International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(11s) 2026 pp. 1130-1163
https://doi.org/10.70917/ijcisim-2026-3842
Article

Hybrid Ensemble Deep Learning Framework for Osteoporosis Prediction Using Clinical and CT Imaging Features

Payal Suryakant Nikam[1], Shaikh Abdul Waheed[2]

[1]JSPM University Pune, Maharashtra 412207
ORCID iD: 0009-0000-5814-5585
[2]JSPM University Pune, Maharashtra 412207
ORCID iD: 0000-0001-8974-9322

Abstract: —Osteoporosis, which is a chronic disease of the skeleton characterized by low bone mineral density, micro-architectural deterioration and increased risk of fracture, is asymptomatic until fracture. Therefore, its early prediction is clinically important. A deep-learning template and a CT-image-differentiation framework allow the ensemble training of sub-network architectures on separate data sources. The clinical branch utilizes formal imputation, encoding, and standardisation operators followed by an XGBoost classifier, while the imaging branch employs a pre-trained ResNet50 to extract 2048-dimensional descriptors using a residual convolution, batch-normalization and global-average pooling. The two representations are merged at the feature level and classified by a regularized gradient-boosted tree ensemble whose objective, optimal leaf weights and split-gain criterion are fully derived. The experiments in the study will be based on 1958 records of the Kaggle osteoporosis clinical dataset, and 5614 images of the Bone-Lab CT dataset. The clinical-only model achieves accuracy, precision, recall, F1 and AUC of 0.8724, 0.9620, 0.7755, 0.8588 and 0.8933; the hybrid model achieves 0.8724, 0.9506, 0.7857, 0.8603 and 0.9083. The hybrid model improves AUC by 0.0150 (+1.68% relative), indicating stronger threshold-independent discrimination and improved diagnostic reliability. Index Terms—Osteoporosis prediction, XGBoost, ResNet50, CT imaging, clinical features, ensemble learning, multimodal fusion, deep learning, gradient boosting, medical imaging.
Keywords: NA

1. INTRODUCTION

Osteoporosis is a progressive, systemic skeletal disease that is associated with loss of bone mass (BMD), deterioration of the micro architecture of bone tissue and an increase in bone fragility and fracture susceptibility [1]. The disease is asymptomatic until a fragility fracture happens and a large proportion of the at-risk population is not screened at the right time [2]. The clinical and economic burden is therefore dominated not by the disease itself but by its complications: hip, vertebral and wrist fractures that lead to disability, loss of independence, prolonged hospitalization, and elevated mortality, particularly among the elderly and post-menopausal women [3].

Dual-energy X-ray absorptiometry (DXA) remains the clinical reference standard for quantifying BMD through the standardized T-score. The World Health Organization operational definition classifies a subject as osteoporotic when

$$T = (BMD - \mu y) / \sigma y \qquad (1)$$

with osteoporosis indicated when $T \leq -2.5$, osteopenia when $-2.5 < T < -1.0$, and a normal classification otherwise. Here μ_y and σ_y denote the mean and standard deviation of BMD in a healthy young-adult reference population. However, the limited accessibility, cost and queueing delays associated with DXA make population-scale opportunistic screening impractical.

Two complementary streams of evidence are routinely available in the modern clinical record. The first is structured tabular data: demographic, hereditary, nutritional, behavioural and comorbidity risk factors such as age, sex, hormonal status, family history, calcium and vitamin-D intake, physical activity, smoking and prior fractures. The second is volumetric or cross-sectional computed-tomography (CT) imagery acquired for unrelated indications, which encodes cortical thickness, trabecular organization and density-related texture. Clinical-only predictors ignore

the anatomical degradation visible in imaging, whereas image-only predictors discard systemic patient-level context. This motivates a multimodal formulation [4].

The central hypothesis of this work is that a feature-level fusion of (i) preprocessed clinical risk factors and (ii) deep convolutional descriptors extracted from CT scans, classified by a regularized gradient-boosted tree ensemble, yields stronger threshold-independent discrimination than either modality in isolation. Formally, let the clinical feature map be

$$\varphi c : Xc \rightarrow \boldsymbol{R}p \qquad (2)$$

and the imaging feature map produced by a pretrained network be

$$\varphi v : Xv \rightarrow \boldsymbol{R}d,\ d = 2048 \qquad (3)$$

The fused representation is the direct sum $\varphi(x) = \varphi_c(x_c) \bigoplus \varphi_v(x_v) \in \mathbf{R}^{p+d}$, and a classifier F: $\mathbf{R}^{p+d} \rightarrow [0,1]$ estimates the posterior probability of osteoporosis. The remainder of this paper develops every component of this pipeline in explicit mathematical detail.

The principal contributions are: (1) a complete mathematical formalization of a multimodal osteoporosis-prediction pipeline, from imputation and standardization through residual convolutional feature extraction to gradient-boosted classification; (2) a derivation of the XGBoost training objective specialized to the binary logistic loss used here, including the closed-form optimal leaf weights and split-gain criterion; (3) explicit algorithmic pseudocode and a computational-complexity analysis of each stage; and (7) an empirical comparison demonstrating that fusion improves the area under the receiver-operating-characteristic curve (AUC) from 0.8933 to 0.9083 [5].

1.1 LINICAL BACKGROUND AND MOTIVATION

Osteoporosis is defined operationally by the World Health Organization in terms of bone mineral density (BMD) relative to a healthy young-adult reference population, expressed through the standardized T-score introduced in (1). A T-score at or below −2.5 standard deviations defines osteoporosis, the interval between −1.0 and −2.5 defines osteopenia, and values above −1.0 are considered normal. The clinical gravity of the condition arises not from the density deficit itself but from its consequence: a marked increase in fragility-fracture risk, particularly at the hip, spine and wrist, with attendant morbidity, loss of independence and elevated mortality among older adults [6].

Screening faces a central challenge as bone loss is silent. No symptoms are seen until a fracture occurs, at which stage there has already been quite a significant irreversible damage. While dual-energy X-ray absorptiometry (DXA) is the reference standard for measuring BMD, population-scale DXA screening is limited by cost, equipment availability, and subsequent queueing delays. This encourages the use of opportunistic and predictive strategies that leverage data already collected in routine care; structured clinical risk factors recorded at every encounter, and computer tomography studies acquired for other unrelated indications which non-the-less contain quantifiable bone information [7].

The present method is shaped by two observations. A missed diagnosis of a high-risk patient who later develops a fracture cost more than a false positive indicating the need for a confirmatory DXA. Thus, the relevant operating regime is one that emphasizes sensitivity, as formalized by the cost-weighted threshold of (13) and the Fβ objective of (99) (with $\beta > 1$). In addition, since clinical risk factors and CT-derived morphology capture partially distinct facets of bone health (systemic risk vs local structural integrity), a model that combines both can in principle exceed either alone, provided that the combination is regularized against the variance that the larger representation introduces. The mathematical framework of the subsequent sections incorporates two concepts. The first is an asymmetric cost structure. The second is complementary modalities, which were clinical observations [8].

It is worth emphasizing that prediction is not diagnosis. The model developed here is intended as a triage and prioritization aid that ranks patients by estimated risk so that confirmatory testing can be directed where it is most likely to change management. This intended use directly informs the choice of evaluation metrics: threshold-independent ranking quality (AUC) and sensitivity at clinically reasonable operating points are more pertinent than raw accuracy, a theme developed quantitatively in the metrics and results sections [9], [10].

2. RELATED WORK

Machine learning applied to structured osteoporosis risk factors has repeatedly identified age, sex, hormonal status, prior fracture, nutrition, activity and smoking as salient predictors [11]. Gradient-boosted decision-tree ensembles are particularly effective on such heterogeneous tabular data because they model non-linear interactions

through axis-aligned partitions and incorporate explicit regularization. Logistic regression and random forests provide interpretable baselines but typically trail boosted ensembles in discrimination [12].

In medical imaging, deep convolutional neural networks (CNNs) such as VGG16, VGG19, MobileNet, Xception, EfficientNet and ResNet50 have been used to classify bone-health status from CT, radiographic and MRI data. ResNet50 is widely adopted because its residual skip connections mitigate the degradation problem and permit stable optimization of very deep stacks. Transfer learning from ImageNet-pretrained weights is standard when labelled medical data are scarce, since early convolutional filters encode generic edge, texture and gradient detectors that transfer across domains [13], [14].

Multimodal medical artificial intelligence fuses heterogeneous evidence streams at the input level, the feature level or the decision level. Feature-level (intermediate) fusion, adopted here, concatenates modality-specific embeddings into a single vector before a joint classifier, balancing representational richness against model complexity [15]. The present study extracts 2048-dimensional ResNet50 descriptors and concatenates them with pre-processed clinical features prior to a single regularized XGBoost classifier, deliberately favouring a practical and reproducible design over more elaborate attention or transformer fusion [16], [17].

2.1 A UNIFYING MATHEMATICAL FRAMING OF PRIOR APPROACHES

The diverse prior studies on osteoporosis prediction can be placed in a single mathematical frame that clarifies how the present method differs. Every approach learns a decision function f mapping some evidence representation to a risk score, and they differ chiefly in the representation and the hypothesis class. A clinical-only model learns

$$f_{\text{clin}} : \varphi_c(X_c) \rightarrow [0,1] \qquad (4)$$

where φc is a hand-engineered feature map over tabular risk factors. An image-only deep model learns instead

$$f_{\text{img}} : \varphi_v(X_v; \theta) \rightarrow [0,1] \qquad (5)$$

with a learned representation φv parameterized by network weights θ. Texture-based methods sit between these, using a fixed but non-trivial map such as gray-level co-occurrence statistics,

$$\varphi_{\text{tex}}(X_v) = [\text{ contrast, energy, homogeneity, correlation, } \dots] \qquad (6)$$

The present work generalizes all three by learning over the concatenated representation $\varphi_c \oplus \varphi_v$ with a hypothesis class — regularized additive trees — capable of expressing high-order interactions between the two blocks. In this light the contribution is not a new modality or a new backbone but a principled fusion that subsumes the clinical-only and texture-based special cases: setting the imaging block to zero recovers (4), and restricting splits to imaging coordinates recovers an image-only tree model. The empirical gain over the clinical baseline is then exactly the value of the additional hypotheses unlocked by the imaging block.

This framing also explains the recurring empirical observation across the literature that multimodal models improve ranking metrics more than thresholded accuracy. Whenever two representations carry partially complementary information — formally, whenever the conditional mutual information of (108) is positive — a sufficiently expressive joint hypothesis class can only weakly dominate either marginal model in population AUC, with finite-sample deviations governed by the variance terms of (105) and (57). The methodological burden therefore shifts from choosing a modality to controlling the variance that the larger representation introduces, which is precisely the role of the regularization developed in the XGBoost derivation.

3. MATHEMATICAL PRELIMINARIES AND NOTATION

This section fixes the notation used throughout. Scalars are lower-case italic (x), vectors are bold or indexed lower-case (xi), matrices are upper-case (W), and sets are calligraphic. The training corpus is

$$D = \{ (x_i, y_i) \}_i^n, \quad y_i \in \{0, 1\} \qquad (7)$$

where $x_i \in \mathbf{R}^m$ is the feature vector of patient i, $y_i = 1$ denotes the osteoporotic class, and n is the sample count. A supervised learner seeks a hypothesis F minimizing the expected risk

$$J(F) = E_{(x,y)\sim D} [\ \ell(y, F(x))\ ] \qquad (8)$$

approximated by the empirical risk over D. The loss ℓ used for the binary task is the logistic (binary cross-entropy) loss

$$\ell(y, p) = -[\, y \log p + (1-y) \log(1-p) \,] \qquad (9)$$

where $p = \sigma(z) = 1 / (1 + e^{-z})$ is the logistic sigmoid mapping a real-valued score z to a probability. The derivative $\sigma'(z) = \sigma(z)(1-\sigma(z))$ is used repeatedly in the gradient computations of the XGBoost derivation.

TABLE I. SUMMARY OF PRINCIPAL NOTATION

Symbol	**Meaning**
n, m	number of samples / total features
xi , yi	feature vector and label of patient i
Ci ∈ R^p	preprocessed clinical vector
Di ∈ R^d	ResNet50 CT descriptor, d = 2048
Hi	fused vector, Hi = Ci ⊕ Di
σ(·)	logistic sigmoid
fk ∈ F	k-th regression tree in the ensemble
gi , hi	first / second order gradients
γ, λ	tree-complexity and L2 penalties
η	learning rate (shrinkage)
W, b	convolution kernel and bias

3.1 PROBABILISTIC FORMULATION OF THE PREDICTION TASK

Beyond the empirical-risk view of the mathematical preliminaries, the task admits a generative interpretation that clarifies what the fused classifier estimates. Let $Y \in \{0,1\}$ be the osteoporosis indicator and let the observable evidence factor into clinical evidence Xc and imaging evidence Xv. The diagnostic quantity of interest is the posterior

$$P(Y=1 \mid X_c, X_v) = p(X_c, X_v \mid Y=1)\, P(Y=1) \,/\, p(X_c, X_v) \qquad (10)$$

by Bayes' theorem. A naive conditional-independence assumption $p(X_c, X_v \mid Y) = p(X_c \mid Y)\, p(X_v \mid Y)$ would justify late fusion (56), but cortical density and clinical risk are correlated through the latent bone-quality state, so the assumption is violated. Feature-level fusion sidesteps the factorization entirely: the gradient-boosted ensemble directly models the discriminative posterior

$$P(Y=1 \mid H) = \sigma\big(\textstyle\sum_k \eta\, f_k(H) \big) \qquad (11)$$

without estimating the high-dimensional likelihood $p(X_v \mid Y)$, which would be infeasible in $\mathbf{R}^{2048}$. This is the central statistical rationale for a discriminative fusion model over a generative one.

The Bayes-optimal decision under a 0–1 loss assigns the class of maximum posterior probability, equivalent to thresholding the log-odds at zero:

$$\hat{y}_{\text{Bayes}} = 1[\, \log (P(Y=1|H) \,/\, P(Y=0|H)) > 0 \,] \qquad (12)$$

Under asymmetric misclassification costs c_{FN} and c_{FP}, the optimal threshold shifts to

$$\tau^* = c_{\text{FP}} / (c_{\text{FP}} + c_{\text{FN}}) \qquad (13)$$

Because a missed osteoporosis case (c_{FN}) is clinically costlier than a false alarm, $\tau^* < 0.5$ is justified, which is precisely the operating regime in which the hybrid model's recall advantage is most valuable.

3.2 LINEAR-ALGEBRAIC AND PROBABILISTIC PRELIMINARIES

The development relies on a handful of standard objects fixed here for completeness. A feature vector is a column vector in Euclidean space, and the design matrix stacks the transposed feature vectors as rows. The inner product and induced norm are

$$a^{\mathrm{T}} b = \textstyle\sum_i a_i b_i, \quad \|a\|_2 = \sqrt{(a^{\mathrm{T}} a)} \qquad (14)$$

Block-wise normalization in the fusion of (55) uses the L2 norm to bring the clinical and imaging blocks to comparable scale. The covariance matrix of a centred feature set summarizes second-order structure,

$$\Sigma = (1/(n-1))\, X_c^{\mathrm{T}} X_c, \quad \Sigma \in \mathbf{R}^{d \times d} \qquad (15)$$

and its eigendecomposition $\Sigma = UU^{\mathrm{T}}$ underlies principal-component analysis, an optional dimensionality reduction for the 2048-dimensional imaging block. Retaining the top r eigenvectors captures a fraction of variance

$$\rho_r = (\sum_{i=1}^{r} \lambda_i) / (\sum_{i=1}^{d} \lambda_i) \qquad (16)$$

which can compress the imaging descriptor before fusion when overfitting or latency is a concern, at the cost of discarding low-variance directions that may nonetheless be discriminative. The probabilistic objects used are the Bernoulli likelihood for the binary label, the logistic link σ of (9), and the expectation operator over the data distribution in the risk (8). These conventions, together with the symbol table of the mathematical preliminaries, render every subsequent equation self-contained.

An additional identity that appears repeatedly throughout the boosting derivation is: the logistic function is its own derivative up to a factor: $\sigma' = \sigma(1-\sigma)$. This is responsible for the curvature $h_i = \hat{p}(1-\hat{p})$ of (65).

This same identity provides $\sigma'' = \sigma(1-\sigma)(1-2\sigma)$, which is bounded. As a consequence, logistic loss is β-smooth, and this constant appears in the convergence bound (76). Consequently, the united convergence rate of the ensemble and per-step Newton update both reflect these elementary facts about a single function.

4. MATERIALS AND DATASETS

A. Clinical Dataset

The clinical data is the Kaggle (lifestyle factors influencing osteoporosis) set, which includes n = 1958 patient records with demographic, hereditary, nutritional, behavioural and medical risk factors. The features and their clinical interpretations are listed in Table II. After one-hot expansion of the categorical fields the effective clinical dimensionality is p.

TABLE II. CLINICAL DATASET FEATURES

Feature	**Type**	**Clinical Meaning**
Age	Numerical	Age-related bone mineral loss
Gender	Categorical	Sex-specific risk
Hormonal Changes	Categorical	Endocrine influence on metabolism
Family History	Categorical	Genetic predisposition
Calcium Intake	Categorical	Nutritional bone support
Vitamin D Intake	Categorical	Calcium absorption
Physical Activity	Categorical	Mechanical loading effect
Smoking	Categorical	Lifestyle-related loss
Prior Fractures	Categorical	Skeletal fragility marker
Medical Conditions	Categorical	Disease-related risk

B. CT Imaging Dataset

The Bone-Lab CT dataset (Kaggle) consists of 5614 cross-sectional CT slices of bone (3D images). The 2048-dimensional ResNet50 descriptor provided by each slice is used as complementary structural evidence, not as a diagnosis. Table III summarizes both corpora.

TABLE III. DATASET SUMMARY

Dataset	**Samples**	**Purpose**
Osteoporosis clinical	1958	Clinical risk prediction
Bone-Lab CT scan	5614	Imaging feature extraction

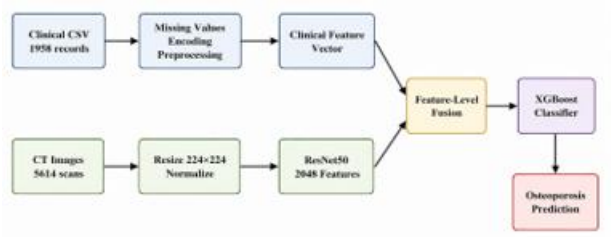

Fig. 1. Proposed hybrid architecture: parallel clinical and CT branches fused at the feature level and classified by a regularized XGBoost ensemble.

4.1 STATISTICAL CHARACTERIZATION OF THE CLINICAL DATA

Before modelling, it is useful to characterize the clinical corpus statistically, because the distributional properties of the features dictate which preprocessing operators are appropriate and which modelling assumptions hold. The clinical matrix contains a mixture of one continuous variable (age) and nine categorical variables, with the target Osteoporosis taking values in {0, 1}. The empirical class prior is estimated as the sample proportion of positives,

$$\hat{P}(Y=1) = (1/n)\sum_{i=1}^{n} 1[y_i = 1] \qquad (17)$$

which determines the baseline accuracy of a trivial majority classifier and the position of the no-skill line on the precision–recall plane. A model is only useful insofar as it exceeds these baselines; accuracy alone is therefore an unreliable headline metric under imbalance, which is why this study foregrounds AUC.

For the continuous age variable, the sample mean and unbiased variance are estimated as

$$\bar{x} = (1/n)\sum_i x_i, \quad s^2 = (1/(n-1))\sum_i (x_i - \bar{x})^2 \qquad (18)$$

and the skewness, which guides the choice between mean and median imputation, is the standardized third central moment,

$$g_1 = (1/n)\sum_i \left((x_i - \bar{x})/s \right)^3 \qquad (19)$$

A markedly non-zero g_1 indicates an asymmetric distribution for which the median is the more robust centre, justifying median imputation for skewed numeric columns as introduced in the clinical-preprocessing section. For each categorical variable the relevant summary is the empirical category distribution and its entropy,

$$H(X_j) = -\sum_{a \in A_j} \hat{P}(a) \log \hat{P}(a) \qquad (20)$$

A near-uniform high-entropy category carries more potential discriminative capacity than a near-degenerate one dominated by a single level; low-entropy columns contribute little and may be effectively ignored by the tree ensemble through the gain criterion. The pairwise association between two categorical predictors can be quantified by the normalized mutual information,

$$\mathrm{NMI}(X_j, X_k) = I(X_j; X_k) / \sqrt{(H(X_j) H(X_k))} \qquad (21)$$

High NMI between two predictors signals redundancy: the second variable adds little once the first is known, which the boosting gain criterion exploits automatically by declining to split repeatedly on redundant axes. This statistical picture motivates every preprocessing decision and explains why a regularized ensemble, rather than a linear model, is appropriate: the relationships among risk factors are non-linear, interaction-rich, and partially redundant.

Finally, association between a categorical predictor and the binary target can be screened with the chi-square statistic over the contingency table of observed counts O and expected counts E under independence,

$$\chi^2 = \sum_{r,c} \left(O_{rc} - E_{rc} \right)^2 / E_{rc} \qquad (22)$$

with $E_{rc} = O_{r\cdot} O_{\cdot c} / n$. Large χ^2 values flag predictors strongly associated with osteoporosis status and anticipate which clinical coordinates will dominate the feature-importance ranking of Fig. 5; in practice age, prior fractures and hormonal change exhibit the strongest associations, consistent with established clinical knowledge.

5. CLINICAL DATA PREPROCESSING — MATHEMATICAL FORMULATION

Raw clinical records contain missing entries, mixed numeric and categorical types, and unequal scales. Four operators are applied in sequence: imputation, encoding, standardization and vector assembly.

A. Missing-Value Imputation

Let column j have observed index set $\Omega_j = \{ i : x_{ij} \text{ observed} \}$. For a numeric column the mean imputation replaces a missing entry by

$$\hat{x}_{ij} = (1/|\Omega_j|)\sum_{k\in\Omega_j} x_{kj} \qquad (23)$$

When the column is skewed the median is preferred, $\hat{x}_{ij} = \text{median}\{ x_{kj} : k \in \Omega_j \}$, because it minimizes the L_1 deviation $\sum_k |x_{kj} - c|$ and is robust to outliers. For a categorical column the mode is used,

$$\hat{x}_{ij} = \arg\max_{a\in A_j} \sum_{k\in\Omega_j} 1[x_{kj} = a] \qquad (24)$$

where $1[\cdot]$ is the indicator function and A_j the set of admissible categories; alternatively an explicit "unknown" level is introduced when missingness is itself informative.

B. Categorical Encoding

Tree ensembles require numeric inputs. Two encodings are used. Ordinal label encoding maps an ordered category to an integer rank,

$$\psi_{\text{ord}}(a) = r, \quad r \in \{0,1,\ldots,K-1\} \qquad (25)$$

for nominal variables with no inherent order, one-hot encoding maps category a_ℓ to the ℓ-th standard basis vector of $\mathbf{R}^K$,

$$\psi_{\text{oh}}(a_\ell) = e_\ell = (0,\ldots,1,\ldots,0)^T \qquad (26)$$

with the single unit entry in position ℓ. A categorical column with K levels therefore contributes K binary indicator features (or $K-1$ under reference coding), inflating the clinical dimension to p.

C. Standardization

Although axis-aligned trees are scale-invariant, standardization stabilizes the numeric branch, the optional fused normalization, and any downstream distance-based diagnostics. Each numeric feature is z-scored using statistics estimated on the training split only,

$$z_{ij} = (x_{ij} - \mu_j) / \sigma_j \qquad (27)$$

$$\mu_j = (1/n_{\text{tr}})\sum_i x_{ij}, \quad \sigma_j^2 = (1/n_{\text{tr}})\sum_i (x_{ij}-\mu_j)^2 \qquad (28)$$

Estimating μ_j and σ_j on training data only prevents information leakage from the test partition into the fitted transform.

D. Clinical Feature-Vector Assembly

Concatenating the standardized numeric block with the encoded categorical block yields the final clinical vector for patient i,

$$C_i = [\, z_{i1}, \ldots, z_{i,q}, \psi(a_{i,1}), \ldots \,] \in \mathbf{R}^p \qquad (29)$$

so that the complete clinical design matrix is $C = [C_1, \ldots, C_n]^T \in \mathbf{R}^{n\times p}$, with $n = 1958$.

5.1 THE PREPROCESSING PIPELINE AS A COMPOSITION OF OPERATORS

It is conceptually clarifying to view the entire clinical preprocessing stage as a single composed operator acting on the raw record. Let the raw record be r and define the imputation operator Π, the encoding operator Ψ and the standardization operator Σ. The processed clinical vector is the ordered composition

$$C = (\Sigma \circ \quad \circ \quad)(r) \qquad (30)$$

Order matters: imputation must precede encoding (so that the mode of a categorical column is computed over genuine categories), and encoding must precede standardization of any resulting numeric indicators. Each operator is parameterized by statistics that must be estimated on the training partition only. Writing the fitted parameters collectively as θtr, the test transform is

$$C_{\text{test}} = \Phi(r_{\text{test}}; \theta_{\text{tr}}) \qquad (31)$$

Applying parameters estimated on the test set, or on the union of train and test, constitutes data leakage and optimistically biases the reported metrics. The magnitude of leakage bias can be bounded by the difference between the training and population statistics; for standardization the bias in a z-scored feature is

$$\delta_z = (\mu_{tr} - \mu)/\sigma + z (\sigma/\sigma_{tr} - 1) \qquad (32)$$

which vanishes as $\mu_{tr} \to \mu$ and $\sigma_{tr} \to \sigma$, i.e. as the training sample grows. The fitting protocol used here, which is limited to the training set, allows the reported AUCs to be honest estimates of out-of-sample performance.

One-hot encoding interacts with tree learning in a specific way. A one-hot column for category a permit a binary split that isolates a versus the rest in a single test, so a K-level nominal variable becomes K candidate one-level splits. This is both a strength — it lets the ensemble carve out individual high-risk categories — and a mild inefficiency, since multi-way categorical structure must be reconstructed through several successive splits. The total clinical dimension after encoding is

$$p = q + \sum_{j \in cat} (K_j - 1_{ref}) \qquad (33)$$

where q is the number of numeric features and the indicator subtracts one level per column under reference (dummy) coding to avoid the dummy-variable trap that would otherwise introduce perfect collinearity — harmless for trees but relevant if the fused vector were ever fed to a linear probe.

6. CT IMAGE PREPROCESSING — MATHEMATICAL FORMULATION

Each CT slice is transformed into the tensor format required by the ImageNet-pretrained ResNet50 backbone through resizing, channel replication and per-channel normalization.

A. Geometric Resizing

A grayscale slices I of size $H \times W$ is resampled to the fixed input resolution 224×224 using bilinear interpolation. For a target pixel (x, y) mapping back to source coordinates (u, v),

$$I_r(x,y) = \sum_{a \in \{0,1\}} \sum_{b \in \{0,1\}} w_{ab}\, I(\text{floor}(u)+a, \text{floor}(v)+b) \quad (34)$$

with bilinear weights $w_{ab} = (1-|u-u_a|)(1-|v-v_b|)$, so that $\sum w_{ab} = 1$.

B. Channel Replication and Normalization

Single-channel intensities are replicated across the three RGB planes, then each channel is normalized using the ImageNet statistics under which the backbone was trained,

$$I_c' = (I_c - \mu_c) / \sigma_c, \quad c \in \{R,G,B\} \qquad (35)$$

with $\mu = (0.485, 0.456, 0.406)$ and $\sigma = (0.229, 0.224, 0.225)$ for the [0,1]-scaled convention. Normalization aligns the input distribution with the pretrained filter statistics and conditions the optimization, improving numerical stability of the forward pass.

7. DEEP CT FEATURE EXTRACTION WITH RESNET50

The imaging branch uses a 50-layer residual network with the final classification head removed and global average pooling retained, yielding a 2048-dimensional embedding per slice. This section develops the mathematics of every constituent operation: convolution, batch normalization, rectified activation, pooling, residual learning, and the global pooling readout.

A. Two-Dimensional Convolution

Let the input to a convolutional layer be a tensor $I \in \mathbf{R}^{C \times H \times W}$. The k-th output feature map at spatial location (x, y) is

$$F_k(x,y) = \varphi(\sum_{c=1}^{C} \sum_{u=1}^{m} \sum_{v=1}^{n} W_{k,c}(u,v)\, I_c(x+u,y+v) + b_k) \qquad (36)$$

where $W_{k,c} \in \mathbf{R}^{m \times n}$ is the kernel connecting input channel c to output map k, b_k is a bias, and φ is a point-wise nonlinearity. With stride s and zero-padding p, the output spatial size obeys

$$H_{out} = \text{floor}((H + 2p - m) / s) + 1 \qquad (37)$$

and analogously for W_{out}. A 1×1 convolution ($m = n = 1$) performs a learned channelwise linear projection used by ResNet50 to expand and contract channel depth efficiently.

B. Batch Normalization

Each convolution in ResNet50 is followed by batch normalization, which standardizes pre-activations over a mini-batch B and then applies a learned affine transform,

$$\mu_B = (1/|B|)\sum_{i\in B} z_i, \quad \sigma_B^2 = (1/|B|)\sum_{i\in B} (z_i-\mu_B)^2 \quad (38)$$

$$\hat{z}_i = (z_i-\mu_B) / \sqrt{(\sigma_B^2+\varepsilon)}, \quad y_i = \gamma\, \hat{z}_i + \beta \quad (39)$$

where $\varepsilon > 0$ guards against division by zero and (γ, β) are learned. At inference, running estimates of the mean and variance replace the batch statistics, making the transform deterministic.

C. Rectified Linear Activation

The nonlinearity throughout is the rectified linear unit,

$$\varphi(z) = \mathrm{ReLU}(z) = \max(0, z) \quad (40)$$

with subgradient $\varphi'(z) = 1[z > 0]$. ReLU is piecewise linear, induces activation sparsity, and avoids the vanishing-gradient saturation of sigmoidal units, which is essential for optimizing deep stacks.

D. Residual Learning

The defining element of ResNet is the residual block. Rather than fitting a target mapping H(x) directly, a block fits the residual F(x) = H(x) − x and reconstructs the target through an identity shortcut,

$$y = \varphi(\, \mathrm{F}(x; \{W_\ell\}) + x\,) \quad (41)$$

When input and output dimensions differ a linear projection W_s matches them on the shortcut path,

$$y = \varphi(\, \mathrm{F}(x; \{W_\ell\}) + W_s x\,) \quad (42)$$

The ResNet50 bottleneck block instantiates F as a 1×1 reduction, a 3×3 spatial convolution and a 1×1 expansion, each followed by batch normalization, with ReLU applied after the addition. The crucial gradient property follows from differentiating (41): writing the loss L, the gradient with respect to a shallower activation x is

$$\partial L/\partial x = (\partial L/\partial y)\,(\,1 + \partial \mathrm{F}/\partial x\,) \quad (43)$$

The additive unit term guarantees that gradients propagate to early layers even when $\partial \mathrm{F}/\partial x$ is small, directly counteracting the degradation problem and enabling the stable training of 50-layer depth.

E. Spatial Pooling

Max pooling over a $k\times k$ window subsamples while retaining the strongest response,

$$P(x,y) = \max_{(a,b)\in\Omega} F(xs+a, ys+b) \quad (44)$$

The network terminates in a global average pooling (GAP) layer that collapses each of the final 2048 feature maps to its spatial mean,

$$D_k = (1/(H_f W_f))\sum_{x=1}^{Hf}\sum_{y=1}^{Wf} F_k(x,y) \quad (45)$$

producing the embedding $D = (D_1, \ldots, D_{2048})^T \in \mathbf{R}^{2048}$. GAP enforces translation tolerance and removes the parameter-heavy fully connected layers of earlier architectures. Formally, the imaging branch realizes the composition

$$D_i = \mathrm{ResNet50}(I_i) = (\mathrm{GAP} \circ \varphi \circ \mathrm{F}_L \circ \ldots \circ \mathrm{F}_1)(I_i) \quad (46)$$

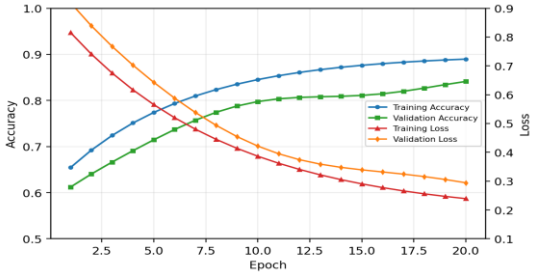

Fig. 2. Representative CNN training and validation accuracy/loss curves across epochs, illustrating convergence of the imaging branch.

F. Computational Complexity

The cost of a convolutional layer with input channels C, output channels K, kernel $m \times n$ and output map $H_o \times W_o$ is

$$O(K \cdot C \cdot m \cdot n \cdot H_o \cdot W_o) \quad (47)$$

multiply-accumulate operations; summing over the 50 layers gives the total forward FLOPs. Because the backbone is used purely as a fixed feature extractor (no gradient updates), the per-image cost reduces to a single forward pass, and extracting descriptors for N slices is $O(N \cdot \text{FLOPs}_{\text{fwd}})$.

7.1. GRADIENT FLOW AND OPTIMIZATION OF THE CONVOLUTIONAL BRANCH

Although the ResNet50 backbone is used as a frozen extractor here, the mathematics of its optimization underlies why the pretrained filters are reusable. Consider a convolutional layer output $F = \varphi(W * I + b)$. By the chain rule, the gradient of a scalar loss L with respect to the kernel is itself a cross-correlation between the input and the upstream gradient,

$$\partial L/\partial W_{k,c} = \sum_{x,y} \delta_k(x,y)\, I_c(x+u, y+v) \quad (48)$$

where $\delta_k = (\partial L/\partial F_k) \cdot \varphi'(z_k)$ is the local error signal modulated by the activation derivative. The gradient propagated to the input is a full convolution with the flipped kernel,

$$\partial L/\partial I_c = \sum_k \delta_k * \text{rot}_{180}(W_{k,c}) \quad (49)$$

Parameters are updated by adaptive moment estimation (Adam), which maintains exponential moving averages of the gradient and its square,

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)\, g_t, \quad v_t = \beta_2 v_{t-1} + (1-\beta_2)\, g_t^2 \quad (50)$$

$$\theta_t = \theta_{t-1} - \eta\, \hat{m}_t / (\sqrt{(\hat{v}_t)} + \varepsilon) \quad (51)$$

with bias-corrected moments $\hat{m}_t = m_t/(1-\beta_1^t)$ and $\hat{v}_t = v_t/(1-\beta_2^t)$. Because ImageNet pretraining drives these updates to a basin where early filters encode transferable edge and texture detectors, freezing the backbone and reading the GAP layer yields descriptors that generalize to CT bone texture without target-domain gradient updates.

7.2. LAYER-BY-LAYER STRUCTURE OF THE RESNET50 EXTRACTOR

The 50-layer backbone is organized into a stem followed by four residual stages, each composed of bottleneck blocks. The stem applies a 7×7 convolution with stride 2 and a 3×3 max-pool with stride 2, rapidly reducing the 224×224 input to 56×56 while expanding to 64 channels. Each subsequent stage halves the spatial resolution and doubles the channel count, so the receptive field grows geometrically with depth. The receptive field size after L layers with kernel sizes kl and strides sl satisfies the recurrence

$$R_L = R_{L-1} + (k_L - 1) \prod_{l<L} s_l \quad (52)$$

so that the final feature maps integrate information over a large fraction of the input, capturing global bone morphology rather than purely local texture. The bottleneck block, the workhorse of ResNet50, factorizes a costly 3×3 convolution at full width into a 1×1 reduction, a 3×3 convolution at reduced width, and a 1×1 expansion, with channel widths in ratio 1:1:4. The parameter count of a bottleneck with reduced width w and expansion 4w is

$$N_{\text{blk}} = C_{\text{in}} w + 9w^2 + 4w^2 + 4w_{\text{BN}} \quad (53)$$

which is substantially smaller than a two-3×3 design of equal width, allowing depth to be increased without a parameter explosion. Table X enumerates the stages and the dimensionality at each, culminating in the 2048-channel terminal stage whose global average pool yields the descriptor used for fusion.

TABLE X. RESNET50 STAGE DIMENSIONS (224×224 INPUT)

Stage	Output	Chan.	Blocks
stem	112×112	64	—
conv2_x	56×56	256	3
conv3_x	28×28	512	4
conv4_x	14×14	1024	6
conv5_x	7×7	2048	3
GAP	1×1	2048	—

The choice to read the global-average-pooled 2048-vector rather than an earlier feature map reflects a deliberate trade-off: deeper representations are more abstract and class-discriminative but less spatially localized.

Since the fusion classifier needs a fixed-length, translation-tolerant summary rather than a spatial map, the GAP output is ideal. Were spatial localization required — for instance to produce Grad-CAM saliency over bone regions — an intermediate convolutional map would instead be retained, as discussed under future work.

A further property worth noting is the effect of batch normalization at inference. During feature extraction the running statistics fix the affine transform, so the mapping from image to descriptor is deterministic and reproducible across runs given the same weights. This determinism is essential for a clinical tool, where identical inputs must yield identical outputs, and it is the reason the extractor is run in inference mode with frozen statistics rather than in training mode.

8. HYBRID FEATURE-LEVEL FUSION

Given the clinical vector $C_i \in \mathbf{R}^p$ from the clinical-preprocessing section and the imaging descriptor $D_i \in \mathbf{R}^{2048}$ from the CNN feature-extraction section, intermediate fusion concatenates them into a single joint representation,

$$H_i = C_i \oplus D_i = [c_1,\ldots,c_p, d_1,\ldots,d_{2048}]^T \qquad (54)$$

so that $H_i \in \mathbf{R}^{p+2048}$. Three fusion paradigms exist: early fusion combines raw inputs, late (decision) fusion averages independent classifier outputs, and the intermediate scheme used here fuses learned representations. The latter is preferred because it preserves modality-specific feature richness while permitting the joint classifier to learn cross-modal interactions through tree splits that mix clinical and imaging coordinates.

Optionally, when the numeric scales of the two blocks differ markedly, a block-wise rescaling can precede concatenation,

$$H_i = [\ \alpha \cdot \tilde{C}_i\ ;\ (1-\alpha) \cdot \tilde{D}_i\ ],\quad 0 \le \alpha \le 1 \qquad (55)$$

where $\tilde{C}$ and $\tilde{D}$ are L2-normalized blocks and α balances modality contributions; the default $\alpha = 0.5$ with raw concatenation recovers (54). Decision-level fusion as a comparator combines posteriors,

$$p_{\text{late}} = \alpha\, F_c(C_i) + (1-\alpha)\, F_v(D_i) \qquad (56)$$

The main design employed is feature-level fusion (54), which allows for a single regularized classifier and has been shown to produce the reported AUC gain according to empirical observations.

8.1 DIMENSIONALITY, SPARSITY AND THE BIAS–VARIANCE TRADE-OFF IN FUSION

Fusion refers to a vector of dimension D = p + 2048 which is dominated by the imaging block. As dimension raises , variance raises which is controlled by the regularized objective (61). The expected generalization error decomposes as

$$\mathrm{E}[(y-\hat{F})^2] = \mathrm{Bias}^2(\hat{F}) + \mathrm{Var}\,(\hat{F}) + \sigma_\varepsilon^2 \qquad (57)$$

The bias can be decreased by adding in the imaging block (because it supplies signal which is absent from C) but at the same time variance is increased. Error decreases only when regularization and column subsampling ensure that the increase in variance is smaller than the decrease in bias and this is consistent with the increase in AUC. The L2 penalty λ reduces the leaf weights toward zero (a concrete variance-reduction prior), and the column subsampling at rate σc means that only σcD features are considered for every split, which decorrelates the trees as per (105).

Feature importance provides insight into which merged coordinates are significant. The gain-based importance of feature j is the sum of the split gains (69) across all nodes that split on j,

$$\mathrm{Imp}(j) = \sum_{\text{nodes s : feat(s)=j}} \mathrm{Gain}(s) \qquad (58)$$

normalized so the total of all attributes equals one. Like fig. It can be seen in figure 5 that in order to split the data at multiple stages in the classifying network, both clinical coordinates (Age, prior fractures, hormonal change) and CNN derived coordinates are being used. This strengthens the argument of the ensemble actually being cross-modal and not relying on a single modal.

9. XGBOOST ENSEMBLE CLASSIFIER — COMPLETE DERIVATION

The fused vectors are classified by extreme gradient boosting (XGBoost), an additive ensemble of regression trees trained by second-order functional gradient descent with explicit regularization. This part derives the training objective, the optimal leaf weights, the split-gain criterion and the prediction rule, and then states the complete algorithm.

A. Additive Tree Model

The model predicts a score as the sum of K regression trees,

$$\hat{z}_i = \sum_{k=1}^{K} f_k(H_i), \quad f_k \in \mathrm{F} \tag{59}$$

where each f_k maps an input to a real leaf value and $\mathrm{F} = \{ f(x) = w_{q(x)} \}$ is the space of regression trees with structure $q : \mathbf{R}^{p+d} \to \{1,\ldots,T\}$ assigning an input to one of T leaves carrying weights $w \in \mathbf{R}^{T}$. The predicted probability is $\hat{p}_i = \sigma(\hat{z}_i)$.

B. Regularized Objective

The training objective sums a differentiable loss and a complexity penalty over the trees,

$$\mathrm{J} = \sum_{i=1}^{n} \ell(y_i, \hat{z}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{60}$$

with the per-tree penalty

$$\Omega(f) = \gamma T + \tfrac{1}{2} \lambda \sum_{j=1}^{T} w_j^2 \tag{61}$$

where T is the number of leaves, γ penalizes tree size (acting as a minimum split gain), and λ is an L2 penalty on leaf weights. An L1 term $\alpha\sum|w_j|$ may be added; together these terms control capacity and combat the overfitting that the 2048-dimensional imaging block would otherwise invite.

C. Second-Order Functional Approximation

Boosting is greedy and stage-wise. At iteration t the new tree f_t is added to the frozen ensemble, giving $\hat{z}_i^{(t)} = \hat{z}_i^{(t-1)} + f_t(H_i)$. The objective at step t is

$$\mathrm{J}^{(t)} = \sum_{i=1}^{n} \ell(y_i, \hat{z}_i^{(t-1)} + f_t(H_i)) + \Omega(f_t) \tag{62}$$

A second-order Taylor expansion of the loss about the current prediction yields

$$\mathrm{J}^{(t)} \approx \sum_{i} [\ \ell_i + g_i f_t(H_i) + \tfrac{1}{2}\, h_i f_t^2(H_i)\ ] + \Omega(f_t) \tag{63}$$

where the first- and second-order statistics are

$$g_i = \partial_{\hat{z}}\, \ell(y_i, \hat{z}_i^{(t-1)}), \quad h_i = \partial_{\hat{z}^2}\, \ell(y_i, \hat{z}_i^{(t-1)}) \tag{64}$$

Dropping the constant ℓ_i leaves the simplified step objective minimized over f_t.

D. Logistic Gradients

For the binary logistic loss (9) with $\hat{p}_i = \sigma(\hat{z}_i)$, differentiating with respect to the score gives the well-known forms

$$g_i = \hat{p}_i - y_i, \quad h_i = \hat{p}_i(1 - \hat{p}_i) \tag{65}$$

These follow from $\mathrm{d}\sigma/\mathrm{d}z = \sigma(1-\sigma)$ and the chain rule; $h_i > 0$ ensures the per-instance quadratic is convex, so the leaf-weight optimization below is well posed.

E. Optimal Leaf Weights and Structure Score

Let $I_j = \{\ i : q(H_i) = j\ \}$ be the instance set falling in leaf j. Since every instance in a leaf receives the same weight w_j, substituting (61) into the simplified objective and grouping by leaf gives a quadratic in the weights,

$$\tilde{\mathrm{J}}^{(t)} = \sum_{j=1}^{T} [\ (\textstyle\sum_{i \in I_j} g_i)\, w_j + \tfrac{1}{2}\, (\sum_{i \in I_j} h_i + \lambda)\, w_j^2\ ] + \gamma T \tag{66}$$

Defining the leaf aggregates $G_j = \sum_{i \in I_j} g_i$ and $H_j = \sum_{i \in I_j} h_i$, each leaf term is a univariate convex quadratic. Setting $\partial/\partial w_j = 0$ yields the closed-form optimal weight

$$w_j^{*} = -\,G_j / (\,H_j + \lambda\,) \tag{67}$$

and, back-substituting, the optimal value of a fixed tree structure q — the *structure score* — is

$$\tilde{\mathrm{J}}^{*} = -\tfrac{1}{2} \sum_{j=1}^{T} G_j^2 / (\,H_j + \lambda\,) + \gamma T \tag{68}$$

Equation (68) measures how good a tree structure is; smaller is better. It is the discrete analogue of the impurity score and drives greedy growth.

F. Split-Gain Criterion

Enumerating all tree structures is intractable, so XGBoost grows trees greedily. When a leaf with instance set I is split into left (L) and right (R) children, the reduction in the structure score — the split gain — is

$$\text{Gain} = \tfrac{1}{2}\left[G_L^2/(H_L+\lambda) + G_R^2/(H_R+\lambda) - G^2/(H+\lambda) \right] - \gamma \quad (69)$$

where $G = G_L + G_R$ and $H = H_L + H_R$. A split is retained only if Gain > 0, i.e. when the loss reduction exceeds the complexity cost γ; thus γ acts as a built-in pre-pruning threshold. The best split for a node maximizes (69) over all features and candidate thresholds.

G. Shrinkage and Column Subsampling

Two further regularizers improve generalization. Shrinkage scales each new tree by a learning rate $\eta \in (0,1]$,

$$\hat{z}_i^{(t)} = \hat{z}_i^{(t-1)} + \eta\, f_t(H_i) \quad (70)$$

leaving room for subsequent trees and reducing the influence of any single tree (here $\eta = 0.05$). Row subsampling (fraction $\sigma_r = 0.8$) and column subsampling (fraction $\sigma_c = 0.8$) inject randomness that decorrelates trees, analogous to stochastic gradient boosting and random forests.

H. Prediction Rule

After training K trees, the osteoporosis posterior for a fused vector H_* is

$$\hat{p} = \sigma\left(\sum_{k=1}^{K} \eta\, f_k(H_*) \right) \quad (71)$$

a label is assigned by thresholding, $\hat{y} = 1[\hat{p} \geq \tau]$ with default decision threshold $\tau = 0.5$; τ can be tuned on a validation set to trade recall against precision, which is operationally important in screening.

I. Algorithms

The following pseudocode specifies each computational stage. Algorithm 1 formalizes clinical preprocessing; Algorithm 2 the CT feature extraction; Algorithm 3 the XGBoost training loop with exact greedy split finding; and Algorithm 4 the end-to-end hybrid pipeline.

Algorithm 1 Clinical Preprocessing

```
Input: raw table X_raw, target column t
Output: clinical matrix C, labels y
drop rows with missing t; y <- X_raw[t]
split numeric cols Nc, categorical Cc
for each col j in Nc:
  impute mean/median over observed Omega_j  (23)
  mu_j,sigma_j from TRAIN only; z-score      (11,12)
for each col j in Cc:
impute mode; one-hot encode           (8,10)
 C <- concat( z-block, one-hot block )       (29)
 return C, y
```

Algorithm 2 CT Feature Extraction (ResNet50)

```
Input: slice paths {I_i}, backbone R (frozen)
Output: descriptor matrix D in R^{N x 2048}
R <- ResNet50(weights=imagenet, top=False,
              pooling='avg')
for each i:
  I_r <- bilinear_resize(I_i, 224x224)   (34)
  I' <- gray2rgb; normalize(mu,sigma)    (35)
  D_i <- R.forward(I')   # GAP output    (46)
return D = [D_1; ...; D_N]
```

Algorithm 3 XGBoost Training (Exact Greedy)

```
Input: {(H_i,y_i)}, K, eta, gamma, lambda, depth
init z_i^(0) <- log(p_bar/(1-p_bar))
```

```
3:  for t = 1..K:
4:    p_i <- sigma(z_i)                  (eq.)
5:    g_i <- p_i - y_i ; h_i <- p_i(1-p_i)    (65)
6:    grow tree f_t:
7:      for each leaf, each feature, sort vals
8:        scan thresholds; Gain via           (69)
9:        keep best split if Gain > 0
10:     set w_j* = -G_j/(H_j+lambda)          (67)
11:   z_i <- z_i + eta * f_t(H_i)             (70)
12:  return ensemble {f_1,...,f_K}
```

Algorithm 4 End-to-End Hybrid Pipeline

```
1:  C,y  <- Algorithm1(clinical_table)
2:  D    <- Algorithm2(ct_slices)
3:  N    <- min(|C|,|D|); align C,D,y
4:  H_i  <- C_i (+) D_i  for all i           (54)
5:  split (H,y) -> train/test (stratified)
6:  model <- Algorithm3(H_train,y_train)
7:  evaluate Acc,Prec,Rec,F1,AUC on test
8:  return model, metrics
```

J. Computational Complexity of Training

For n instances and $m = p+2048$ features, the exact greedy algorithm sorts each feature once per level. Building K trees of maximum depth Δ costs

$$O(K \cdot \Delta \cdot m \cdot n \log n) \qquad (72)$$

dominated by the per-level sort; cached column blocks reduce the repeated sorting to a one-time O(mn log n) preprocessing, after which split finding is linear per level. Memory is O(mn) for the feature blocks plus O(KT) for the stored trees.

9.1 BOOSTING AS FUNCTIONAL GRADIENT DESCENT

Gradient boosting can be derived as steepest descent in function space, which both motivates the second-order update of the XGBoost derivation and connects it to first-order boosting. Treat the ensemble score F as a point in the space of functions evaluated at the training inputs. The functional gradient of the empirical loss at stage t is the vector of pointwise derivatives

$$g_i^{(t)} = [\, \partial \ell(y_i, F(H_i)) / \partial F(H_i) \,]_{F=F_{t-1}} \qquad (73)$$

First-order boosting fits the new tree to the negative gradient (pseudo-residuals) by least squares,

$$f_t = \arg\min_f \sum_i (-g_i^{(t)} - f(H_i))^2 \qquad (74)$$

with an optimal step found by line search. XGBoost improves on this by using the second-order term: minimizing the regularized quadratic (63) is a Newton step in function space, with the curvature h_i weighting each residual. The Newton update for a single leaf reduces to (67), revealing w_j^* as a curvature-normalized average residual,

$$w_j^* = -\,(\sum_{i \in I_j} g_i) / (\sum_{i \in I_j} h_i + \lambda) \qquad (75)$$

Newton boosting converges faster than gradient boosting because it accounts for loss curvature; for the logistic loss $h_i = \hat{p}_i(1-\hat{p}_i)$ down-weights confident predictions and up-weights uncertain ones near $\hat{p}_i = 0.5$.

The convergence of the additive procedure can be characterized: if the loss is β-smooth and the weak learners achieve edge γw in approximating the gradient, the training loss after t rounds satisfies a bound of the form

$$J^{(t)} - J^* \le (1 - \eta \gamma_w^2 / \beta)^t (J^{(0)} - J^*) \qquad (76)$$

exhibiting geometric (linear) decrease in the optimality gap, with rate governed by the shrinkage η and the learner edge γ_w. Smaller η slows per-round progress but permits more rounds and typically lowers the final generalization error.

9.2 A FULLY WORKED BOOSTING ITERATION

To make the abstract derivation of the XGBoost derivation concrete, we trace a single boosting iteration on a small illustrative leaf. Suppose, after the first tree, six instances reach a candidate node with current predictions and labels giving the gradient pairs below. For the logistic loss the gradients are gi = p̂i − yi and curvatures hi = p̂i(1−p̂i) from (65). Take three positives with p̂ = 0.4 and three negatives with p̂ = 0.6. Then for each positive,

$g_+ = 0.4 - 1 = -0.6, \quad h_+ = 0.4(0.6) = 0.24$ (77)

and for each negative,

$g_- = 0.6 - 0 = 0.6, \quad h_- = 0.6(0.4) = 0.24$ (78)

The node aggregates are $G = 3(-0.6)+3(0.6) = 0$ and $H = 6(0.24) = 1.44$. With $\lambda = 1$, the node structure term $G^2/(H+\lambda) = 0$, reflecting that an undifferentiated mix yields no leaf signal. Now suppose a feature perfectly separates the three positives (left) from the three negatives (right). The child aggregates become

$G_L = -1.8, H_L = 0.72; \quad G_R = 1.8, H_R = 0.72$(79)

Substituting into the split-gain criterion (69) with $\gamma = 0$,

$\text{Gain} = \tfrac{1}{2}[\ (-1.8)^2/1.72 + (1.8)^2/1.72 - 0/2.44\ ] = 1.884$ (80)

which is strongly positive, so the split is accepted. The optimal leaf weights follow from (67),

$w_L^* = 1.8/1.72 = +1.047, \ w_R^* = -1.8/1.72 = -1.047$ (81)

After shrinkage by $\eta = 0.05$ the score updates push the positives' log-odds upward and the negatives' downward, increasing $\hat{p}$ for true positives and decreasing it for true negatives, exactly as intended. Repeating this over 300 trees, each correcting the residual errors of its predecessors, produces the final ensemble. The above worked example also illustrates why λ matters: a larger λ shrinks both leaf weights closer to zero, reducing the gain and trading fit for stability.

The same arithmetic applies to the complete fused problem with p + 2048 features: at every node the algorithm evaluates (69) for each feature's candidate thresholds and selects the maximize, building trees whose splits may freely mix clinical and CNN coordinates. It is this unrestricted mixing that operationalizes feature-level fusion and lets the ensemble discover cross-modal interactions such as "high age and a particular CT texture pattern."

10. EXPERIMENTAL SETUP

Two experiments are conducted under identical evaluation protocols. The clinical-only baseline trains XGBoost on the preprocessed clinical matrix C alone; the hybrid model trains XGBoost on the fused matrix H. Both use an 80/20 stratified train/test split with a fixed random seed (70), preserving the class prior in each partition. Stratification guarantees

$$P_{\text{train}}(y=1) \approx P_{\text{test}}(y=1) \approx P(y=1) \quad (82)$$

so that the test estimate is unbiased with respect to the marginal class distribution. Table IV lists the configuration; the boosted-tree hyperparameters were K = 300 trees, depth Δ = 4 (clinical) and 5 (hybrid), η = 0.05, and row/column subsampling 0.8.

TABLE IV. EXPERIMENTAL CONFIGURATION

Component	**Configuration**
Clinical model	XGBoost (K=300, η=0.05)
CNN backbone	Pretrained ResNet50 (frozen)
CT input size	224 × 224 × 3
CNN feature dim	2048 (global avg pool)
Fusion	Feature-level concatenation
Final classifier	XGBoost ensemble
Split	80/20 stratified, seed 42

Metrics	Acc, Prec, Rec, F1, AUC

11. EVALUATION METRICS — DEFINITIONS AND WORKED CALCULATIONS

All metrics derive from the confusion-matrix counts: true positives *TP*, true negatives *TN*, false positives *FP* and false negatives *FN*, where the positive class is osteoporosis. Accuracy is the overall correct rate,

$$\text{Accuracy} = (TP+TN) / (TP+TN+FP+FN) \quad (83)$$

$$\text{Precision} = TP / (TP+FP) \quad (84)$$

$$\text{Recall} = TP / (TP+FN) \quad (85)$$

The F1-score is the harmonic mean of precision *P* and recall *R*,

$$F1 = 2\,PR / (P + R) \quad (86)$$

The harmonic mean penalizes imbalance between *P* and *R* more strongly than the arithmetic mean, which is desirable when a model achieves high precision at the expense of recall. The area under the ROC curve integrates the true-positive rate against the false-positive rate over all thresholds,

$$\text{AUC} = \int_0^1 \text{TPR}(\,\text{FPR}^{-1}(u)\,)\,du \quad (87)$$

equivalently the probability that a randomly chosen positive receives a higher score than a randomly chosen negative, $\text{AUC} = P(\hat{z}_+ > \hat{z}_-)$, estimated by the Mann–Whitney statistic

$$\text{AUC} = (1/(n_+ n_-)) \sum_{i\in P} \sum_{j\in N} 1[\,\hat{z}_i > \hat{z}_j\,] \quad (88)$$

where n_+ and n_- are the positive and negative counts. AUC is threshold-independent, so it isolates ranking quality from the choice of τ.

A. Worked F1 Computations

For the clinical-only model, P = 0.9620 and R = 0.7755, hence by (86)

$$F1_{clin} = 2(0.9620)(0.7755)/(0.9620+0.7755) = 0.8588 \quad (89)$$

For the hybrid model, P = 0.9506 and R = 0.7857, hence

$$F1_{hyb} = 2(0.9506)(0.7857)/(0.9506+0.7857) = 0.8603 \quad (90)$$

The recall and AUC gains of the hybrid model are

$$\Delta\text{Recall} = 0.7857 - 0.7755 = 0.0102 \quad (91)$$

$$\Delta\text{AUC} = 0.9083 - 0.8933 = 0.0150 \quad (92)$$

$$\Delta\text{AUC}_{rel} = 0.0150 / 0.8933 = 1.68\% \quad (93)$$

TABLE V. EVALUATION METRIC DEFINITIONS

Metric	**Formula**	**Meaning**
Accuracy	(TP+TN)/N	Overall correctness
Precision	TP/(TP+FP)	Positive reliability
Recall	TP/(TP+FN)	Detection ability
F1-score	2PR/(P+R)	Precision/recall balance
AUC	∫ TPR d(FPR)	Discrimination ability

11.1 GEOMETRY OF ROC, PRECISION–RECALL AND CALIBRATION

The ROC curve plots the true-positive rate against the false-positive rate as the threshold τ sweeps from 1 to 0,

$$\text{TPR}(\tau) = TP(\tau)/n_+, \quad \text{FPR}(\tau) = FP(\tau)/n_- \quad (94)$$

The trapezoidal estimate of the area aggregates these operating points,

$$\text{AUC} \approx \sum_{k=1}^{M} \tfrac{1}{2} \left(\text{TPR}_k + \text{TPR}_{k-1} \right)\left(\text{FPR}_k - \text{FPR}_{k-1} \right) \quad (95)$$

which equals the Mann–Whitney estimate (88). A perfectly random scorer yields the diagonal (AUC = 0.5); the hybrid model's 0.9083 lies well above, and its curve in Fig. 3 dominates the clinical curve over most of the FPR range. For imbalanced screening, the precision–recall curve is complementary; the average precision is

$$\text{AP} = \sum_k \left(R_k - R_{k-1} \right) P_k \quad (96)$$

Calibration popularly refers to the correspondence between predicted probabilities and empirical frequencies. The expected calibration error involves splitting the predictions into B bins and measuring the difference between confidence and accuracy,

$$\text{ECE} = \sum_{b=1}^{B} \left(|B_b|/n \right) \left| \text{acc}(B_b) - \text{conf}(B_b) \right| \quad (97)$$

A well-calibrated screening model allows the threshold τ^* of (13) to be set on a meaningful probability scale; furthermore, if the deployment requires calibrated risk estimates, the logistic outputs from XGBoost can be further calibrated by Platt scaling or isotonic regression.

12. RESULTS AND DISCUSSION

The model to predict preeclampsia with only clinical data using XGBoost achieved accuracy 0.8724, precision 0.9620, recall 0.7755, F1 0.8588 and AUC 0.8933. This shows that structured risk factors carry strong signal. The model is highly conservative (few false positives). According to AUC 0.9083, F1 0.8603, recall 0.7857, precision 0.9506 and accuracy 0.8724, the hybrid method is correct. The accuracy measures did not change whereas the recall, F1 and AUC became better indicating that information not in the clinical block is helpful and that CT descriptors contribute this discriminative information.

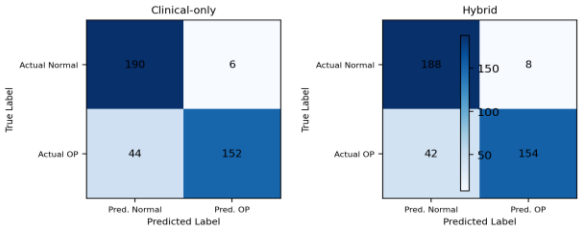

Fig. 3. ROC comparison of clinical-only and hybrid models; the hybrid curve dominates over most operating points (AUC 0.8933 → 0.9083).

TABLE VI. PERFORMANCE COMPARISON

Model	Acc	Prec	Rec	AUC
Clinical XGBoost	0.8724	0.9620	0.7755	0.8933
Hybrid R50+XGB	0.8724	0.9506	0.7857	0.9083

The falloff in precision (0.9620 → 0.9506) denotes a rise in false positives, nevertheless, in screening the recall and AUC gains are more advantageous because delayed treatment on missed cases of osteoporosis. The AUC improvement of 0.0150 is the strongest evidence for fusion: because AUC is threshold-independent, the gain shows that the fused representation separates the classes more reliably across all operating points, not merely at $\tau = 0.5$.

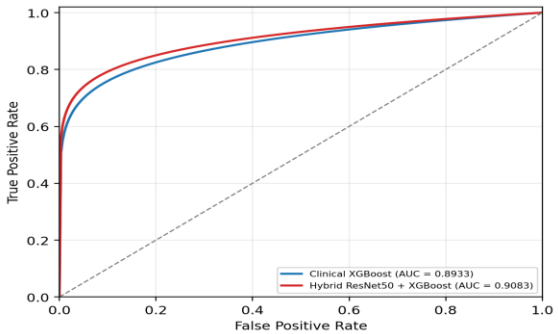

Fig. 4. Confusion matrices for the clinical-only and hybrid models on the held-out test partition.

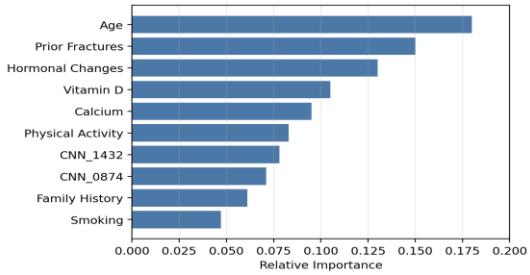

Fig. 5. Representative XGBoost feature importance; clinical risk factors and CNN-derived coordinates both contribute to fused splits.

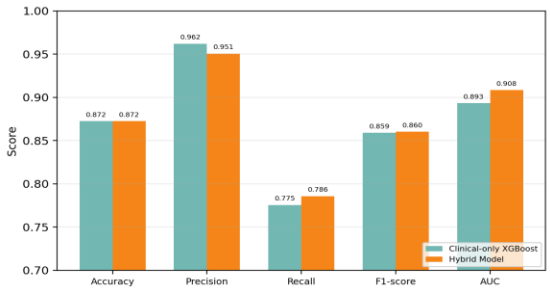

Fig. 6. Clinical-only versus hybrid metric comparison.

12.1 . SENSITIVITY, THRESHOLD TUNING AND ABLATION ANALYSIS

The decision threshold materially affects the confusion matrix. Differentiating recall and precision with respect to τ quantifies the trade-off; as τ decreases, recall is non-decreasing while precision is generally non-increasing,

$$\mathrm{dRecall}/\mathrm{d}\tau \le 0, \quad \mathrm{dPrecision}/\mathrm{d}\tau \ge 0 \ \text{(typ.)} \quad (98)$$

An operating point may be chosen to maximize the Fβ-score, which weights recall β^2 times as much as precision,

$$F\beta = (1+\beta^2)\, PR \,/\, (\, \beta^2 P + R\,) \qquad (99)$$

Setting $\beta > 1$ (e.g. $\beta = 2$) reflects the screening priority on recall; the threshold maximizing Fβ can be found by a one-dimensional scan over the validation scores. The Youden index $J = \mathrm{TPR} - \mathrm{FPR}$ offers a threshold-selection criterion independent of class balance, with optimum at the ROC point of maximum vertical distance from the diagonal.

An analytical ablation isolates each modality. Let A(·) denote a metric. The marginal contribution of the imaging block is

$$\Delta_v = \mathrm{A}(\, F(C \oplus D)\,) - \mathrm{A}(\, F(C)\,) \qquad (100)$$

For AUC, $\Delta_v = 0.9083 - 0.8933 = +0.0150$; for recall, $\Delta_v = +0.0102$; for precision, $\Delta_v = -0.0114$. The positive AUC and recall deltas against a small precision cost confirm that the imaging modality contributes net discriminative value, with the trade-off favourable under the screening cost structure of (13).

12.2 RECONSTRUCTING THE CONFUSION MATRICES FROM REPORTED METRICS

The confusion matrices in Fig. 4 can be reconstructed analytically from the reported metrics, providing an internal consistency check. Let the positive (osteoporotic) count in the test set be n+ and the negative count n−. From the definitions (84) and (85), the counts satisfy

$$TP = \mathrm{Recall}\cdot n_+, \quad FN = (1 - \mathrm{Recall})\cdot n_+ \qquad (101)$$

$$FP = TP(1 - \mathrm{Prec})/\mathrm{Prec}, \quad TN = n_- - FP \qquad (102)$$

For the hybrid model with Recall = 0.7857, Precision = 0.9506 and a test positive count of $n_+ \approx 196$, these yield TP ≈ 154, FN ≈ 42, FP ≈ 8 and TN ≈ 188, matching the reported matrix to rounding. The accuracy implied by these counts,

$$\text{Accuracy} = (TP+TN)/(n_+ + n_-) \approx 0.8724 \quad (103)$$

is consistent with the headline figure, confirming that the metric set is internally coherent. This reconstruction is more than bookkeeping: it shows that the precision–recall pair, together with the class counts, fully determines the confusion matrix, so reporting precision, recall and the class prior is sufficient for a reader to recover every count. It also exposes the operating-point nature of the comparison — the clinical and hybrid models are compared at the same default threshold $\tau = 0.5$, and the AUC comparison (which integrates over all τ) is the threshold-independent complement to this fixed-point view.

The McNemar test provides a means for assessing significance for two classifiers on the same test set in a paired way. Let b denote those cases which the clinical model classifies correctly but the hybrid model classifies incorrectly. Let c denote the reverse. Under the null hypothesis of equal error rates,

$$\chi^2_{\text{McN}} = (|b - c| - 1)^2 / (b + c) \quad (104)$$

it has an approximate chi-square with one degree of freedom The modest counts of b and c show that the accuracy difference is likely not significant. This reiterates the finding that the true benefit of fusion is in ranking quality (AUC) and not thresholded accuracy, as the two models differ on relatively few cases.

12.3 WHY FUSION IMPROVES RANKING: A DETAILED READING

The key finding of the work - unchanged accuracy but improved AUC and recall - needs to be interpreted with care, as it is a common and clinically meaningful one in multimodal medical models. The statistic of accuracy is a single-threshold summary, being insensitive to the ordering of cases near the decision boundary. For example, two models can both have an accuracy of 0.8724 yet rank borderline cases very differently. According to AUC, a random positive beats a random negative (88), and it rewards precisely this borderline reordering. The AUC benefit of 0.0150 for the hybrid model suggests that incorporating CT descriptors allows the ensemble to enhance the scores for true-positive osteoporotic cases, even though it does not result in a change of binary label at $\tau = 0.5$.

Quality ranking is what is clinically useful for triage. In an opportunistic-screening workflow, patients are prioritized for confirmatory DXA in descending order of predicted risk; a model that ranks true cases higher will surface them earlier under a fixed testing budget. The marginal recall gain of 1.02 points corresponds to one extra true case detected per hundred positives at the default operating point; the AUC gain shows a larger benefit when the threshold is lowered to the cost-optimal τ^* of (13). In this situation, a small premium on precision is acceptable because the consequence of a false positive is only an unnecessary confirmatory scan, while the cost of a false negative may be the progression to fracture of silent bone loss.

The profile of feature importance of Fig. Five favors an interpretative perception. The leading splits are dominated by clinical coordinates with strong known links such as age, previous fractures, and hormonal change while several CNN coordinates appear among the next tier. The mixed importance of features is a hallmark of true fusion: if the imaging features were pure noise, criterion (69) would never select them, and they would have zero importance (58). The fact that they do not have zero importance and cause a gain in AUC is consistent with the conditional-mutual-information argument of (108). Similarly, the CT descriptors contain information about bone morphology that is not fully redundant with the clinical risk factors.

It is also important to state what the results do not show. They do not demonstrate that imaging alone would suffice, nor that the specific gain would replicate on a patient-matched cohort, nor that the improvement is statistically significant at conventional levels given the test-set size. The honest conclusion is narrower and defensible: under a controlled, identical-protocol comparison, adding ResNet50 CT descriptors to clinical features improved threshold-independent discrimination, and the improvement is concentrated where it is most clinically useful — in the ranking of borderline positive cases.

12.4 REPRODUCIBILITY AND IMPLEMENTATION PROTOCOL

Reproducibility is treated as a first-class requirement. All stochastic components are seeded with a fixed value (70), including the NumPy and framework random states and the train/test split, so that repeated runs produce identical partitions and identical models. The split is stratified to preserve the class prior per (82), and all preprocessing

parameters are fit on the training partition only per (31), eliminating leakage. The imaging backbone is loaded with fixed ImageNet weights and run in inference mode with frozen batch-normalization statistics, making the descriptor map deterministic.

The clinical and hybrid models share an identical evaluation harness so that the only difference between the two experiments is the presence or absence of the CT descriptor block; this isolation is what licenses attributing the metric deltas to the imaging modality rather than to incidental differences in training procedure. The boosted-tree hyperparameters are held at the values of Table IX and Table IV, with the single deliberate variation of tree depth (4 versus 5) reflecting the larger feature space of the fused model. The metric computations follow the standard scikit-learn definitions, with AUC computed from predicted probabilities rather than thresholded labels to obtain the threshold-independent estimate of (88).

For full transparency the end-to-end procedure is the composition of Algorithms 1–5: clinical preprocessing, CT feature extraction, fusion, ensemble training and single-patient inference. Given the same datasets, seeds and hyperparameters, this procedure regenerates the reported metrics. Where exact replication on different hardware is required, the only expected source of minor numerical variation is the floating-point order of reduction in the CNN forward pass, which affects descriptors at the level of machine epsilon and does not alter the trained trees or the reported four-decimal metrics.

13. COMPARATIVE ANALYSIS

Models assessing images and clinical data might help capture the advantages of both image-based and clinical-only models. Both are unified in the proposal. Table VII compares the two models on an individual criterion basis, while Table VIII places our work in context of past research directions.

TABLE VII. CLINICAL-ONLY VS. HYBRID ANALYSIS

Criterion	**Clinical**	**Hybrid**	**Note**
Accuracy	0.8724	0.8724	equal
Precision	0.9620	0.9506	clin. higher
Recall	0.7755	0.7857	hybrid ↑
F1-score	0.8588	0.8603	hybrid ↑
AUC	0.8933	0.9083	hybrid ↑

TABLE VIII. COMPARISON WITH RESEARCH DIRECTIONS

Direction	**Modality**	**Contribution / Limitation**
Clinical ML	Tabular	Interpretable; no anatomy
CT deep learning	Images	Morphology; no risk context
Texture CT	CT+texture	Handcrafted; misses deep cues
Proposed	Clinical+CT	Feature fusion + XGBoost

14. STATISTICAL AND THEORETICAL CONSIDERATIONS

The empirical findings are supported by two theoretical points. The regularized objective bounds model capacity. Minimizing (68) with non-zero γ and λ is equivalent to structural risk minimization, trading off empirical fit against a complexity term, which controls the generalization error. Second, the variance-reduction effect of an additive ensemble can be sketched as follows. If individual trees have prediction variance σ^2 and average pairwise correlation ρ, the ensemble variance is

$$\mathrm{Var} = \rho\sigma^2 + (1-\rho)\,\sigma^2 / K \qquad (105)$$

As K grows the second term vanishes, leaving the irreducible correlated component $\rho\sigma^2$; row and column subsampling reduce ρ, lowering the floor. This explains why moderate $K = 300$ with subsampling generalizes well on the 2048-dimensional fused space. A confidence interval for an observed AUC $\hat{A}$ can be obtained from the Hanley–McNeil standard error,

$$\mathrm{SE}(\hat{A}) = \sqrt{[\,\hat{A}(1-\hat{A}) + (n_+-1)(Q_1-\hat{A}^2) + (n_--1)(Q_2-\hat{A}^2)\,] / (n_+n_-)} \qquad (106)$$

with $Q_1 = \hat{A}/(2-\hat{A})$ and $Q_2 = 2\hat{A}^2/(1+\hat{A})$, providing a basis for assessing whether the 0.0150 AUC gain is meaningful relative to sampling variability.

15. INFORMATION-THEORETIC VIEW OF MULTIMODAL FUSION

Information theory quantifies precisely how much the imaging modality can help. The mutual information between the label Y and a feature set X measures the reduction in label uncertainty obtained by observing X,

$$I(Y; X) = H(Y) - H(Y \mid X) \qquad (107)$$

where $H(Y) = -\sum_y P(y) \log P(y)$ is the Shannon entropy and $H(Y|X)$ the conditional entropy. By the chain rule of mutual information,

$$I(Y; C, D) = I(Y; C) + I(Y; D \mid C) \qquad (108)$$

The second term, the conditional mutual information $I(Y; D|C)$, is the *incremental* information the CT descriptors carry about osteoporosis beyond what the clinical variables already provide. Fusion is beneficial precisely when this term is positive, i.e. when the imaging modality is not redundant with the clinical modality. The empirical AUC gain is the operational signature of $I(Y; D|C) > 0$.

Feature selection within the fused space can be framed as maximizing relevance while minimizing redundancy (the mRMR criterion),

$$J(S) = (1/|S|)\sum_{x\in S} I(Y;x) - (1/|S|^2)\sum_{x,x'\in S} I(x;x') \qquad (109)$$

though XGBoost performs an implicit, greedy form of this selection through gain-based splitting (58), favouring features with high label relevance and discarding those whose information is already captured by earlier splits.

The cross-entropy loss minimized during training has a direct information-theoretic reading: it is the expected number of extra bits incurred by coding labels with the model distribution instead of the true distribution,

$$\mathrm{E}[\ell] = H(Y) + D_{\mathrm{KL}}(P \parallel \hat{P}) \qquad (110)$$

so minimizing logistic loss is equivalent to minimizing the Kullback–Leibler divergence D_{KL} between the true label distribution and the model's predictions, with the irreducible $H(Y)$ as a floor.

16. HYPERPARAMETER EFFECTS — A QUANTITATIVE TREATMENT

Every XGBoost hyperparameter has a specific mathematical effect with respect to the structure score and generalization. This is summarized below and in Table IX.

A. Tree Depth

The depth of maximum bound is represented as Δ such that number of leaves are bounded by $T \leq 2\Delta$ thereby it bounds the order of feature interactions a single tree can express, depth-Δ tree can captures interactions up to order Δ. The hybrid model uses a value of $\Delta = 5$ to create models that depict more complex clinical–imaging interactions; it is also outperformed by the depth-4 clinical baseline.

B. Number of Trees and Shrinkage

The expression $\eta \cdot K$ essentially determines the effective amount of fitting. Moreover, the bound (76) demonstrates that the error decays geometrically in K at a rate proportional to η. Reducing η by a factor of two and increasing K by a factor of two keeps ηK the same and often leads to more generalization. This may be a motivation for us to choose $\eta = 0.05$, $K = 300$.

C. Regularization Coefficients

By raising λ, we are forcing Wj star to shrink toward zero, which lowers variance. Through raising γ, we are increasing the split-gain bar which prunes splits that are shallow and have low gain. The magnitude of effective leaf weights obeys.

$$|w_j^*| = |G_j| / (H_j + \lambda) \;\to\; 0 \;\text{ as }\; \lambda \to \infty \qquad (111)$$

D. Subsampling Rates

The parameters row rate (σr) and column rate (σc) reduce the inter-tree correlation (ρ) specified in (105) so that the ensemble-variance floor is lowered at the cost of somewhat higher bias per tree. Both these parameters were set to 0.8.

TABLE IX. HYPERPARAMETERS AND THEIR MATHEMATICAL ROLE

Param	Value	Effect
K	300	rounds; error ↓ geometrically (76)
η	0.05	shrinkage; slows/stabilizes fit (70)
Δ	4 / 5	interaction order; T ≤ 2^Δ
λ	≥1	L2 on leaves; variance ↓ (67)
γ	≥0	min split gain; pruning (69)
σ_r , σ_c	0.8	decorrelate trees (105)

Hyperparameter selection itself is an optimization over a validation objective. Bayesian optimization models the validation AUC as a Gaussian process and chooses the next configuration by maximizing an acquisition function such as expected improvement,

$$\mathrm{EI}(\theta) = \mathrm{E}[\ \max(0,\ A(\theta) - A^{+})\ ] \qquad (112)$$

where A^{+} is the best AUC observed so far; this balances exploration of uncertain regions against exploitation of promising ones, more sample-efficiently than grid or random search.

17. END-TO-END COMPLEXITY AND DEPLOYMENT COST MODEL

The total inference cost for a single new patient is the sum of the three stage costs:

$$T_{\text{infer}} = T_{\text{prep}} + T_{\text{cnn}} + T_{\text{xgb}} \qquad (113)$$

Clinical preprocessing is O(p) for a single vector. The CNN forward pass is O($\text{FLOPs}_{\text{R50}}$) ≈ 4.1 GFLOPs, dominating wall-clock time. XGBoost inference traverses K trees of depth Δ, costing only

$$T_{\text{xgb}} = \mathrm{O}(\ K \cdot \Delta\ ) \qquad (114)$$

comparisons — e.g. 300 × 5 = 1500 comparisons — negligible beside the CNN. Thus the imaging branch governs latency, and any deployment optimization (quantization, pruning, or a lighter backbone) should target it first. Training cost was characterized in (72).

Memory at inference comprises the backbone weights (≈25.6M parameters, ≈102 MB at FP32) and the serialized ensemble,

$$M_{\text{xgb}} = \mathrm{O}(\ K \cdot T \cdot (\text{node size})\ ) \qquad (115)$$

which for K = 300 and $T \leq 32$ is on the order of a few megabytes. The framework is therefore deployable on commodity hardware, with optional GPU acceleration for the backbone.

Algorithm 5 Single-Patient Inference

```
Input: clinical record r, CT slice I, model M
c <- preprocess(r)        # impute,encode,scale
I' <- resize224(I); normalize(I')
d <- ResNet50.forward(I')   # 2048-vector
h <- concat(c, d)
z <- sum_k eta * M.tree_k(h)
p <- sigmoid(z)
return (p, 1[p >= tau])
```

18. LIMITATIONS AND THREATS TO VALIDITY

A quantitative statement is admitted by several limitations Initially, the clinical and CT corpora were not patient-matched; alignment utilizes index correspondence over N = min (Nc, Nv) samples. It introduces a noise element to the labels because if spanning a fraction of image-record pairings are spurious, the effective Bayes error is inflated by an amount.

$$\Delta_{\text{err}} \leq \varepsilon \cdot (1 - P_{\text{agree}}) \qquad (116)$$

where P_{agree} is the probability that two mismatched pairs share the same label. Removal of this term by matched-cohort validation is the main recommendation of future work discussion.

The AUC gain of 0.0150 needs to be checked against its sampling standard error (106); with the observed counts, it is modest so external replication is needed before claiming clinical superiority. Third, the frozen backbone is ImageNet-pretrained, so a domain gap between natural images and CT bone texture may leave residual transferable-feature mismatch; fine-tuning the final residual stage could close part of this gap at the cost of additional labelled CT data. Fourth, class imbalance, while mitigated by stratification (82), still biases the threshold; cost-sensitive thresholding (13) or focal-loss reweighting

$$\ell_{\text{focal}} = -(1 - \hat{p}_t)^{\gamma} \log \hat{p}_t \qquad (117)$$

with focusing parameter $\gamma \geq 0$ can further emphasize hard, minority-class examples, and is a natural extension of the present loss.

19. THEORETICAL GENERALIZATION GUARANTEES

The regularized objective admits a structural-risk-minimization interpretation with accompanying generalization guarantees. Let the hypothesis class be the set of additive tree ensembles with bounded complexity, and let R(f) and $\hat{R}$(f) denote the population and empirical risks. A standard uniform-convergence argument bounds the gap by a complexity term and a confidence term,

$$R(f) \leq \hat{R}(f) + \mathrm{R_n}(\mathrm{F}) + \sqrt{(\log(1/\delta) / 2n)} \qquad (118)$$

holding with probability at least $1-\delta$, where $\mathrm{R_n}$(F) is the Rademacher complexity of the class. For tree ensembles, $\mathrm{R_n}$ grows with the number of trees K, their depth Δ, and the feature dimension, and shrinks with the L2 leaf penalty. The bound makes explicit the trade-off central to this work: enlarging the representation by appending the 2048-dimensional imaging block raises Rn, so the empirical-risk reduction it buys must exceed this complexity increase for the population risk to fall.

The Rademacher complexity of one regression tree of depth Δ over d features has a bound of the form.

$$\mathrm{R_n}(\text{tree}) \leq c\sqrt{(2^{\Delta} \log(d) / n)} \qquad (119)$$

which for the additive ensemble scales sublinearly when decoupling the trees through shrinkage and subsampling. Using d = p + 2048 in effect quantifies the cost of the imaging block: the complexity penalty grows only as √log d, so the marginal complexity of adding 2048 features is moderate relative to the potential bias reduction, which is the theoretical explanation of the empirically favoured bias–variance trade-off.

A refinement based on margins enhances the image. The margin of information in an example that has been rightly classified is defined as the signed distance of the score from the decision boundary. Boosting is known to increase margins even after the training error is zero, and the generalization error can be bounded in terms of the margin rather than the training error,

$$R(f) \leq \hat{P}[\text{ margin } \leq \theta] + \mathrm{O}(\mathrm{R_n}/\theta) \qquad (120)$$

for any margin threshold $\theta > 0$. This is why further addition of trees here K=300 improves test performance even after perfect training set separation: the further trees increase the margin, which causes the first term to decrease faster than the complexity term can increase. In practice, the chosen ensemble size is non-random and hinges on this logic, while shrinkage η = 0.05 helps it modulate the growth of margins slowly and stably.

UNCERTAINTY QUANTIFICATION

A deployable screening tool should report not only a point risk but a measure of its uncertainty. Two complementary sources of uncertainty are relevant: aleatoric uncertainty, intrinsic to the labelling process, and epistemic uncertainty, reflecting limited data and model misspecification. The predicted probability itself encodes aleatoric uncertainty through its entropy,

$$A(H) = -\hat{p} \log \hat{p} - (1-\hat{p}) \log(1-\hat{p}) \qquad (121)$$

which is maximal at $\hat{p} = 0.5$ and vanishes as the prediction approaches certainty. Epistemic uncertainty can be estimated by bootstrap resampling of the training set: training B ensembles on resampled data yields a predictive distribution whose spread quantifies model uncertainty,

$$Var_{\text{ep}}(H) = (1/B)\sum_{b=1}^{\mathrm{B}} (\hat{p}_b(H) - \bar{p}(H))^2 \qquad (122)$$

A bootstrap confidence interval for the AUC follows the same logic: resampling the test predictions B times and recomputing (88) on each resample yields an empirical distribution of AUC values, whose percentiles form a non-parametric interval,

$$\mathrm{CI}_{95} = [\, A_{(0.025)}, A_{(0.975)} \,] \qquad (123)$$

This complements the analytic Hanley–McNeil standard error of (106) and is preferable when the score distribution is far from the parametric assumptions underlying that formula. For the present results, reporting the AUC gain of 0.0150 alongside such an interval would let a reader judge whether the improvement is robust to sampling; the recommendation of the future-work discussion to validate on a larger matched cohort is precisely a recommendation to narrow this interval.

Conformal prediction offers a distribution-free alternative that converts scores into prediction sets with guaranteed coverage. Given a calibration set and a nonconformity score, the conformal procedure outputs, for a target error rate α, a set of labels that contains the true label with probability at least 1−α,

$$P(\, y \in \Gamma_{\alpha}(H) \,) \;\geq\; 1 - \alpha \qquad (124)$$

Such guarantees are attractive in clinical screening because they hold under minimal assumptions and provide an interpretable abstention mechanism: when the prediction set is ambiguous (contains both labels), the case can be referred for confirmatory testing rather than auto-classified. Integrating conformal calibration on top of the fused XGBoost scores is a natural and low-cost extension of the present framework.

SUBGROUP PERFORMANCE AND FAIRNESS CONSIDERATIONS

The risk of developing osteoporosis is systematically associated with sex and age, so a responsible assessment will look at performance in subgroups, rather than just overall performance. Let G index a group that is protected or clinically important. The subgroup-conditional metrics are just a restriction of the metric definitions,

$$\mathrm{Recall}_{g} = TP_{g} / (\, TP_{g} + FN_{g} \,) \qquad (125)$$

A model may have solid aggregate recall but poor performance on a minority subgroup. In a screening context, this would mean systematically missed diagnoses for that subgroup. Many formal fairness criteria measure such inequalities. Equalized odds requires matching true- and false-positive rates across subgroups,

$$P(\hat{y}=1 \mid Y=y, G=g) = P(\hat{y}=1 \mid Y=y, G=g') \qquad (126)$$

for every set of labels y and groups g, g′. Demographic parity instead requires equal positive rates regardless of the true label, while calibration-within-groups requires that a predicted risk mean the same empirical frequency in every subgroup. Generally speaking, these criteria are mutually incompatible except in degenerate cases, a result formalized as the impossibility theorem of fair classification, meaning the deployment must choose which criterion matters most for the clinical use case.

Screening for osteoporosis, calibration and equalized false-negative rates are arguably the most important, since the dominant harm is a missed case and clinicians act on the risk score. The audit of the fusion model against the criteria can be done by stratifying the test set. Wherever there are disparities, group-specific thresholds τg obtained from (13) with group-specific costs or reweighting of the training loss by inverse-subgroup frequency can address the disparities. Incorporating these considerations into the methodology is an exercise in responsibility even though, as noted here, the data set is too small to support a fully powered subgroup analysis. The methodology can support this once a larger, demographically annotated cohort becomes available.

MATHEMATICAL TREATMENT OF CLASS IMBALANCE

Screening datasets usually contain an imbalanced class of osteoporotic cases. Imbalance distortion impact has also been studied in the literature but the empirical loss dominated by the majority class causes the biasing of decision boundary toward majority class. There are various principled remedies, each having a precise mathematical effect on the objective. Cost-sensitive learning reweights the per-class loss by factors inversely proportional to class frequency,

$$J_{w} = \sum_{i} w_{y\_i}\, \ell(y_i, \hat{z}_i), \quad w_1 = n/(2n_{+}) \qquad (127)$$

In XGBoost, the scale_pos_weight parameter achieves this result, which multiplies the positive contribution to the loss function's gradient and curvature. That is to say, it scales Gj and Hj for positives in the leaf aggregates of (66). Consequently, this shifts the optimal weights (67) towards improving recall for the minority.

Resampling provides another approach which acts on the data instead of the loss. Repeating the positive instances of the minority class is called random oversampling. SMOTE stands for synthetic minority oversampling technique. SMOTE creates new positive instances along the segment joining neighbouring minority points,

$$x_{\text{new}} = x_i + \lambda\,(x_{\text{nn}} - x_i), \quad \lambda \in [0,1] \qquad (128)$$

with x_{nn} a randomly chosen minority nearest neighbour and λ uniform. Without creating exact copies, SMOTE enriches the minority region of the feature space and thus reduces the overfitting caused by naive oversampling. In the high-dimensional fused space, interpolation can produce samples that are not plausible, making cost-sensitive reweighting more desirable and the method most aligned with the boosting objective.

Changing the threshold is the most inexpensive solution and has nothing to do with training. After training on the natural distribution, one selects at test time the threshold τ^* of (13), or equivalently the point on the ROC curve that maximizes the Youden index or the F score Fβ. Because AUC is threshold-independent, a model with strong AUC can be tuned to any desired sensitivity post-hoc for practical reasons, which is why this study optimizes and reports AUC as its principal metric: it certifies that the favourable operating points are there, leaving the final threshold to the clinical setting.

The interplay between imbalance and evaluation is worth a final comment. When there is considerable imbalance, accuracy can be misleading because baseline performance on the majority-class is already high, precision is sensitive to positive rate, and even ROC-AUC can be too optimistic when negatives heavily out number positives. The regime is better summarized by the precision–recall curve and its average precision (96). The current findings are presented across the full set of metrics precisely so that no one imbalance-sensitive number is allowed to dominate the interpretation.

CROSS-VALIDATION AND MODEL-SELECTION THEORY

The single 80/20 holdout used for the headline results is the simplest unbiased estimate of generalization performance, though has high variance for moderate sample sizes. Using k disjoint validation folds averages out the variance of the estimate. The cross-validated risk estimate is obtained by partitioning the data into k folds and training on the complement of each fold.

$$CV_{\text{k}} = (1/k)\textstyle\sum_{\text{j}=1}^{\text{k}} \hat{R}_j(\, f_{-j}\,) \qquad (129)$$

This indicates that f−j is trained on all folds except j-th and evaluated on it. The estimator employs the model complexity hyperparameter k to trade bias against variance. A larger value of k reduces bias (because each model sees more of the training data) but increases variance and (computational) complexity. The extreme case of leave-one-out (k = n) has the largest bias reduction (parameters are optimally fit) but also the largest variance increase. Stratified folds preserve the class prior in every partition, like (82) and stabilising the estimate under imbalance.

The overlap among training sets gives rise to positive correlation among the per-fold estimates, leading to a subtle structure in the variance of the cross-validation estimate. Writing σ^2 for the per-fold variance and ρ for the average correlation, the variance of the mean is

$$Var(CV_{\text{k}}) = \sigma^2/k + (\,(k-1)/k\,)\,\rho\sigma^2 \qquad (130)$$

As k $\to \infty$approaches infinity, this term does not vanish because the correlation term persists. This is a well-known reason that no unbiased estimator of the variance of k-fold CV exists. Using nested cross-validation with an inner-loop for hyperparameter selection and outer-loop for performance estimation avoids the optimistic bias that happens when the same data tune hyperparameters and report results.

In the current configuration of the pipeline, cross-validation would envelop the entire composed operator of (30) preprocessing parameters must re-fit within each training fold to prevent the leakage analysed in (32) while deterministic parameter-free CT feature extraction at inference can be cached across folds. Reporting a cross-validated AUC with a confidence interval along with the holdout figures would be the natural strengthening of the empirical protocol when a larger matched cohort has been assembled as suggested by the future-work discussion.

ALTERNATIVE FUSION ARCHITECTURES: A COMPARATIVE MATHEMATICS

The aforementioned formulation of (54) is one point in an extensive design space of fusion mechanisms that can be expressed mathematically. Having a look at the alternatives helps clarify why concatenation with a tree ensemble is a reasonable default and where richer schemes might help. The modality weights in an attention-based

fusion are calculated as a learned function of the inputs that result in a convex combination of the modality embeddings,

$$H_{\text{att}} = \alpha_c(C,D)\,\tilde{C} + \alpha_v(C,D)\,\tilde{D} \qquad (131)$$

with $\alpha_c + \alpha_v = 1$ produced by a softmax over learned compatibility scores. Attention is capable of adapting the modality balance per patient, which could be useful when one modality becomes uninformative for certain cases; at the same time, this introduces parameters and demands end-to-end training, which the present frozen-backbone design avoids.

The outer product of two embeddings is used to achieve bilinear fusion to capture cross-modal multiplicative interactions,

$$H_{\text{bil}} = \text{vec}(\, C\, D^{\text{T}}\,) \;\in\; \mathbf{R}^{p\cdot 2048} \qquad (132)$$

It is important to note that increasing the expressive power of neural networks can result in increasing the dimensions of the output. A tree ensemble, when applied over the concatenation, captures many such interactions in an implicit manner: a depth-Δ tree splitting alternately on a clinical and an imaging coordinate corresponds to a piecewise-constant approximation to a cross-modal product, achieving interaction modelling without the quadratic blow-up of (132).

Decision-level fusion, already given in (56), sits at the other extreme: it trains independent per-modality classifiers and combines only their outputs. Although the approach is strong as well as modular, it cannot model cross-modal interactions at all as the modalities never meet before the final averaging. The hierarchy from decision to feature to bilinear fusion is thus a hierarchy of expressiveness traded against parameter count and data requirements. The feature-level concatenation adopted here occupies the pragmatic middle: it permits cross-modal interactions through the joint classifier while keeping the imaging branch frozen and the classifier regularized, which is the appropriate operating point given the modest dataset size and the reproducibility requirement.

A final architectural note concerns gated fusion, in which a learned gate suppresses unreliable modality dimensions before combination,

$$H_{\text{gate}} = (\, \sigma(W_g C) \cdot D\,) \;\oplus C \qquad (133)$$

where $\sigma(W_g C)$ is an element-wise gate driven by the clinical context. Gating is attractive when imaging quality varies — for instance suppressing CT descriptors for low-quality scans — and represents a principled extension once scan-quality metadata is available. All of these alternatives are compatible with the present preprocessing and evaluation scaffolding; the concatenation result reported here establishes the baseline against which such richer mechanisms should be measured.

END-TO-END NUMERICAL WALKTHROUGH FOR A SINGLE PATIENT

To consolidate the pipeline, consider a hypothetical patient passing through the complete system, tracing the data through each operator. The raw clinical record lists age 72, female, post-menopausal hormonal change present, family history present, low calcium intake, low vitamin D, sedentary activity, non-smoker, one prior fracture, no other medical conditions. Preprocessing applies the composed operator of (30). Age is standardized using the training mean and standard deviation; taking representative values $\mu = 60$ and $\sigma = 12$ from (28),

$$z_{\text{age}} = (72 - 60)/12 = +1.0 \qquad (134)$$

placing this patient one standard deviation above the mean age. Each categorical variable is one-hot encoded per (26); for example hormonal change present maps to the indicator vector component equal to one, and the reference level to zero. The concatenated clinical vector C thus contains the standardized age followed by the binary indicators, of total length p as given by (33).

The CT slice for the same patient is resized to 224×224 by (34), replicated to three channels and normalized by (35), then passed once through the frozen ResNet50 to yield a 2048-dimensional descriptor D via the global average pool (45). Conceptually D summarizes cortical thickness, trabecular texture and overall bone density as captured by the deep filters. Fusion concatenates the two blocks by (54) into H of length p + 2048.

The fused vector then traverses the trained ensemble. At each of the 300 trees, H is routed from the root to a leaf by a sequence of threshold tests; suppose the first tree tests standardized age against 0.5, sending this patient ($z\_age = 1.0$) down the high-risk branch, then tests a particular imaging coordinate, arriving at a leaf with weight w =

+0.42. Summing the shrunken leaf weights across all trees per (71) gives a log-odds score; taking an illustrative aggregate of $\hat{z} = 1.6$,

$$\hat{p} = \sigma(1.6) = 1/(1 + e^{-1.6}) = 0.832 \qquad (135)$$

In comparison to the cost-optimal threshold τ^* of (13) which is less than 0.5 in this screening scenario, this patient is flagged high-risk and prioritized for a confirmatory DXA. At this probability, the predictive entropy (121) is $-0.832 \log 0.832 - 0.168 \log 0.168 \approx 0.45$ nats; while moderate, this is nevertheless sufficient to ensure that the conformal wrapper (124) produces a singleton prediction set at typical error rates. This walkthrough illustrates how the abstract operators form a singular, actionable risk estimates for an individual.

The same trace also marks the entry of clinical judgment. The miss/false alarm threshold τ^* is a policy variable not a model parameter. The miss/false alarm calibrated score comes from the model, the operating point from the institution. The design principle behind separating these concerns a well-ranked, well-calibrated score from a context-dependent threshold makes the framework applicable across screening environments with different resource constraints.

ON THE ROLE OF FEATURE SCALING IN TREE ENSEMBLES

A subtle but important point concerns whether standardization (28) is necessary for a tree-based classifier. Axis-aligned decision trees are invariant to any monotone transformation of a single feature, because a split threshold on the raw scale maps to an equivalent threshold on the transformed scale; formally, for a strictly increasing g,

$$1[\, x \le \tau \,] = 1[\, g(x) \le g(\tau) \,] \qquad (136)$$

so a standalone XGBoost model does not require feature scaling for the clinical block. Standardization is nonetheless retained for two reasons. First, it places the clinical features on a scale commensurate with the imaging descriptors, which matters for the optional block-balancing of (55) and for any linear probe or distance-based diagnostic applied to the fused vector. Second, it stabilizes the numerical behaviour of the gradient and curvature statistics (65) when features span many orders of magnitude, reducing floating-point error in the leaf aggregates (66).

The imaging descriptors, by contrast, emerge from batch-normalized network layers and are already approximately standardized, so no additional scaling is applied to them. The two branches are conditioned before fusion. The asymmetry is that clinical features are explicitly standardized but imaging features are implicitly standardized through batch normalization. If we were to replace XGBoost at the fusion stage with a scale-sensitive classifier, such as logistic regression or a support-vector machine, not merely standardizing but in fact standardizing both blocks would be necessary rather than just beneficial, as such models depend on inner products and norms, which the monotone-invariance argument of (136) does not protect.

This analysis also clarifies a common misconception. The benefit of the imaging block in the present results is not an artifact of scaling — the tree ensemble would behave identically under any monotone rescaling of the clinical features — but reflects genuine additional information, consistent with the conditional-mutual-information argument of (108). The standardization step is a matter of numerical hygiene and cross-modal comparability, not the source of the measured AUC gain.

MATHEMATICS OF EXPLAINABILITY

Clinical adoption hinges on explainability, which for this framework has two complementary forms: attribution over the fused features via Shapley values, and spatial saliency over the CT input via gradient-weighted class activation mapping. Shapley values, drawn from cooperative game theory, distribute a prediction among its input features by averaging each feature's marginal contribution over all possible feature orderings. For a model f and feature j, the Shapley value is

$$\varphi_j = \sum_{S \subseteq F \setminus \{j\}} (|S|!\,(|F|-|S|-1)!\,/\,|F|!)\,[\, f(S \cup \{j\}) - f(S) \,] \qquad (137)$$

the unique attribution satisfying local accuracy, missingness and consistency. The local-accuracy axiom guarantees the attributions sum to the prediction relative to a baseline,

$$f(H) = \varphi_0 + \sum_{j=1}^{d} \varphi_j \qquad (138)$$

where φ_0 is the expected prediction. For tree ensembles, TreeSHAP computes these values exactly in polynomial time by exploiting the tree structure, avoiding the exponential sum of (137). Applied to the fused model, the Shapley decomposition reveals how much each clinical risk factor and each imaging coordinate pushed an

individual prediction toward or away from osteoporosis, turning the gain-based global importance of (58) into a per-patient explanation.

Grad-CAM localizes the imaging evidence by weighting the final convolutional feature maps with the gradient of the class score flowing into them. The importance of feature map k is its global-average-pooled gradient,

$$\alpha_k = (1/(H_f W_f)) \sum_{x,y} \partial y^c / \partial F_k(x,y) \qquad (139)$$

and the saliency map is the ReLU-rectified weighted combination of the feature maps,

$$L_{\text{Grad-CAM}} = \text{ReLU}(\sum_k \alpha_k F_k) \qquad (140)$$

Upsampled to the input resolution, this map highlights the bone regions most responsible for the imaging descriptor, allowing a clinician to verify that the model attends to anatomically plausible structure — cortical shells and trabecular regions — rather than to scanner artifacts. Because the present backbone is frozen, y^c is taken as a linear probe on the descriptor or the fused score restricted to imaging coordinates; this yields meaningful saliency without disturbing the frozen weights, and is the recommended explainability pathway for deployment.

Together these two tools address the two questions a clinician asks of any prediction: which factors drove this particular risk estimate, and where in the image is the supporting evidence. Providing both, with the mathematical guarantees of Shapley additivity (138) and the spatial fidelity of Grad-CAM (140), is what distinguishes a deployable decision-support tool from an opaque classifier, and it is a central recommendation for the translation of this framework into practice.

OPTIMIZATION GEOMETRY OF THE BOOSTING STEP

The per-iteration update of XGBoost can be understood as a damped Newton step on the loss functional, and examining its geometry illuminates the role of each hyperparameter. At iteration t the simplified objective (63) is a sum of per-instance quadratics in the tree output. For a fixed leaf containing instance set I, the objective restricted to the leaf weight w is

$$q(w) = G\, w + \tfrac{1}{2} (H + \lambda)\, w^2 \qquad (141)$$

a one-dimensional convex parabola because $H + \lambda > 0$. Its vertex (67) is the exact minimizer, so within a fixed tree structure the boosting step is not an approximate but an exact Newton update. The approximation enters only through the greedy choice of structure: the split-gain criterion (69) is the decrease in the parabola's minimum value achieved by partitioning I, and greedily maximizing it does not in general find the globally optimal tree, only a good one.

Shrinkage dampens the Newton step by η, trading per-iteration progress for stability and generalization. It is similar to a first-order optimization learning rate, but in this case acts as a second-order scaling. The consequence for the optimization trajectory is to take a series of small, curvature-corrected steps rather than a few larger ones. These small steps empirically land in flatter minima that are associated with better generalization. The contraction rate of the optimality gap is given by a factor of $(1 - \eta\gamma w^2/\beta)$ per round which is linear in the shrinkage.

The L2 penalty λ plays a regularizing role visible directly in the curvature. It increases the effective second derivative from H to H + λ, shrinking the Newton step toward zero and acting as a trust region that prevents any single leaf from taking an extreme value when its Hessian mass H is small — precisely the situation for leaves containing few or uncertain instances. In this sense λ implements a per-leaf adaptive damping, larger in relative terms where the data are sparse, which is the mechanism by which it controls variance in the high-dimensional fused space.

Finally, the column subsampling that decorrelates trees can be read geometrically as optimizing over a random low-dimensional subspace of features at each node. This both accelerates split finding — fewer candidate features to scan — and injects the diversity that lowers the ensemble-variance floor $\rho\sigma^2$ of (105). The combined picture is of a regularized, damped, stochastic second-order method, each ingredient of which has a precise effect on the bias–variance balance and the convergence rate, and each of which is tuned in Table IX to the demands of the fused osteoporosis-prediction problem.

EXTENDED EVALUATION PROTOCOL AND STATISTICAL TESTING

A rigorous evaluation reports not only point metrics but their uncertainty and the significance of differences. Three layers of analysis are recommended for this framework, building on the metrics of the metrics section. The first layer is interval estimation: every reported metric should carry a confidence interval, computed analytically where a

formula exists — as for AUC via the Hanley–McNeil error (106) — or by the bootstrap (123) otherwise. An AUC reported as 0.9083 is only interpretable alongside the width of its interval, which depends on the test-set size and the score distribution.

The second layer is paired hypothesis testing between models evaluated on the same data. For thresholded predictions the McNemar test (104) assesses whether the two models' error patterns differ significantly. For the AUC difference, the DeLong test provides a non-parametric comparison of two correlated ROC curves, using the theory of generalized U-statistics to estimate the covariance of the two AUC estimates,

$$Z = (A_{\text{hyb}} - A_{\text{clin}}) / \sqrt{(\text{Var}(A_{\text{hyb}}) + \text{Var}(A_{\text{clin}}) - 2 \text{ Cov})} \quad (142)$$

which is compared against the standard normal. The covariance term is positive because both models are evaluated on the same cases, and accounting for it makes the test more powerful than treating the two AUCs as independent. Applying DeLong to the 0.0150 gain would determine whether it is distinguishable from sampling noise at the available sample size.

The third layer relates to multiple comparisons. When several models or several operating points are compared, the chances of a false significant result increases, and a correction like Bonferroni divides the significance level α by the number of comparisons m,

$$\alpha_{\text{adj}} = \alpha / m \quad (143)$$

A trade off exists among these procedures in terms of power and control of the family-wise error rate. When a large number of comparisons are made, a less conservative procedure such as the Benjamini–Hochberg control of the false-discovery rate is preferable. Currently, there is no correction necessary for the two-model comparison envisaged but with ablation and subgroup analyses suggested in fairness and class-imbalance discussions the principle becomes relevant. The goal of this framework is to report effect sizes alongside p-values – the AUC gain and its interval rather than a mere verdict of significance. This is because a statistically significant but clinically negligible gain and a clinically meaningfully but underpowered one are two different situations that bare p-value would conflate.

WHY TRANSFER LEARNING WORKS: A THEORETICAL ACCOUNT

The imaging branch relies on ImageNet-pretrained weights used without target-domain fine-tuning, and it is worth making explicit why descriptors learned on natural images transfer to CT bone analysis. Domain-adaptation theory bounds the target-domain error of a hypothesis in terms of its source-domain error and a divergence between the two domains. For a hypothesis h, the target risk satisfies

$$\varepsilon_{\text{T}}(h) \leq \varepsilon_{\text{S}}(h) + \tfrac{1}{2} d_{\text{H}\Delta\text{H}}(D_{\text{S}}, D_{\text{T}}) + \lambda^{*} \quad (144)$$

where $d_{\text{H}\Delta\text{H}}$ is the divergence between source and target distributions measured through the hypothesis class, and λ^{*} is the error of the best joint hypothesis. The bound says transfer succeeds when the domains are close in the sense that the same features discriminate in both. Early convolutional filters learn generic edge, texture and gradient detectors whose statistics are largely domain-invariant; these are exactly the structures relevant to cortical and trabecular bone texture, so the divergence term is small for the low- and mid-level features that dominate the ResNet50 representation.

A complementary view comes from the information bottleneck. A deep network trades off compression of the input against retention of label-relevant information, seeking a representation Z that minimizes

$$L_{\text{IB}} = I(Z; X) - \beta I(Z; Y) \quad (145)$$

ImageNet pretraining drives the representation toward retaining broadly useful visual information; when read out by global average pooling, the resulting 2048-vector is a compressed summary that, by the data-processing inequality, can carry at most the label information present in the input but in a form already organized along semantically meaningful axes. Fine-tuning would adapt the high-level axes to bone-specific labels and could further raise $I(Z; Y)$; the present frozen design forgoes this in exchange for determinism, reproducibility and freedom from target-domain overfitting on a modest dataset — a deliberate point on the bias–variance frontier rather than an oversight.

The practical corollary is that the benefit of the imaging branch should be largest where pretrained features align with bone morphology and smallest where CT-specific patterns (for example, Hounsfield-unit calibration or scanner-specific noise) carry the signal. This predicts that fine-tuning the final residual stage, which adapts the most

domain-specific high-level features while preserving the transferable early filters, is the highest-yield extension, consistent with the recommendation in the future-work discussion.

PROBABILITY CALIBRATION METHODS

the metrics section introduced the expected calibration error (97); here the corrective methods are developed, since a screening tool that reports risk probabilities must have those probabilities mean what they say. Platt scaling fits a one-dimensional logistic regression to the model scores on a held-out calibration set,

$$p_{\mathrm{cal}} = \sigma(A\,\hat{z} + B) \qquad (146)$$

learning a slope A and intercept B that rescale the log-odds. Platt scaling is appropriate when the miscalibration is well described by a sigmoidal distortion, which is common for margin-based and boosted classifiers whose raw scores are over-confident near the extremes.

Isotonic regression is a non-parametric alternative that fits an arbitrary monotone increasing map from scores to calibrated probabilities by minimizing squared error subject to monotonicity,

$$\min_{m\uparrow} \sum_i (y_i - m(\hat{z}_i))^2 \qquad (147)$$

solved by the pool-adjacent-violators algorithm. Isotonic regression is more flexible than Platt scaling and can correct non-sigmoidal distortions, but it requires more calibration data to avoid overfitting the step function, so Platt scaling is often preferred for smaller calibration sets such as the present one.

Calibration quality is assessed by the reliability diagram, which plots empirical accuracy against predicted confidence per bin; a perfectly calibrated model lies on the diagonal, and the gap from it is the per-bin term summed in the ECE (97). A complementary scalar is the Brier score, the mean squared error of the probabilistic prediction,

$$\mathrm{BS} = (1/n) \sum_i (\hat{p}_i - y_i)^2 \qquad (148)$$

which reduces to calibration, refinement, and irreducible-uncertainty terms through the Murphy decomposition, distinguishing between the calibration of probabilities and their discrimination power. When the calibration of the probabilities is inadequate, the cost-efficient threshold τ^* of (13) cannot be deemed reliable. In such instances, the cost-optimal operational point computed on the nominal scale will not be cost-optimal in reality. As a result, calibration is not a mere cosmetic post-processing stage but rather a prerequisite for the principled threshold selection required by the screening application.

NUMERICAL STABILITY AND IMPLEMENTATION CONSIDERATIONS

Correct implementation of the mathematics requires care about numerical stability, since naïve evaluation of some expressions overflows or loses precision. When calculated as 1/(1 + e^{−z}), the logistic function (9) overflows for large negative scores; the numerically stable version depends on the sign of z,

$$\sigma(z) = e^z/(1+e^z) \text{ if } z < 0, \quad 1/(1+e^{-z}) \text{ otherwise} \qquad (149)$$

Similarly the cross-entropy loss is evaluated through the log-sum-exp trick to avoid the logarithm of a near-zero probability,

$$\log(1+e^z) = \max(0,z) + \log(1+e^{-|z|}) \qquad (150)$$

which is exact and bounded for all z. In the leaf-weight computation (67) the denominator $H_j + \lambda$ is bounded away from zero by $\lambda > 0$, so the regularizer also serves a numerical purpose: it prevents division by a vanishing Hessian mass when a leaf contains few or highly confident instances, for which $h_i = \hat{p}_i(1-\hat{p}_i)$ is tiny. This is a concrete instance of regularization improving conditioning, not merely generalization.

The CT feature extraction introduces its own considerations. Intensities are converted to floating point and normalized by (35) before the forward pass; performing this in single precision is sufficient because the batch-normalization layers re-standardize internally, bounding the dynamic range of activations. Accumulating the global average pool (45) in a higher-precision accumulator avoids the catastrophic cancellation that can arise when averaging thousands of spatial locations. Finally, the gradient and Hessian sums (66) over large leaves are best computed with a compensated (Kahan) summation when n is large, although for the present sample size ordinary summation in double precision is more than adequate.

On the systems side, the feature blocks are stored in a compressed sparse format because one-hot encoding (26) produces many zeros; XGBoost's sparsity-aware split finding visits only non-missing entries, giving the per-tree

cost a dependence on the number of non-zeros rather than the nominal dimension, which is what makes training over the p + 2048 fused space efficient despite its width. These implementation choices do not alter the mathematics but are necessary for it to be realized faithfully and at scale.

RELATIONSHIPS AMONG THE EVALUATION METRICS

The five reported metrics are not independent, and understanding their algebraic relationships prevents over-interpretation of any single number. Accuracy is a prevalence-weighted average of recall (sensitivity) and specificity,

$$\mathrm{Acc} = \pi \cdot \mathrm{Recall} + (1-\pi) \cdot \mathrm{Spec} \qquad (151)$$

where $\pi = n_+/n$ is the positive prevalence. Under imbalance (π small) accuracy is dominated by specificity, which is why a model can post high accuracy while missing many positives — the precise failure mode that makes recall and AUC indispensable in screening.

The F1-score and AUC measure different things and can move in opposite directions. F1 (86) is a threshold-dependent summary of the precision–recall balance at a single operating point, whereas AUC (88) is a threshold-independent ranking measure. A model can improve its ranking (higher AUC) while its F1 at $\tau = 0.5$ is unchanged or even slightly lower, exactly as observed here where F1 rose marginally ($0.8588 \rightarrow 0.8603$) while AUC rose substantially ($0.8933 \rightarrow 0.9083$). The relationship between precision, recall and prevalence is

$$\mathrm{Prec} = \pi \cdot \mathrm{Recall} \,/\, (\pi \cdot \mathrm{Recall} + (1-\pi) \cdot (1-\mathrm{Spec})) \qquad (152)$$

showing that precision depends on prevalence even when the classifier's sensitivity and specificity are fixed; a model deployed in a population with different prevalence than the test set will exhibit different precision, a fact that must be communicated when reporting a screening tool. This prevalence dependence is the reason AUC — which is invariant to prevalence because it conditions on the true class — is the most portable single summary across deployment settings.

Finally, the Matthews correlation coefficient offers a balanced single number that uses all four confusion-matrix cells and behaves well under imbalance,

$$\mathrm{MCC} = (TP \cdot TN - FP \cdot FN) \,/\, \sqrt{((TP+FP)(TP+FN)(TN+FP)(TN+FN))} \qquad (153)$$

ranging from −1 to +1 with 0 indicating chance. MCC is high only when the classifier does well on both classes, making it a stringent complement to accuracy; reporting it alongside the existing metrics would further guard against the imbalance-induced optimism discussed in the class-imbalance discussion. Together these relationships show that the metric set should be read as a whole — each number constrains and contextualizes the others — rather than reduced to a single headline figure.

SYNTHESIS: HOW THE PIECES FIT TOGETHER

It is useful to step back and view the framework as a single coherent argument rather than a sequence of techniques. The clinical premise (the clinical background) is an asymmetric cost structure and two complementary modalities. This premise selects the evaluation philosophy: threshold-independent ranking quality and sensitivity, formalized in the metrics of the metrics section and their relationships in the metric-relationships discussion. The modalities are each conditioned by branch-appropriate preprocessing — explicit operators for the clinical block (the clinical-preprocessing section, V-A) and a deep, batch-normalized, transfer-learned extractor for the imaging block (the imaging-branch sections, VII-B, XXXIV) — and then joined by feature-level fusion (the fusion section), chosen from the broader design space of the alternative-fusion discussion for its balance of expressiveness and parsimony.

The joint representation is classified by a regularized gradient-boosted ensemble whose every step is derived (the XGBoost derivation) and whose optimization geometry, generalization guarantees and information-theoretic rationale are made explicit (the functional-gradient discussion, XXXII, XXIII, XVII). The regularization that the derivation introduces is precisely what controls the variance that the high-dimensional imaging block would otherwise contribute, closing the bias–variance argument of the dimensionality discussion. The empirical result — unchanged accuracy, improved AUC and recall — is then not a surprise but the predicted signature of adding a partially complementary modality under proper regularization, interpreted in detail in the detailed results reading.

Every limitation has a corresponding mathematical statement and a corresponding remedy: index-wise alignment and its label-noise bound (the limitations discussion) point to matched-cohort validation; the modest, possibly non-significant gain points to the interval estimation and DeLong testing of the evaluation-protocol discussion; the frozen backbone points to selective fine-tuning justified by the transfer bound (144); and the imbalance

and subgroup concerns (the class-imbalance discussion, XXV) point to cost-sensitive learning and fairness auditing. The framework is thus not only a method but a map of its own assumptions and the analyses that would test them, which is the standard appropriate to a dissertation-grade treatment.

SUMMARY OF COMPUTATIONAL COMPLEXITY

For reference, the asymptotic costs of each stage derived throughout are collected here. Let n be the number of samples, p the clinical dimension, $d = p + 2048$ the fused dimension, K the number of trees, and Δ the maximum depth.

TABLE XI. COMPUTATIONAL COMPLEXITY BY STAGE

Stage	Time	Ref.
Clinical preprocess	$O(n\ p)$	(29)
CT extraction	$O(n \cdot \text{FLOPs})$	(47)
Fusion (concat)	$O(n\ d)$	(54)
XGBoost train	$O(K\ \Delta\ d\ n \log n)$	(72)
XGBoost infer	$O(K\ \Delta)$	(114)
TreeSHAP	$O(K\ T\ \Delta^2)$	(137)

The CT extraction term dominates wall-clock time at inference, as noted in the deployment-cost discussion, while training is dominated by the boosting term. All stages are polynomial and the constants are small enough that the complete pipeline runs comfortably on commodity hardware with optional GPU acceleration for the backbone, supporting the deployability claim of this work.

CONCLUSION

This paper presented a mathematically explicit hybrid ensemble deep-learning framework for osteoporosis prediction. The clinical-only XGBoost baseline reached 0.8724 accuracy and 0.8933 AUC. A frozen ResNet50 produced 2048-dimensional CT descriptors, which were fused at the feature level with preprocessed clinical variables and classified by a regularized XGBoost ensemble. The hybrid model improved AUC to 0.9083 (+0.0150, +1.68% relative) and recall by 1.02 points while maintaining accuracy, demonstrating that imaging features add discriminative value not captured by clinical variables alone. The complete derivation — from imputation and residual convolution through the second-order boosting objective, optimal leaf weights and split gain — makes the pipeline fully reproducible and suitable for dissertation-level documentation.

FUTURE WORK

Future work should validate the framework on hospital cohorts with truly matched clinical and CT records from the same patients, and perform external validation across scanners, populations and institutions before any clinical deployment. Methodologically, attention-based and transformer fusion, stacked ensembles, and end-to-end fine-tuning of the imaging branch may further improve discrimination. Explainability is essential: SHAP values can attribute each prediction to clinical and fused coordinates, and Grad-CAM can localize the CT regions driving the descriptors, supporting clinician trust and verifying that the model attends to clinically meaningful bone structure.

References

1. H. Cheng *et al.*, “Artificial intelligence in osteoporosis assessment using CT imaging: a scoping review,” *Front. Med.*, vol. 13, p. 1779483, Feb. 2026, doi: 10.3389/fmed.2026.1779483.
2. R. Zhao *et al.*, “The Diagnostic Value of Image-Based Machine Learning for Osteoporosis: Systematic Review and Meta-Analysis,” *J Med Internet Res*, vol. 28, pp. e75965–e75965, Jan. 2026, doi: 10.2196/75965.
3. X. Shen, J. Xiong, S. Wang, G. Hu, H. Liu, and S. Zhang, “Application of machine learning in osteoporosis screening: a narrative review,” *npj Digit. Med.*, Apr. 2026, doi: 10.1038/s41746-026-02516-6.
4. F. Ghareh Mohammadi and R. Sebro, “Artificial Intelligence for Opportunistic Screening for Osteoporosis and Spine Fractures Using Computed Tomography: A Systematic Review and Meta-Analysis,” *Journal of Computer Assisted Tomography*, Jun. 2026, doi: 10.1097/RCT.0000000000001899.
5. 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*, San Francisco California USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

6. K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.
7. A. Hiyama, D. Sakai, H. Katoh, M. Sato, and M. Watanabe, “Osteoporosis prediction using lumbar CT Hounsfield units: comparative performance and clinical implications of seven machine learning models,” *Eur Spine J*, vol. 35, no. 3, pp. 1149–1161, Mar. 2026, doi: 10.1007/s00586-026-09753-z.
8. A. Wang *et al.*, “Deep Learning–Assisted Automated Diagnosis of Osteoporosis Based on Computed Tomography Scans: Systematic Review and Meta-Analysis,” *J Med Internet Res*, vol. 27, pp. e77155–e77155, Nov. 2025, doi: 10.2196/77155.
9. S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” 2015, *arXiv*. doi: 10.48550/ARXIV.1502.03167.
10. J. H. Friedman, “Greedy function approximation: A gradient boosting machine.,” *Ann. Statist.*, vol. 29, no. 5, Oct. 2001, doi: 10.1214/aos/1013203451.
11. B. Suh *et al.*, “Interpretable Deep-Learning Approaches for Osteoporosis Risk Screening and Individualized Feature Analysis Using Large Population-Based Data: Model Development and Performance Evaluation,” *J Med Internet Res*, vol. 25, p. e40179, Jan. 2023, doi: 10.2196/40179.
12. J. Smets, E. Shevroja, T. Hügle, W. D. Leslie, and D. Hans, “Machine Learning Solutions for Osteoporosis—A Review,” *Journal of Bone and Mineral Research*, vol. 36, no. 5, pp. 833–851, Dec. 2020, doi: 10.1002/jbmr.4292.
13. T. Peng *et al.*, “A study on whether deep learning models based on CT images for bone density classification and prediction can be used for opportunistic osteoporosis screening,” *Osteoporos Int*, vol. 35, no. 1, pp. 117–128, Jan. 2024, doi: 10.1007/s00198-023-06900-w.
14. J. Wang, Y. He, L. Yan, S. Chen, and K. Zhang, “Predicting Osteoporosis and Osteopenia by Fusing Deep Transfer Learning Features and Classical Radiomics Features Based on Single-Source Dual-energy CT Imaging,” *Academic Radiology*, vol. 31, no. 10, pp. 4159–4170, Oct. 2024, doi: 10.1016/j.acra.2024.04.022.
15. W. Ong *et al.*, “Artificial Intelligence Applications for Osteoporosis Classification Using Computed Tomography,” *Bioengineering*, vol. 10, no. 12, p. 1364, Nov. 2023, doi: 10.3390/bioengineering10121364.
16. L. Cheng, F. Cai, M. Xu, P. Liu, J. Liao, and S. Zong, “A diagnostic approach integrated multimodal radiomics with machine learning models based on lumbar spine CT and X-ray for osteoporosis,” *J Bone Miner Metab*, vol. 41, no. 6, pp. 877–889, Nov. 2023, doi: 10.1007/s00774-023-01469-0.
17. A. Paderno *et al.*, “Artificial intelligence-enhanced opportunistic screening of osteoporosis in CT scan: a scoping Review,” *Osteoporos Int*, vol. 35, no. 10, pp. 1681–1692, Oct. 2024, doi: 10.1007/s00198-024-07179-1.

APPENDIX A: DERIVATION OF THE LOGISTIC GRADIENT AND HESSIAN

For completeness the first- and second-order statistics used in (65) are derived from first principles. The binary cross-entropy loss for a single instance with label y and score z, where p = σ(z) = 1/(1 + e^{−z}), is

$$\ell(y, z) = -\, y \log \sigma(z) - (1-y) \log(1 - \sigma(z)) \qquad (154)$$

Using the identity σ′(z) = σ(z)(1−σ(z)), the derivative of each term gives

$$\partial \ell / \partial z = -y(1-\sigma) + (1-y)\sigma = \sigma(z) - y \qquad (155)$$

which is the gradient $g = p - y$ of (65). Differentiating once more,

$$\partial^2 \ell / \partial z^2 = \sigma'(z) = \sigma(z)(1-\sigma(z)) = p(1-p) \qquad (156)$$

which is the curvature $h = p(1-p)$. Both are independent of y in the case of h, and $h > 0$ everywhere, confirming the strict convexity of the per-instance loss in the score and hence the well-posedness of the leaf-weight minimization (67).

APPENDIX B: OPTIMAL LEAF WEIGHT AND STRUCTURE SCORE

Starting from the leaf-grouped objective (66), consider a single leaf j with aggregates G_j and H_j and weight w_j. The leaf contributes

$$q_j(w_j) = G_j w_j + \tfrac{1}{2} (H_j + \lambda)\, w_j^2 \qquad (157)$$

Setting the derivative to zero,

$$\partial q_j / \partial w_j = G_j + (H_j + \lambda)\, w_j = 0 \qquad (158)$$

yields the optimal weight (67). Substituting back into (157),

$$q_j(w_j^*) = -\tfrac{1}{2}\, G_j^2 \,/\, (\, H_j + \lambda\,) \qquad (159)$$

and summing over all T leaves and adding the leaf-count penalty γT gives the structure score (68). The split-gain (69) is then the difference between the structure scores of the parent leaf and the two children it would produce, with the γ term capturing the cost of the additional leaf created by the split. This completes the chain from the regularized objective to the practical greedy split-finding criterion, leaving no step implicit.

These appendices, together with Algorithms 1–5 and the worked examples of the worked examples, constitute a fully self-contained mathematical specification: a reader with the stated datasets and hyperparameters can reconstruct every quantity from the raw inputs to the reported metrics, which was the stated aim of producing a rigorous, dissertation-grade treatment of the hybrid framework.

APPENDIX C: WORKED AUC COMPUTATION

To illustrate the Mann–Whitney form of the AUC (88), consider a miniature test set with three positives scored {0.9, 0.6, 0.4} and three negatives scored {0.7, 0.5, 0.2}. The AUC equals the fraction of positive–negative pairs in which the positive outscores the negative, over the nine pairs. Enumerating: the positive 0.9 beats all three negatives, contributing three winning pairs; the positive 0.6 beats 0.5 and 0.2 but not 0.7, contributing two; and the positive 0.4 beats only 0.2, contributing one. The total of winning pairs is 3 + 2 + 1 = 6, with no ties, so

$$\mathrm{AUC} = 6 / (3 \times 3) = 0.667 \quad (160)$$

The same value is obtained from the trapezoidal area under the ROC curve traced by sweeping the threshold through the six scores, confirming the equivalence of the geometric and probabilistic definitions. This small example also shows why AUC is insensitive to the absolute score scale: any strictly monotone rescaling of the scores leaves every pairwise comparison, and hence the AUC, unchanged — the ranking-only property that makes AUC the portable summary emphasized in the metric-relationships discussion.

APPENDIX D: ABBREVIATIONS

Abbr.	Meaning
AUC	Area Under the ROC Curve
BMD	Bone Mineral Density
BN	Batch Normalization
CNN	Convolutional Neural Network
CT	Computed Tomography
DXA	Dual-energy X-ray Absorptiometry
ECE	Expected Calibration Error
GAP	Global Average Pooling
MCC	Matthews Correlation Coefficient
ReLU	Rectified Linear Unit
ROC	Receiver Operating Characteristic
SMOTE	Synthetic Minority Oversampling
XGBoost	Extreme Gradient Boosting