Heart	0.667	0.723	0.407	0.275
prune_90	Heart	0.431	0.283	0.071	0.084
quant_8bit	Heart	0.970	0.997	0.996	0.185
quant_4bit	Heart	0.961	0.985	0.879	0.145
quant_3bit	Heart	0.904	0.896	0.800	0.129
quant_2bit	Heart	0.782	0.654	0.921	-0.006
full	BCR	1.000	1.000	1.000	-0.157
prune_30	BCR	0.918	0.963	0.951	-0.129
prune_50	BCR	0.814	0.834	0.882	-0.104
prune_70	BCR	0.659	0.634	0.304	-0.204
prune_90	BCR	0.278	0.500	-0.152	-0.038
quant_8bit	BCR	0.954	0.999	0.971	-0.156
quant_4bit	BCR	0.873	0.945	0.775	-0.086
quant_3bit	BCR	0.800	0.906	0.880	-0.221
quant_2bit	BCR	0.508	0.798	0.135	-0.249
full	CSI	1.000	1.000	1.000	-0.198
prune_30	CSI	0.847	0.925	0.964	-0.213
prune_50	CSI	0.688	0.821	0.815	-0.244
prune_70	CSI	0.530	0.689	0.473	-0.252
prune_90	CSI	0.442	0.237	0.496	-0.021
quant_8bit	CSI	0.897	0.999	0.999	-0.189
quant_4bit	CSI	0.843	0.932	0.964	-0.168
quant_3bit	CSI	0.724	0.863	0.802	-0.219
quant_2bit	CSI	0.538	0.649	0.498	-0.174

Three observations stand out from Table 2. First, mild compression preserves explanations across all five datasets: at 30 percent pruning and 8 bit quantization, SHAP and LIME rank correlations against the full model are at least 0.84 on every dataset (with the minimum 0.847 on CSI PECARN under 30 percent pruning). Second, the response to aggressive

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15604

compression is compression-type-dependent. Under 90 percent pruning, both AUC retention and explanation rank correlations decline together on four of the five datasets (Pima, Heart, Breast Cancer Recurrence, CSI PECARN); only Wisconsin Breast Cancer exhibits the dissociation in which AUC retention remains high (0.947) while SHAP and LIME rank correlations fall sharply (to 0.573 and 0.558 respectively). Under 2 bit quantization, by contrast, AUC retention is essentially preserved or even slightly elevated on four of the five datasets (Wisconsin Breast Cancer, Pima, Heart, Breast Cancer Recurrence), while SHAP rank correlation falls to 0.841, 0.685, 0.782, and 0.508 respectively. The transparency cost of compression is therefore most consistently observed under aggressive quantization. Third, the SHAP versus LIME rank correlation on the same model is uniformly low on three of the five datasets, including negative values on Breast Cancer Recurrence and CSI PECARN, indicating substantial baseline disagreement between SHAP and LIME on these datasets independent of compression.

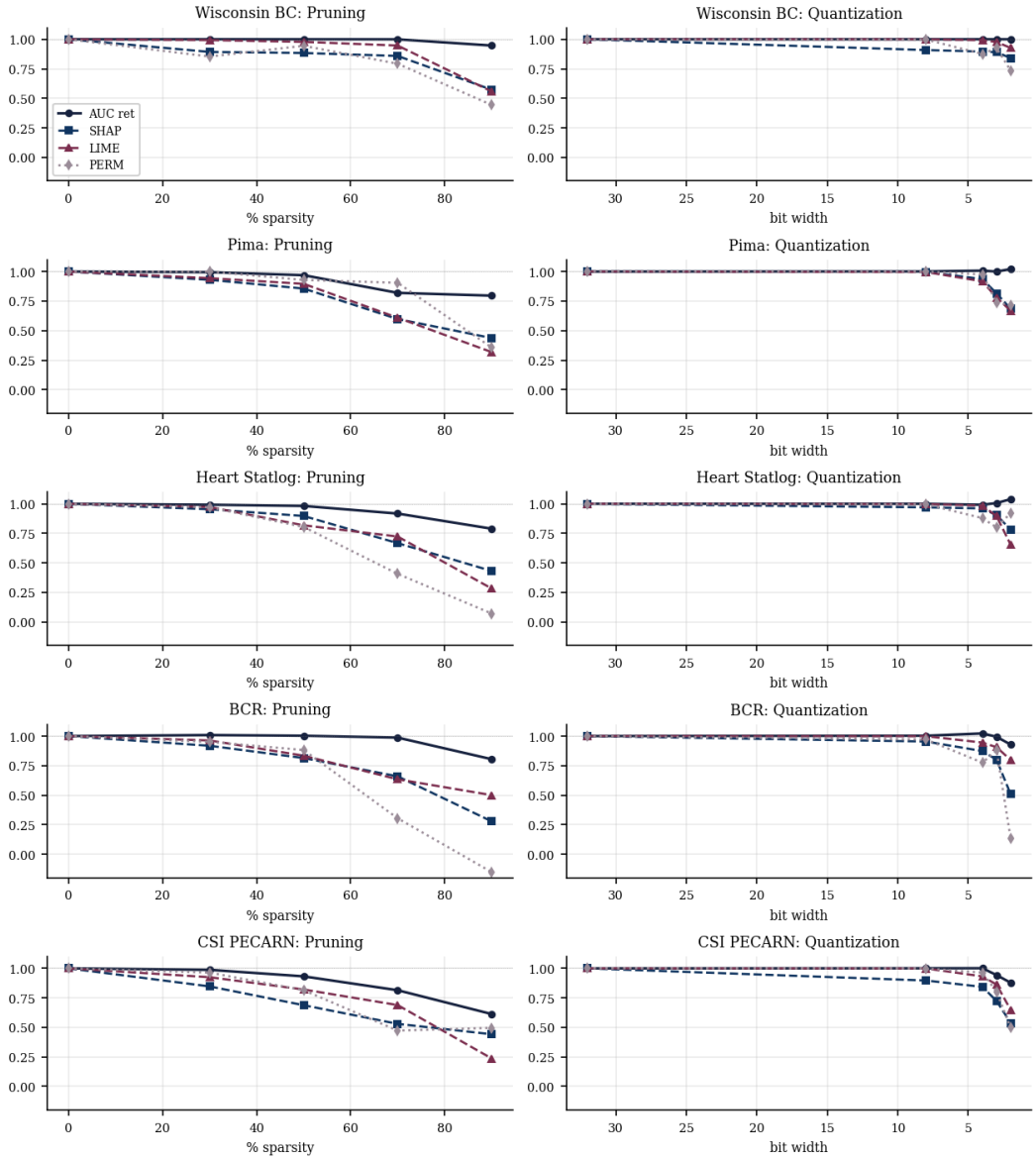


Figure 3. Predictive accuracy retention compared with SHAP, LIME, and permutation explanation fidelity across the pruning sweep (left column) and quantization sweep (right column) for each of the five datasets. Each marker is one model variant. The horizontal grey reference line at y=1.0 indicates parity with the full model.

4.3. Drift Relative to the Explainer Noise Floor

The explainer noise floors, computed as the rank correlation between two independent runs of each stochastic explainer on the full model under different random seeds, vary across datasets and are reported in Table 3. The SHAP noise floor ranges from 0.882 (CSI PECARN) to 1.000 (Pima Indians Diabetes and Heart Statlog); on Wisconsin Breast Cancer it is

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15604

0.891 and on Breast Cancer Recurrence 0.952. The LIME noise floor ranges from 0.864 (Wisconsin Breast Cancer) to 0.958 (Breast Cancer Recurrence). Permutation importance with a fixed random seed is deterministic and is therefore not assigned a stochastic noise floor in the present study. Comparison of the observed SHAP and LIME rank correlations to these noise floors is shown in Figure 4. Across all five datasets, the rank correlations on aggressively compressed variants fall well below the corresponding noise floor, confirming that the observed explanation drift is attributable to genuine model change rather than to explainer stochasticity. On mild compression variants, observed rank correlations on several datasets sit close to or just below the noise floor, reflecting near-parity drift that may not be distinguishable from explainer noise.

Table 3. Explainer noise floors on the full model for each dataset.

Dataset	SHAP noise floor	LIME noise floor
BC	0.891	0.864
Pima	1.000	0.872
Heart	1.000	0.944
BCR	0.952	0.958
CSI	0.882	0.951

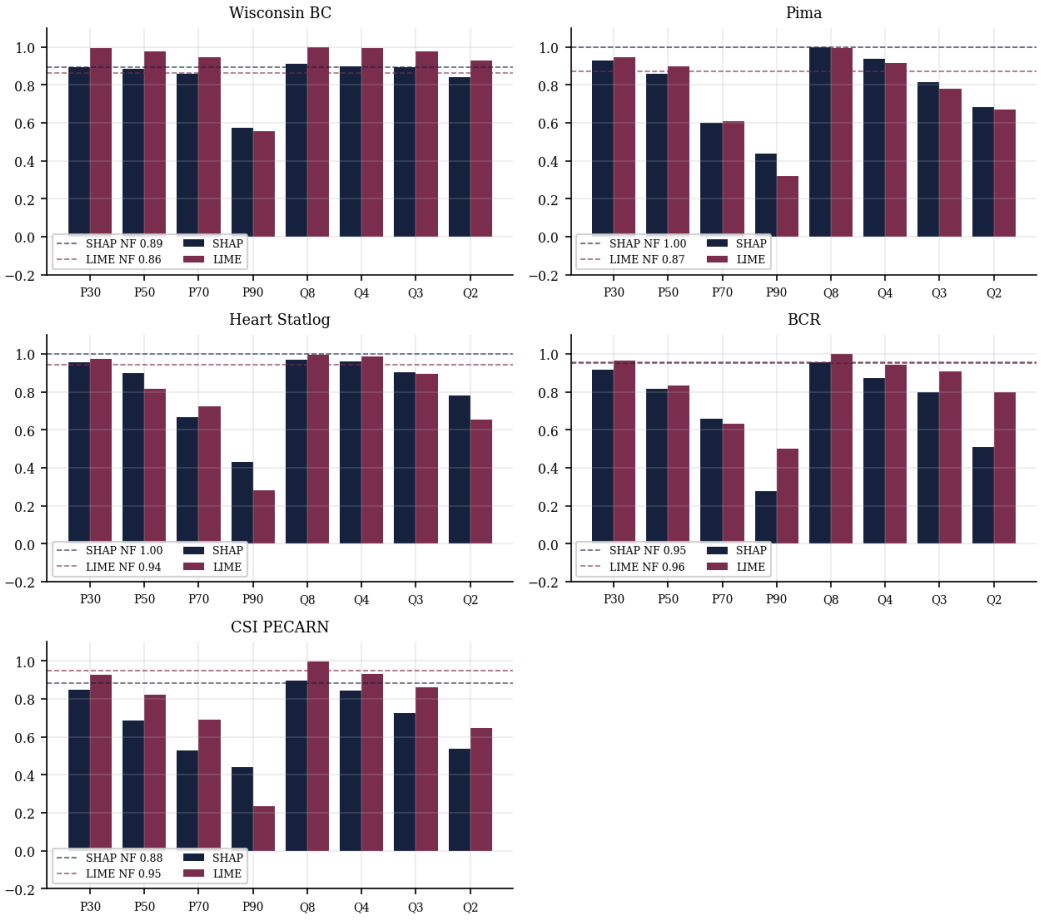


Figure 4. SHAP and LIME explanation rank correlation against the full model, for each of the eight compressed variants on each of the five datasets, compared to the explainer's own noise floor (dashed horizontal line) computed on the full model under different random seeds. Bars below the noise floor indicate genuine drift attributable to model change.

4.4. Per-Instance Attribution Behaviour

Beyond aggregate metrics, the per-instance behaviour of SHAP attributions provides further illustration of how compression alters individual explanations. Figure 5 shows SHAP attribution vectors for a single deterministically selected test instance on the Wisconsin Breast Cancer and Pima Indians Diabetes datasets across the nine model variants. On Wisconsin Breast Cancer, the leading features for the full model retain their sign across all compression variants, with

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15604

the most visible distortion occurring at 90 percent pruning, where several mid-ranked features (such as worst symmetry and area error) are weakened or effectively zeroed. On Pima Indians Diabetes, the dominant negative attribution of glucose is preserved across most variants, but at 90 percent pruning skin thickness changes sign and the magnitude of glucose itself drops noticeably, whereas under 2 bit quantization the attribution becomes more concentrated on glucose with the remaining features pushed toward zero.

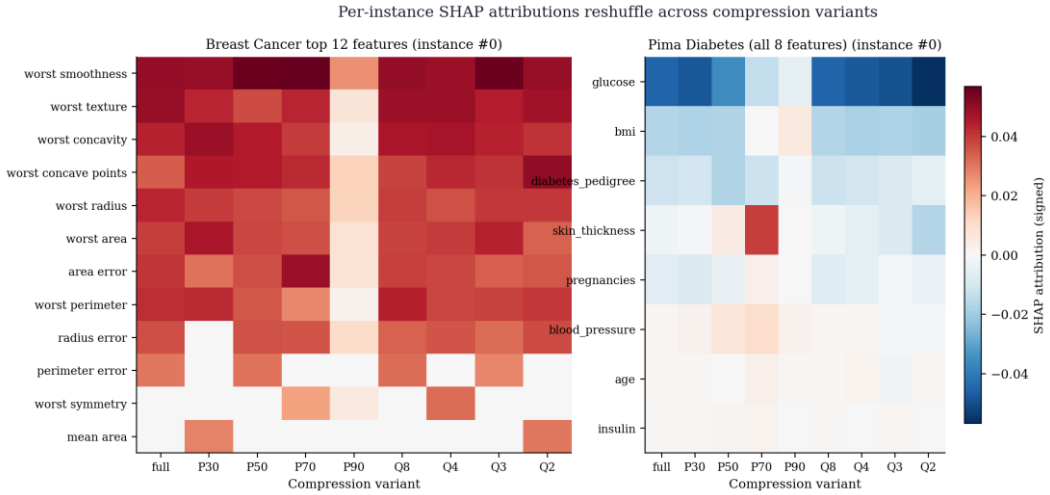


Figure 5. SHAP attribution vectors for a single deterministically selected test instance on Wisconsin Breast Cancer (left, top 12 features by mean absolute attribution) and Pima Indians Diabetes (right, all 8 features) across the nine model variants. Colour encodes signed attribution magnitude.

4.5. SHAP versus LIME Disagreement Under Compression

Table 2 also reports the SHAP versus LIME rank correlation on the same model for each variant. This quantity is conceptually distinct from explanation drift, since it measures how two different explanation methods describe the same model rather than how one method describes two models. The baseline level of SHAP versus LIME disagreement varies substantially across datasets. On Wisconsin Breast Cancer and Pima Indians Diabetes the rank correlation on the full model is moderate (0.564 and 0.447). On Heart (Statlog) it is low and positive (0.173). On Breast Cancer Recurrence and CSI PECARN it is slightly negative (-0.157 and -0.198), indicating that for these two datasets SHAP and LIME systematically disagree about which features matter on the same model. Across compressed variants the SHAP versus LIME rank correlation does not exhibit a strong monotonic relationship with compression intensity across the five datasets, although values fluctuate substantially; the one exception is a small monotonic increase from 0.564 to 0.618 on Wisconsin Breast Cancer as pruning intensity rises from 30 to 90 percent, which is not reproduced under quantization on the same dataset nor on the other four datasets. These observations are consistent with prior systematic quantification of SHAP versus LIME disagreement on tabular benchmarks, where agreement was reported to vary markedly with model family and dataset [36]; the present results extend that body of work by showing that this baseline is not, in general, further amplified by accuracy preserving compression of a fixed neural architecture.

5. Discussion

The empirical results support a refined version of the central hypothesis of the study. The transparency cost of compression, by which we mean the gap between accuracy retention and explanation fidelity retention, is real and measurable, but its visibility is compression-type-dependent rather than universal across compression operators and datasets.

5.1. Compression-Type-Dependent Transparency Cost

Aggressive pruning at 90 percent sparsity degrades both predictive accuracy and explanation fidelity on four of the five datasets studied (Pima Indians Diabetes, Heart Statlog, Breast Cancer Recurrence, CSI PECARN). On these datasets the standard practice of evaluating compression by accuracy metrics alone would already flag the model as inadequate for deployment, so explanation drift does not add a separate signal. Wisconsin Breast Cancer is the exception: at 90 percent pruning, AUC retention remains at 0.947 while SHAP and LIME rank correlations against the full model fall to 0.573 and 0.558. Aggressive quantization at 2 bit, in contrast, preserves AUC retention on four of the five datasets (values at or above 0.928), while SHAP rank correlation falls to 0.841, 0.685, 0.782, and 0.508 on those four datasets respectively. On these datasets the standard accuracy protocol does not flag a problem, but the explanations have nonetheless drifted

DOI: 10.17148/IJARCCE.2026.15604

materially. The transparency cost of compression is therefore most consistently observed under aggressive quantization, where it represents a real and potentially undetected risk; under aggressive pruning, on most datasets, accuracy and explanation signals are aligned and the standard protocol is adequate.

5.2. Interpretation in the Clinical Context

The clinical implications follow directly from the structure of clinical decision support workflows. A model that is selected for deployment on the basis of preserved area under the receiver operating characteristic curve or F1 score after quantization may nevertheless rely on a materially different feature basis when explaining individual predictions to clinicians. In a domain in which the explanation contributes to clinical reasoning, regulatory documentation, and patient communication [5, 21, 41], an unfaithful explanation is not a neutral artefact but an active source of risk. Recent systematic reviews of explainable artificial intelligence in healthcare [42, 43, 44] document a growing body of clinical applications in which the validity of post hoc explanations is treated as a deployment prerequisite, and one recent review emphasises that interpretability methods for medical image and tabular models must be evaluated quantitatively rather than accepted on visual appeal [45]. The present results sharpen the critical perspective that current post hoc methods are unlikely to deliver patient level transparency without additional safeguards [46] and that open challenges remain in deploying explainable artificial intelligence in clinical radiology [47]. They also support the call for caution in the use of post hoc explanations in high stakes domains [27]. The narrower practical recommendation that follows from the present empirical evidence is that explanation fidelity should be reported alongside accuracy specifically when quantization is used, since this is the regime in which accuracy alone is most likely to mask a substantive change in the model's explanatory basis.

5.3. Compression, Green Artificial Intelligence, and Explainer Differences

The motivations behind compression, including reduced computational and energy costs and deployment on resource-limited devices [10, 13, 14], remain valid. The present findings do not argue against compression but against the assumption that accuracy preservation is, in all regimes, a sufficient signal of model preservation; in the quantization case in particular, a more complete evaluation protocol would jointly report accuracy retention, explanation fidelity retention, and an explainer noise floor. Despite the impossibility results that bound the theoretical guarantees of SHAP and similar complete and linear methods [25], SHAP and LIME both exhibit drift that increases with compression intensity, with the largest drops at the most aggressive intensities. Their relative sensitivity depends on the operator: at 90 percent pruning the mean LIME rank correlation across the five datasets (0.379) is lower than the mean SHAP rank correlation (0.432), while at 2 bit quantization LIME (0.739) is higher than SHAP (0.671). Permutation importance, as a global rather than local method, captures changes in overall feature dependence but does not pinpoint per-instance attribution drift.

5.4. On the Disagreement Hypothesis

The hypothesis that compression would amplify the disagreement between SHAP and LIME on the same model was not supported by the data: across the five datasets, the SHAP versus LIME rank correlation shows no consistent monotonic trend with compression intensity. A small monotonic increase from 0.564 to 0.618 is observed on Wisconsin Breast Cancer under increasing pruning intensity, but the magnitude of this change is modest and the pattern is not reproduced on the other four datasets nor under quantization on Wisconsin Breast Cancer itself. The baseline level of SHAP versus LIME disagreement varies substantially across datasets, ranging from moderate positive rank correlation on Wisconsin Breast Cancer and Pima Indians Diabetes to mildly negative rank correlation on Breast Cancer Recurrence and CSI PECARN. Earlier work has separately quantified the substantial baseline disagreement between SHAP and LIME on uncompressed tabular models and shown that it varies systematically with model family [36]; the present observation is complementary to that finding. This secondary observation is reported with appropriate caution and is not advanced as a primary contribution.

5.5. Limitations

Several limitations of the study must be acknowledged transparently. The five datasets analysed are moderately sized canonical benchmarks and do not capture the scale or complexity of operational electronic health record data. The experiments use a single training seed, and no error bars are reported. The architecture studied is a single multilayer perceptron of modest size, leaving the behaviour of larger architectures and of other model families widely used on tabular clinical data, in particular random forests [48] and gradient boosted trees [49], uncharacterised. The evaluation is limited to agreement based measures of explanation fidelity; perturbation based faithfulness metrics may yield complementary insights. The Heart (Statlog) dataset features in the public mirror used here are anonymised, limiting clinical interpretability of per-instance explanations on that dataset, although the drift analysis is unaffected. These limitations narrow the scope of the conclusions but do not undermine the core observations.

DOI: 10.17148/IJARCCE.2026.15604

6. Conclusions and Recommendations

6.1. Summary of Findings

This study has quantified the effect of routine compression operators, namely magnitude based L1 pruning and uniform affine post training quantization, on the post hoc explanations produced by SHAP, LIME, and permutation importance on five binary clinical benchmark datasets. The principal empirical finding is that the transparency cost of compression is compression-type-dependent. Under aggressive pruning at 90 percent sparsity, both predictive accuracy and explanation fidelity decline together on four of the five datasets; only one dataset (Wisconsin Breast Cancer) exhibits the dissociation pattern in which accuracy is preserved while explanations drift sharply. Under aggressive quantization at 2 bit, in contrast, four of the five datasets retain AUC at or above 0.928 of the full model while SHAP rank correlation against the full model falls below 0.85 and, on two datasets, below 0.70. The transparency cost of compression is therefore most consistently observed under quantization, where accuracy alone is most likely to be a misleading signal of explanation preservation.

6.2. Recommendations

Three recommendations follow from these findings. First, explanation fidelity retention should be reported alongside accuracy retention whenever a compressed model is being evaluated for deployment in clinical or other high stakes settings, with particular attention to compressed variants produced by quantization [27, 50]. A small set of agreement metrics, such as those used in this study, provides a low cost addition to existing evaluation protocols. Second, an explainer noise floor should be computed and reported to bound the contribution of explainer stochasticity to apparent drift. Third, when explanation fidelity is found to degrade substantially under quantization, decision makers should weigh the gain in efficiency against the loss in explanatory reliability rather than assuming that accuracy parity is sufficient. Future work should extend the present analysis to larger architectures, to other compression operators including structured pruning and mixed precision quantization, to additional model families, and to faithfulness based explanation metrics that complement the agreement based metrics reported here.

Acknowledgements

The authors thank colleagues at the Department of Computer Science and Informatics, Federal University Otuoke, for helpful discussions and the Federal University Otuoke for institutional support.

References

[1] Esteva, A. et al. (2017) "Dermatologist-level classification of skin cancer with deep neural networks", Nature, Vol. 542, No. 7639, pp115-118. https://doi.org/10.1038/nature21056

[2] Topol, E.J. (2019) "High performance medicine: the convergence of human and artificial intelligence", Nature Medicine, Vol. 25, No. 1, pp44-56. https://doi.org/10.1038/s41591-018-0300-7

[3] Rajkomar, A., Dean, J. & Kohane, I. (2019) "Machine learning in medicine", New England Journal of Medicine, Vol. 380, No. 14, pp1347-1358. https://doi.org/10.1056/NEJMra1814259

[4] Caruana, R. et al. (2015) "Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission", Proc. 21st ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, pp1721-1730. https://doi.org/10.1145/2783258.2788613

[5] Antoniadi, A.M. et al. (2021) "Current challenges and future opportunities for XAI in machine learning based clinical decision support systems: a systematic review", Applied Sciences, Vol. 11, No. 11, p5088. https://doi.org/10.3390/app11115088

[6] Selbst, A.D. & Powles, J. (2017) "Meaningful information and the right to explanation", International Data Privacy Law, Vol. 7, No. 4, pp233-242. https://doi.org/10.1093/idpl/ipx022

[7] Ribeiro, M.T., Singh, S. & Guestrin, C. (2016) "Why should I trust you?: explaining the predictions of any classifier", Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, pp1135-1144. https://doi.org/10.1145/2939672.2939778

[8] Lundberg, S.M. & Lee, S.I. (2017) "A unified approach to interpreting model predictions", Advances in Neural Information Processing Systems 30, pp4765-4774. https://doi.org/10.5555/3295222.3295230

[9] Lundberg, S.M. et al. (2020) "From local explanations to global understanding with explainable AI for trees", Nature Machine Intelligence, Vol. 2, No. 1, pp56-67. https://doi.org/10.1038/s42256-019-0138-9

[10] Cheng, Y. et al. (2018) "Model compression and acceleration for deep neural networks: the principles, progress, and challenges", IEEE Signal Processing Magazine, Vol. 35, No. 1, pp126-136. https://doi.org/10.1109/MSP.2017.2765695

[11] Liang, T. et al. (2021) "Pruning and quantization for deep neural network acceleration: a survey", Neurocomputing, Vol. 461, pp370-403. https://doi.org/10.1016/j.neucom.2021.07.045

DOI: 10.17148/IJARCCE.2026.15604

[12] Strubell, E., Ganesh, A. & McCallum, A. (2019) "Energy and policy considerations for deep learning in NLP", Proc. 57th Annual Meeting of the Association for Computational Linguistics, pp3645-3650. https://doi.org/10.18653/v1/P19-1355

[13] Schwartz, R. et al. (2020) "Green AI", Communications of the ACM, Vol. 63, No. 12, pp54-63. https://doi.org/10.1145/3381831

[14] Patterson, D. et al. (2022) "The carbon footprint of machine learning training will plateau, then shrink", Computer, Vol. 55, No. 7, pp18-28. https://doi.org/10.1109/MC.2022.3148714

[15] Sundararajan, M., Taly, A. & Yan, Q. (2017) "Axiomatic attribution for deep networks", Proc. 34th Int. Conf. Machine Learning, pp3319-3328. https://doi.org/10.5555/3305890.3306024

[16] Selvaraju, R.R. et al. (2017) "Grad-CAM: visual explanations from deep networks via gradient-based localization", Proc. IEEE Int. Conf. Computer Vision, pp618-626. https://doi.org/10.1109/ICCV.2017.74

[17] Mothilal, R.K., Sharma, A. & Tan, C. (2020) "Explaining machine learning classifiers through diverse counterfactual explanations", Proc. 2020 Conf. Fairness, Accountability and Transparency, pp607-617. https://doi.org/10.1145/3351095.3372850

[18] Murdoch, W.J. et al. (2019) "Definitions, methods, and applications in interpretable machine learning", Proc. National Academy of Sciences, Vol. 116, No. 44, pp22071-22080. https://doi.org/10.1073/pnas.1900654116

[19] Hassija, V. et al. (2024) "Interpreting black-box models: a review on explainable artificial intelligence", Cognitive Computation, Vol. 16, No. 1, pp45-74. https://doi.org/10.1007/s12559-023-10179-8

[20] Tjoa, E. & Guan, C. (2021) "A survey on explainable artificial intelligence (XAI): toward medical XAI", IEEE Trans. Neural Networks and Learning Systems, Vol. 32, No. 11, pp4793-4813. https://doi.org/10.1109/TNNLS.2020.3027314

[21] Holzinger, A. et al. (2019) "Causability and explainability of artificial intelligence in medicine", WIREs Data Mining and Knowledge Discovery, Vol. 9, No. 4, e1312. https://doi.org/10.1002/widm.1312

[22] Ghorbani, A., Abid, A. & Zou, J. (2019) "Interpretation of neural networks is fragile", Proc. AAAI Conf. Artificial Intelligence, Vol. 33, No. 01, pp3681-3688. https://doi.org/10.1609/aaai.v33i01.33013681

[23] Slack, D. et al. (2020) "Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods", Proc. AAAI/ACM Conf. AI, Ethics, and Society, pp180-186. https://doi.org/10.1145/3375627.3375830

[24] Adebayo, J. et al. (2018) "Sanity checks for saliency maps", Advances in Neural Information Processing Systems 31, pp9505-9515. https://doi.org/10.5555/3327546.3327621

[25] Bilodeau, B. et al. (2024) "Impossibility theorems for feature attribution", Proc. National Academy of Sciences, Vol. 121, No. 2, e2304406120. https://doi.org/10.1073/pnas.2304406120

[26] Lipton, Z.C. (2018) "The mythos of model interpretability", Communications of the ACM, Vol. 61, No. 10, pp36-43. https://doi.org/10.1145/3233231

[27] Rudin, C. (2019) "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead", Nature Machine Intelligence, Vol. 1, No. 5, pp206-215. https://doi.org/10.1038/s42256-019-0048-x

[28] Carvalho, D.V., Pereira, E.M. & Cardoso, J.S. (2019) "Machine learning interpretability: a survey on methods and metrics", Electronics, Vol. 8, No. 8, p832. https://doi.org/10.3390/electronics8080832

[29] Nauta, M. et al. (2023) "From anecdotal evidence to quantitative evaluation methods: a systematic review on evaluating explainable AI", ACM Computing Surveys, Vol. 55, No. 13s, Article 295. https://doi.org/10.1145/3583558

[30] Doshi-Velez, F. & Kim, B. (2018) "Considerations for evaluation and generalization in interpretable machine learning", in Explainable and Interpretable Models in Computer Vision and Machine Learning, Springer, pp3-17. https://doi.org/10.1007/978-3-319-98131-4_1

[31] Saeed, W. & Omlin, C. (2023) "Explainable AI (XAI): a systematic meta-survey of current challenges and future opportunities", Knowledge-Based Systems, Vol. 263, p110273. https://doi.org/10.1016/j.knosys.2023.110273

[32] Bhatt, U. et al. (2020) "Explainable machine learning in deployment", Proc. 2020 Conf. Fairness, Accountability and Transparency, pp648-657. https://doi.org/10.1145/3351095.3375624

[33] Adadi, A. & Berrada, M. (2018) "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)", IEEE Access, Vol. 6, pp52138-52160. https://doi.org/10.1109/ACCESS.2018.2870052

[34] Barredo Arrieta, A. et al. (2020) "Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI", Information Fusion, Vol. 58, pp82-115. https://doi.org/10.1016/j.inffus.2019.12.012

[35] Marcinkevics, R. & Vogt, J.E. (2023) "Interpretable and explainable machine learning: a methods-centric overview with concrete examples", WIREs Data Mining and Knowledge Discovery, Vol. 13, No. 3, e1493. https://doi.org/10.1002/widm.1493

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15604

[36] Stow, M. (2026) "Quantifying explanation disagreement between SHAP and LIME across tabular classification models", Int. Journal of Computer Sciences and Engineering, Vol. 14, No. 1, pp13-21. https://doi.org/10.26438/ijcse.v14i1.7255

[37] He, Y., Zhang, X. & Sun, J. (2017) "Channel pruning for accelerating very deep neural networks", Proc. IEEE Int. Conf. Computer Vision, pp1398-1406. https://doi.org/10.1109/ICCV.2017.155

[38] Jacob, B. et al. (2018) "Quantization and training of neural networks for efficient integer-arithmetic-only inference", Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition, pp2704-2713. https://doi.org/10.1109/CVPR.2018.00286

[39] Saito, T. & Rehmsmeier, M. (2015) "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. https://doi.org/10.1371/journal.pone.0118432

[40] Chicco, D. & Jurman, G. (2020) "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation", BMC Genomics, Vol. 21, No. 1, p6. https://doi.org/10.1186/s12864-019-6413-7

[41] Stiglic, G. et al. (2020) "Interpretability of machine learning based prediction models in healthcare", WIREs Data Mining and Knowledge Discovery, Vol. 10, No. 5, e1379. https://doi.org/10.1002/widm.1379

[42] Loh, H.W. et al. (2022) "Application of explainable artificial intelligence for healthcare: a systematic review of the last decade (2011-2022)", Computer Methods and Programs in Biomedicine, Vol. 226, p107161. https://doi.org/10.1016/j.cmpb.2022.107161

[43] Albahri, A.S. et al. (2023) "A systematic review of trustworthy and explainable artificial intelligence in healthcare: assessment of quality, bias risk, and data fusion", Information Fusion, Vol. 96, pp156-191. https://doi.org/10.1016/j.inffus.2023.03.008

[44] Sadeghi, Z. et al. (2024) "A review of explainable artificial intelligence in healthcare", Computers and Electrical Engineering, Vol. 118, p109370. https://doi.org/10.1016/j.compeleceng.2024.109370

[45] Salahuddin, Z. et al. (2022) "Transparency of deep neural networks for medical image analysis: a review of interpretability methods", Computers in Biology and Medicine, Vol. 140, p105111. https://doi.org/10.1016/j.compbiomed.2021.105111

[46] Ghassemi, M., Oakden-Rayner, L. & Beam, A.L. (2021) "The false hope of current approaches to explainable artificial intelligence in health care", The Lancet Digital Health, Vol. 3, No. 11, pp e745-e750. https://doi.org/10.1016/S2589-7500(21)00208-9

[47] Reyes, M. et al. (2020) "On the interpretability of artificial intelligence in radiology: challenges and opportunities", Radiology: Artificial Intelligence, Vol. 2, No. 3, e190043. https://doi.org/10.1148/ryai.2020190043

[48] Breiman, L. (2001) "Random forests", Machine Learning, Vol. 45, No. 1, pp5-32. https://doi.org/10.1023/A:1010933404324

[49] Chen, T. & Guestrin, C. (2016) "XGBoost: a scalable tree boosting system", Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, pp785-794. https://doi.org/10.1145/2939672.2939785

[50] Gunning, D. et al. (2019) "XAI: Explainable artificial intelligence", Science Robotics, Vol. 4, No. 37, eaay7120. https://doi.org/10.1126/scirobotics.aay7120

