Vertebral Fracture Risk in Spinal Metastases Patients Following Stereotactic Body Radiotherapy Using Quantitative Imaging Data and Machine Learning

Dawit Gulta*, Geoff Klein*‖, Tayler D Ross*, Matthew Rezkalla*, Daniel M- Palhare*, Laura Burgess*, Jay Detsky*, Arjun Sahgal*, Cari M. Whyne*‡§¶, Michael Hardisty*‡¶, Stephen Chen†

*Sunnybrook Research Institute, Toronto, Canada

†School of Information Technology, York University, Toronto, Canada

‡Institute of Biomedical Engineering, University of Toronto, Toronto, Canada

§Institute of Medical Science, University of Toronto, Toronto, Canada

¶Department of Surgery, University of Toronto, Toronto, Canada

‖Department of Medical Biophysics, University of Toronto, Toronto, Canada

***Abstract*—Vertebral compression fractures (VCFs) occur in approximately 14% of patients with spinal metastases following treatment with Stereotactic Body Radiotherapy (SBRT). The Spinal Instability Neoplastic Score (SINS) is the current clinical standard for assessing potential mechanical instability in these patients; however, it has several limitations such as it is manually assessed, has an inconsistent relationship with fracture risk and is only semi-quantitative. This study used quantitative CT imaging biomarkers derived from SBRT treatment planning imaging and machine learning models to predict vertebral compression fractures (VCF) following SBRT in spinal metastases patients (in 300 thoraco-lumbar vertebral segments from 179 patients). Fractures occurred in 18.3% of cases post-SBRT. Machine learning algorithms (Logistic Regression, Random Forest, XGBoost, SVM, Gradient Boosting, AdaBoost, Neural Network) were compared against SINS. The Random Forest model achieved the best performance (sensitivity: 0.64, specificity: 0.76, F1-score: 0.47), showing a 36% improvement in balanced accuracy over SINS. Feature importance analysis identified spinal alignment and tumour composition(lytic or blastic disease) as the strongest predictors. ML models demonstrated meaningful improvements over traditional SINS assessment.**

***Index Terms*—Machine Learning, Vertebral Fractures, SINS, SBRT, Medical Imaging, Fracture Risk Prediction**

I. Introduction

Stereotactic Body Radiotherapy (SBRT) is as a highly effective and precise treatment modality for vertebral metastases, offering significant advantages over conventional radiotherapy [1]. By delivering high doses of radiation in a small number of fractions, SBRT achieves superior local tumour control while minimizing damage to surrounding healthy tissues [2]. This targeted approach has been widely adopted in clinical oncology for its ability to alleviate pain, preserve spinal function, and improve patient quality of life.

This work has been supported by Natural Sciences and Engineering Research Council of Canada (NSERC) through a Discovery Grant – RGPIN-2022-04524 and by MITACS through an Accelerate Grant – IT40018, the Canadian Institutes of Heath Research (CIHR) and the Feldberg Chair for Spine Research, INOVAIT, Varian Medical Systems,

Despite its therapeutic benefits, SBRT is associated with an increased risk of vertebral compression fractures (VCFs)—a serious complication that can lead to structural spinal instability, neurological deficits, and significant morbidity. Clinical studies have reported that 10–15% of non-surgically stabilized patients undergoing SBRT develop VCFs within months to years following treatment [3]–[5]. The risk of fracture is influenced by various factors, including pre-existing bone quality, tumour burden, radiation dose, and the biomechanical integrity of the affected vertebrae. In particular, high-dose SBRT regimens, while effective for tumour ablation, may accelerate bone remodeling, leading to increased bone fragility and post-treatment vertebral collapse [1], [3], [6].

VCFs present a substantial challenge in spinal metastases management, as they not only impact patient mobility and quality of life but also complicate clinical decision-making regarding further treatment interventions [4]. Patients who develop fractures often require surgical stabilization, vertebroplasty, or additional palliative measures to manage pain and preserve spinal function [6]. Therefore, accurately predicting which patients are at heightened risk of developing VCFs post-SBRT is crucial for optimizing treatment planning and improving patient outcomes.

A. The Role of the Spinal Instability Neoplastic Score (SINS) in Clinical Decision-Making

To aid clinicians in assessing spinal stability and fracture risk, the SINS is widely used in clinical practice. Developed by the Spine Oncology Study Group [7], SINS provides a structured, semi-quantitative assessment of vertebral instability by evaluating six key components (see Table I):

1) Location of the tumour (junctional vs. non-junctional regions)
2) Pain characteristics (mechanical vs. non-mechanical pain)
3) Bone lesion type (lytic, blastic, or mixed)
4) Radiographic spinal alignment (normal, subluxation, or translation)

5) Vertebral body collapse severity
6) Posterolateral involvement of spinal elements

Each component is assigned a numerical score (Table I), leading to a total SINS score ranging from 0 to 18, categorizing patients as stable (0–6 points), potentially unstable (7–12 points), or unstable (13–18 points). This classification is used to determine the need for surgical consultation and possible mechanical stabilization (Vertebroplasty, surgical stabilization).

TABLE I
SPINAL INSTABILITY NEOPLASTIC SCORE (SINS) COMPONENTS

Element	Score
Location	
Junctional (occiput-C2, C7–T2, T11–L1, L5–S1)	3
Mobile spine (C3–C6, L2–L4)	2
Semi-rigid (T3–T10)	1
Rigid (S2–S5)	0
Pain with movement/loading of the spine	
Yes	3
No (occasional pain but not mechanical)	1
Pain free lesion	0
Bone Lesion	
Lytic	2
Mixed (lytic/blastic)	1
Blastic	0
Radiographic Spinal Alignment	
Subluxation/translation present	4
De novo deformity (kyphosis/scoliosis)	2
Normal alignment	0
Vertebral Body Collapse	
>50% collapse	3
<50% collapse	2
No collapse with >50% body involved	1
None of the above	0
Posterolateral Involvement	
Bilateral	3
Unilateral	1
None of the above	0

Note: SINS is used to assess spinal instability in oncology patients. Higher scores indicate greater instability risk.

While SINS is widely used as the standard for fracture risk assessment in spinal metastases, its clinical utility is limited by several drawbacks:

1) Binary outcome prediction: SINS categorizes patients into stable, potentially unstable, and unstable groups, but does not provide specific probability estimates for fracture occurrence, limiting risk stratification precision [8].
2) Each of the six SINS components contributes equally to the total score, but not all have equal predictive power [9].
3) Limited use of quantitative imaging biomarkers: Lacks biomechanical modeling and quantitative validation with CT-based structural metrics

B. Bridging the Gap with Machine Learning-Based Quantitative Analysis

Analysis of computed tomography (CT) imaging can quantify features important for mechanical stability (i.e. bone quality, tumour burden) that when combined with machine learning (ML) offer potential solutions to the limitations of manual SINS scoring. By leveraging quantitative biomarkers, ML-based models can provide more objective, accurate, and scalable assessments of spinal instability.

This study aims to develop an automated approach that integrates advanced CT-based feature extraction with machine learning algorithms to enhance fracture risk prediction. Specifically, this approach utilizes computational methods for:

1) Automated vertebrae segmentation to ensure consistent anatomical labeling.
2) Quantification of osteolytic and osteoblastic lesions to assess tumour-related bone changes.
3) Volumetric evaluation of vertebral collapse to provide precise structural integrity measurements.
4) Measurement of spinal alignment using angle-based analysis of vertebral positioning.

By comparing the performance of our machine learning models against traditional SINS scoring, this study aims to determine whether a data-driven, automated approach can enhance fracture risk prediction and reduce observer variability.

C. Aim and Objectives

The primary aim of this study is to develop a machine learning-based framework for vertebral fracture risk prediction in SBRT-treated patients using quantitative imaging biomarkers. The specific objectives are:

1) To automate the quantification of key SINS components as quantitativer imaging biomarkers including vertebral collapse, spinal alignment, and bone lesion characterization—using CT-based quantitative analysis.
2) To develop and validate machine learning models that incorporate both imaging biomarkers and clinical features-specifically mechanical pain and posterolateral involvement- to predict post-SBRT vertebral fractures.
3) To compare the predictive performance of machine learning-based models against traditional SINS scoring and evaluate their potential for improving clinical decision-making.
4) To assess the impact of integrating both quantitative (imaging-derived) and qualitative (clinician-derived) features on fracture risk prediction accuracy.

Based on the integration of these quantitative imaging biomarkers, we anticipate improved fracture risk prediction accuracy compared to conventional SINS assessment. The quantitative features allow greater fidelity and greater precision than the semi-quantitative SINS element. For instance quantitative assessment of tumour burden as a continuous percentage can better represent changes in bone integrity and mechanical compromise surpassing the fidelity possible with the reduced categories (lytic, mixed, blastic) used in SINS.

II. LITERATURE REVIEW

VCFs are a common complication in spinal metastases post-SBRT, with incidence rates of 10% to 40% [4], [6], [10]. SINS [7] was developed to assess spinal instability and guide clinical decision-making by categorizing patients as stable, potentially unstable, or unstable based on six imaging-based criteria. While widely adopted, SINS includes subjective components—such as assessments of alignment, bone lesion quality, and pain—that may lead to inter-observer variability and inconsistent scoring, particularly among less experienced clinicians.

Although inter-observer agreement has been shown to be generally good (e.g., ICC $\approx$ 0.8), limitations remain in terms of reproducibility, scalability, and the categorical nature of the score, which does not yield individualized risk estimates [9]. Moreover, manual SINS assessment is time-intensive and not practical for large-scale or automated workflows. To address these gaps, recent efforts have focused on integrating quantitative imaging biomarkers and machine learning models to improve the precision, objectivity, and scalability of fracture risk prediction.

Studies have explored automated SINS scoring to improve objectivity. Kim et al. [11] developed an AI-based system for fracture risk assessment in spinal radiotherapy. Early automation improved reproducibility but required manual segmentation, highlighting the need for fully automated approaches. Advances in deep learning and computer vision have since enabled more sophisticated assessments of spinal stability. Liu et al. utilized convolutional neural networks (CNNs) to predict primary tumour sites in spinal metastases, demonstrating the feasibility of integrating AI into oncologic imaging [12]. However, challenges remain in standardizing and integrating these automated models into routine clinical practice.

CT-derived imaging biomarkers—vertebral morphology, bone density, and lesion characterization—enhance fracture risk assessment, offering a data-driven alternative to SINS [13]. These models quantify vertebral collapse, osteolytic/blastic lesion involvement, and changes in vertebral structure over time, providing more objective measures of spinal stability [14].

Machine learning and AI-based approaches have demonstrated potential in improving diagnostic accuracy and predictive modeling in spinal imaging. Nia et al. highlighted AI's role in enhancing disease detection and treatment prediction across multiple medical fields, including spinal stability assessment [15]. Wang et al. [16] further emphasized AI's ability to optimize clinical decision-making by integrating imaging biomarkers with patient-specific variables. Recent studies have also shown that AI-driven radiomics can distinguish between malignant vertebral metastases and osteoporotic fractures with high diagnostic accuracy [16]. These findings suggest that AI-based methods can improve specificity and reliability, particularly in assessing individual SINS elements that are subject to subjectivity [17].

III. METHODOLOGY

A. Study Population and Data Source

This retrospective study was conducted after Research Ethics Board approval at Sunnybrook Health Sciences Centre, Toronto, Canada (REB #: 2623). The initial dataset comprised 1,406 records from 590 patients diagnosed with spinal metastases and treated with SBRT. To ensure consistency and clinical relevance, inclusion criteria were applied, selecting cases with complete values of SINS components and vertebral levels between T4 and L5. Patients with prior surgical stabilization were excluded from the study to minimize the confounding effects on fracture risk prediction.

Following data curation, the final data set comprised 300 vertebral segments from 179 patients, with a median age of 65 years (range: 18 to 87 years). The cohort included 44% female and 56% male participants. Fracture incidence was determined through follow-up imaging, identifying 55 vertebral fractures (18.3% of cases). These fractures, including both progressive pre-existing and new-onset fractures, were identified during follow-up, with a median detection time of 570 days post-SBRT.

For baseline comparison, the conventional SINS assessment was evaluated for fracture risk prediction. Vertebrae were classified as fracture-prone if their SINS scores fell within the potentially unstable (7-12) or unstable (13-18) categories, and these predictions were compared against the actual fracture outcomes observed during follow-up.

B. Data Processing and Curation

A structured data pre-processing pipeline was implemented to ensure data completeness, accuracy, and consistency. Records with missing SINS values were excluded to maintain the integrity of the dataset.

The initial data set contained 139 features across multiple domains. The selection of features was guided by clinical relevance to the six SINS components and feature importance rankings despite class imbalance (18.3% fracture cases). Using both domain knowledge and quantitative feature selection methods, we identified seven key predictors: coronal angle, lytic percentage, sagittal alignment, blastic percentage, collapse percentage, and two SINS parameters (location and posterolateral involvement).

This preprocessing framework optimized the quality of the data for downstream machine learning analysis, allowing the integration of clinical and imaging-based predictors for the assessment of fracture risk.

C. Image Analysis and Feature Extraction

CT image processing and quantitative feature extraction were conducted using 3D Slicer (version 5.6.2-linux-amd64, r32448 f10cd8c), leveraging automated segmentation tools for vertebral body analysis. The VertDetect [18] deep learning model was used for vertebral segmentation, providing reliable vertebral instance labels and structural delineations at spinal levels T4 to L5 (Fig. 1).

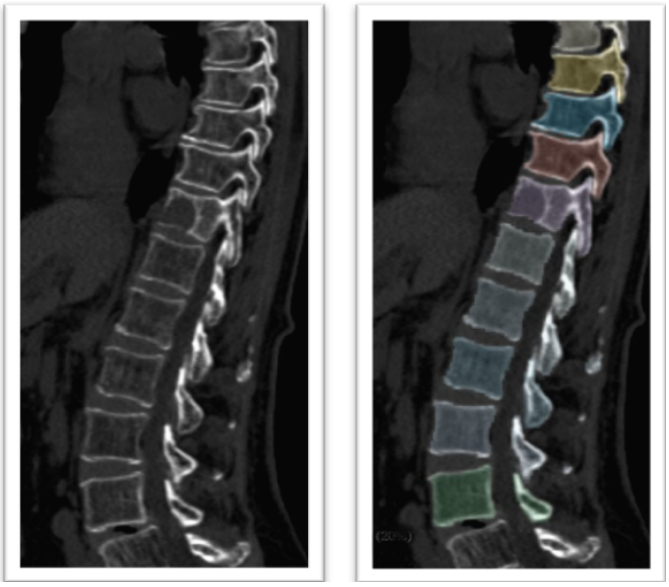

Fig. 1. Spine CT Image Segmentation

Following segmentation, the key quantitative imaging biomarkers were extracted using computational algorithms designed to assess vertebral collapse, tumour burden, and spinal alignment:

1) Vertebral Collapse Metrics: The *Collapse Percentage* was computed by estimating the pre-fracture vertebral volume using a polynomial regression-based approach. Vertebral body volumes were computed from segmentation masks and used to fit a third-degree polynomial model across anatomically ordered vertebrae, excluding the treatment level. This model estimated the expected volume of the treatment vertebra by referencing adjacent levels, allowing calculation of the percent volume loss due to collapse. The collapse percentage ($C\%$) was calculated as:
$$C\% = \frac{V_{\text{estimate}} - V_{\text{measured}}}{V_{\text{estimate}}} \times 100$$
2) Bone Lesion Characterization Lytic Percentage and Blastic Percentage were determined using a previously developed method [13]. Briefly it uses histogram analysis to derive thresholds (Lytic: below 75HU, Blastic: above 330HU) that are applied with the vertebral body segmentation (Fig. 2). The extent of lytic and blastic invovlement are then expressed in percentage terms by dividing by the total vertebral body volume.
3) Spinal Alignment Metrics
Spinal alignment was assessed using a previously described hybrid computational approach that integrates both spline-based and contour-based methods to quantify coronal and sagittal angles [19]:
   - **Spline-Based Method**: A smooth curve was fitted through the vertebral body centroid predictions to evaluate spinal curvature. The gradient of the spline was calculated at two points: at the top of the superior adjacent vertebra and at the bottom of the inferior adjacent vertebra to the vertebra of interest.
   - **Contour-Based Method**: Vertebral surface contours were analyzed to assess local vertebral orientation. Using the vertebral segmentations, superior and inferior end-plates were approximated by identifying points with the greatest Euclidean distance from the centroid while maintaining anatomically correct positioning. The angle was then calculated using the dot product between these end-plate approximations.

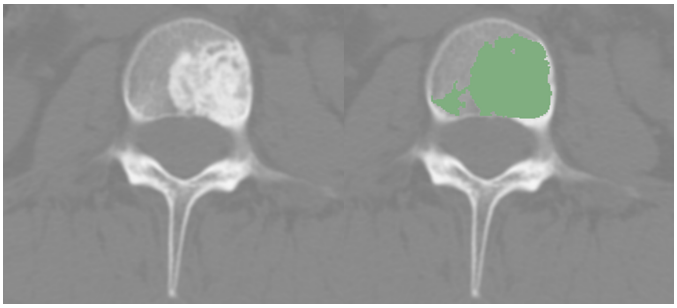

Fig. 2. Example of blastic segmentation used to calculate blastic percentage used for quantifying the bone lesion quantitative image biomarkers

Alignment metrics were derived as follows:

- **Coronal Angle (A_C)**: Computed as the average of coronal angles measured from both the spline-based and contour-based methods:
$$A_C = \frac{A_C^{spline} + A_C^{contour}}{2}$$
where A_{CF}^{spline} is the coronal angle from the spline-based method and $A_{CF}^{contour}$ is the coronal angle from the contour-based method.
- **Relative Sagittal Angle (A_S)**: Defined as the absolute difference between the expected reference angle [20]—which represents the normal sagittal curvature of the spine at a given vertebral level, derived from normative data in patients without spinal deformity— and the measured average sagittal angle. A difference with a normative value was used rather than the angle directly because the spine has a characteristic shape with the sagittal angle being dependent on the spinal level. It is the deviation from the normal that is related to mechanical stability:
$$A_S = |A_{ref} - A_{measured}|$$
where A_{ref} is the reference angle [20] for the specific vertebral level and $A_{measured}$ is the average of sagittal angles from the spline-based and contour-based methods:
$$A_{measured} = \frac{A_{SF}^{spline} + A_{SF}^{contour}}{2}$$

Alignment abnormalities were classified using established thresholds [21]:

- Scoliotic deformity: $A_C > 10°$
- Kyphotic deformity: $A_S > 45°$

This quantitative approach ensures that both coronal and sagittal deviations are objectively assessed using established research-based reference values, eliminating the subjectivity inherent in visual alignment assessment.

D. Machine Learning Model Development

The dataset was preprocessed by handling missing values (mode/mean imputation) and applying StandardScaler for feature normalization. Synthetic Minority Over-sampling Technique (SMOTE) was used to mitigate class imbalance (fracture cases: 18.3%).

Seven machine learning models were used; Logistic Regression, Random Forest, XGBoost, Support Vector Machine (SVM), Gradient Boosting, AdaBoost, and a Neural Network. Model hyperparameters were optimized using grid search, and training was performed using an 80/20 stratified train-test split. Performance evaluation was conducted using 5-fold cross-validation, with the F1-score, sensitivity, specificity, and balanced accuracy as primary metrics.

E. Performance Assessment

Model evaluation was conducted using accuracy, precision, recall (sensitivity), specificity, F1-score, and balanced accuracy to assess predictive performance. Given the imbalance in fracture and non-fracture cases, balanced accuracy and the F1-score were prioritized to provide a fair assessment of classification performance.

To ensure model stability and generalizability, 5-fold cross-validation was performed on the training set, reporting mean performance metrics across folds. Stratified sampling was used to maintain class distribution across splits.

Features used for prediction of fractures were quantitative versions of the SINS elements. Quantitative features were developed based on SINS because of its development through clinical consensus, wide adoption and connection to clinical descision making. Features quantified and used for prediction were all linked to the original SINS elements. Feature importance was computed using Gini importance (mean decrease in impurity) from a Random Forest classifier trained on the one-hot encoded training data. To account for categorical features split into multiple dummy variables, importances were aggregated back to their original feature names before ranking.

To compare machine learning models with SINS, model predictions were evaluated against overall SINS-based classifications using standard classification metrics. Agreement between the model outputs and SINS categories was assessed using Cohen's kappa coefficient.

IV. Results

A. Model Performance

Table II summarizes the comprehensive performance metrics for all models and SINS. Among the models, Random Forest achieved the most balanced performance, with an accuracy of 73.0%, sensitivity of 64.0%, specificity of 76.0%, and an F1-score of 47.0%. The Neural Network demonstrated the highest sensitivity (73.0%) but lower specificity (49.0%), representing an approach that prioritizes fracture detection over false positive reduction. While the AdaBoost model exhibited accuracy (70.0%) and specificity (80.0%), its lower sensitivity (27.0%) indicates some limitations in detecting all fracture cases. Similarly, Gradient Boosting demonstrated accuracy (72.0%) but suffered from very low sensitivity (27.0%), indicating same difficulty in identifying fracture cases.

The SINS manual assessment, which served as a clinical baseline, yielded an accuracy of 61.3%, specificity of 66.9%, and sensitivity of 36.4%. The Random Forest model demonstrated a 36% improvement in balanced accuracy compared to SINS (72.0% vs. 51.7%), supporting its enhanced predictive power. This improvement is clinically significant, as it could potentially lead to earlier intervention for patients at risk of post-SBRT vertebral fractures.

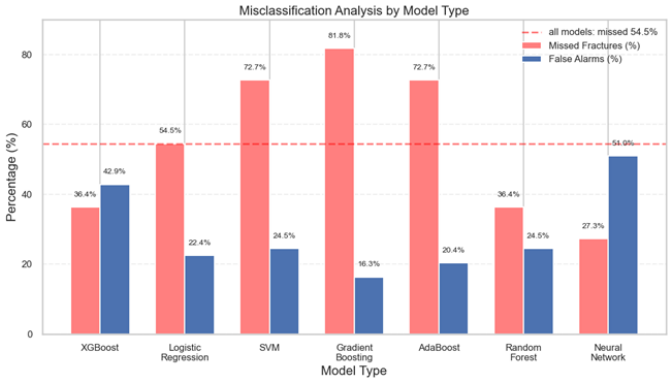

Fig. 3. Misclassification Analysis by Model Types

B. Sensitivity Challenges and Misclassification Analysis

Sensitivity remained a challenge across most models, with values ranging from 9.0% (Gradient Boosting) to 73.0% (Neural Network). Despite addressing class imbalance through appropriate techniques, the distribution disparity (81.7% non-fracture vs. 18.3% fracture cases) remained a limiting factor. The Neural Network demonstrated the highest sensitivity (73.0%), followed by Random Forest and XGBoost (both 64.0%), all showing substantial improvement over the SINS manual assessment baseline (36.4%).

A detailed analysis of misclassification patterns (Fig. 3) revealed that certain fracture cases were consistently missed by multiple models, suggesting inherent challenges in their feature representation. The Neural Network achieved the lowest missed fracture rate (27.3%) but produced the highest false alarm rate (51.0%), illustrating its bias toward fracture detection at the expense of specificity. While Random Forest achieved a balanced compromise between missed fractures (36.4%) and false alarms (24.5%), other models showed distinct trade-offs: XGBoost, Logistic Regression, and SVM all missed over

TABLE II
PERFORMANCE METRICS OF MACHINE LEARNING MODELS FOR VERTEBRAL FRACTURE PREDICTION

Metric	Random Forest	Logistic Regression	XGBoost	SVM	AdaBoost	Gradient Boosting	Neural Network	SINS Manual
Accuracy	0.730	0.720	0.580	0.670	0.700	0.680	0.530	0.613
Precision	0.370	0.310	0.250	0.200	0.230	0.100	0.240	0.198
Sensitivity	0.640	0.450	0.640	0.270	0.270	0.090	0.730	0.364
Specificity	0.760	0.780	0.570	0.760	0.800	0.820	0.490	0.669
F1-Score	0.470	0.370	0.360	0.230	0.250	0.100	0.360	0.256
Kappa	0.305	0.196	0.130	0.024	0.064	-0.096	0.122	0.025
Balanced Accuracy	0.700	0.620	0.610	0.520	0.540	0.460	0.610	0.517

Note. ML models compared against SINS manual assessment. RF: Random Forest, LR: Logistic Regression, SVM: Support Vector Machine. Random Forest achieved the highest balanced accuracy and sensitivity, demonstrating a 36% improvement over SINS baseline.

45.0% of fractures, while AdaBoost and Gradient Boosting both achieved high specificity but at the cost of reduced sensitivity (both missing over 70.0% of fractures). Misclassified fractures often exhibited subtle morphology, mixed lytic-blastic composition, and minimal spinal malalignment, making detection challenging.

The performance metrics highlight the varying strengths of different algorithms. Random Forest and Logistic Regression both achieved the highest accuracy (73.0%), while AdaBoost demonstrated the highest specificity (80.0%) and Neural Network achieved the highest sensitivity (73.%). Notably, Random Forest achieved the best balanced accuracy (70.%) and F1-score (47.0%), suggesting its superior ability to handle the class imbalance problem, while Neural Network offered complementary strengths in fracture detection. In comparison, the SINS manual assessment achieved lower performance across most metrics (balanced accuracy: 51.7%, F1-score: 25.6%).

C. Feature Importance and Predictive Factors

Feature importance analysis identified key predictors of vertebral fracture risk (Fig. 4). The top features were:

1) Spinal alignment: coronal angle (22.7%) and relative sagittal alignment (18.5%)
2) Tumour composition: lytic percentage (21.9%) and blastic percentage (17.5%)
3) Vertebral Collapse: Percentage of Vertebral collapse (10.4%)

SINS parameters (location: 2.2%, posterolateral involvement: 2.8%) showed lower importance.

D. Cross-Validation Stability

Five-fold cross-validation was used to evaluate the stability and generalizability of each model across different data splits (Table III). Modle performance showed moderate variability, with Neural Network achieved the highest mean (71.7% ± 9.4%), followed by Gradient Boosting (70.3% ± 3.6%). but also exhibiting the greatest spread, while Gradient Boosting demonstrated more consistent results (mean 70.3% ± 3.6%). Balanced accuracy, which accounts for class imbalance, also varied between folds; for example, Logistic Regression had a relatively lower balanced accuracy but remained stable across splits. This variability highlights the influence of limited data

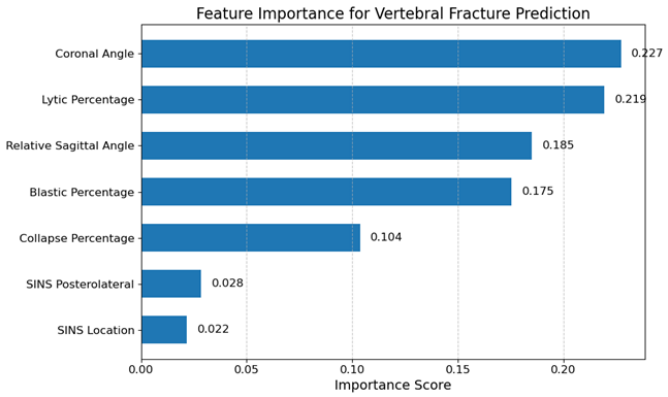

Fig. 4. Feature Importance of the model

and underscores the need for robust, generalizable models. While SINS-based predictions are known to have inter-observer variability (e.g., ICC $\approx$ 0.8 [17]), the quantitative features used by the ML models showed consistent predictive capacity across folds, suggesting a more reproducible framework for fracture risk estimation.

TABLE III
CROSS-VALIDATION RESULTS FOR MACHINE LEARNING MODELS

Model	Mean Acc.	Std Dev	Min Acc.	Max Acc.
Neural Network	0.717	0.094	0.583	0.833
Gradient Boosting	0.703	0.036	0.683	0.767
AdaBoost	0.700	0.066	0.633	0.800
Logistic Regression	0.640	0.057	0.583	0.717
SVM	0.583	0.037	0.533	0.633
Random Forest	0.593	0.072	0.483	0.667
XGBoost	0.480	0.064	0.417	0.567

Note. Cross-validation was performed using 5-fold stratified sampling.

E. Summary of Key Findings

1) Random Forest achieved the best balanced performance on test data (balanced accuracy: 70.0%), while Neural Network offered the highest sensitivity (73.0%);
2) Quantitative imaging biomarkers, particularly spinal alignment (coronal angle: 22.7%) and tumour composi-

tion (lytic percentage: 21.9%), outperformed traditional SINS parameters; and

3) Cross-validation confirmed that machine learning approaches consistently surpass conventional SINS assessment in predictive performance.

V. DISCUSSION

This study aimed to assess the utility of quantitative imaging features combined with machine learning models in predicting VCFs following SBRT and to compare their performance with the SINS. Our findings indicate that while traditional clinical assessments such as SINS provide a structured approach to evaluating fracture risk, quantitative imaging features combined with machine learning models offer a data-driven, alternative with improved predictive performance.

A. Model Performance and Clinical Relevance

The combination of quantitative image features and the random forest classifier showed much greater sensitivity over SINS allowing many more potential fractures to be identified. This can lead to modification in patient treatment, possibly including prophylactic mechanical stabilization, or lower radiation dose. The clinical need being addressed hear requires a balance of performance in terms of both sensitivity and specificity. This is why we suggest that the random forest predictions are the best for this clinical need. Predicting many false positive fractures could lead to unnecessary mechanical stabilization procedures for patients, and if radiation dose is lowered or another minimally invasive technique is used this could lead to less effective local tumour control. Conversely missing many fractures could lead to greater pain, and the possibility of neurological compromise and decreased quality of life, and the need for more invasive surgical interventions because there is a trade off to be made. This pattern likely stems from the limited dataset size (300 vertebral segments), which particularly challenges deep learning models, and class imbalance issues despite SMOTE augmentation.

B. Comparison with SINS

Machine learning models consistently outperformed SINS across evaluation metrics (Figure. 5). Compared to a prior approach that used manually derived, semi-quantitative features, the integration of quantitative imaging biomarkers in this study led to improved performance. For example, Random Forest showed a 36% gain in balanced accuracy over SINS (70.0% vs. 51.7%), while the Neural Network nearly doubled sensitivity (73.0% vs. 36.4%). These improvements demonstrate the added value of quantitative features in enhancing the objectivity, reproducibility, and precision of fracture risk prediction.

C. Cross-Validation and Model Stability

Cross-validation revealed interesting stability patterns: Neural Network achieved highest mean accuracy (71.7%) but with considerable variability (±9.4%), while Gradient Boosting showed more consistent performance (70.3% ±3.6%). This variability highlights a critical clinical implementation challenge: models with higher sensitivity may demonstrate lower consistency across diverse patient populations. For clinical translation, model selection should prioritize the specific clinical need—whether maximizing sensitivity, reliability, or balancing both. External validation with larger multi-institutional datasets remains essential before clinical deployment..

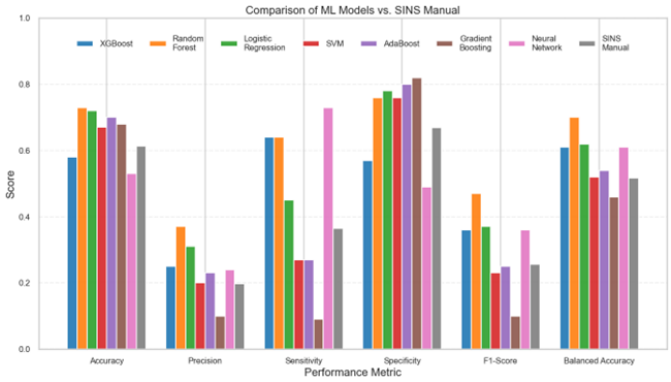

Fig. 5. ML Models Comparison with SINS

D. Predictive Features

Our feature importance analysis identified spinal alignment and tumour composition—specifically coronal angle (22.7%) and lytic percentage (21.9%)—as the strongest predictors of post-SBRT vertebral fractures. These findings align with prior studies demonstrating that tumour-induced bone destruction (e.g., lytic lesions) and structural deformities such as vertebral malalignment are associated with significantly increased fracture risk. For example, hazard ratios ranging from 3.0 to 11.1 have been reported for patients with severe malalignment and compromised vertebral body integrity following SBRT [22]. While SINS and its components, including MRI-assessed bone quality and clinical indicators like mechanical pain, have shown prognostic relevance in previous studies, our results suggest that objective, CT-derived quantitative features may provide more consistent predictive value—particularly in automated models where subjective interpretation is minimized.

E. Clinical Implications

Future research should focus on developing specialized models that prioritize the most predictive features identified in this study—spinal alignment metrics and tumour composition characteristics—while optimizing integration into clinical workflows. Additionally, external validation across diverse patient populations is essential to ensure model generalizability and clinical reliability before implementation.

VI. CONCLUSION

Our findings demonstrate the feasibility of machine learning-based fracture risk assessment as a superior alternative to the current clinical standard scoring systems like SINS. Random Forest showed a 36% improvement in balanced accuracy over SINS, with a 75% increase in sensitivity. Spinal alignment

metrics and tumour burden were the strongest predictors. Cross-validation revealed some variability in model performance across different data splits, however, even with varying performance the machine learning models consistently outperformed SINS assessment. This investigation is a step towards an enhanced, quantitative, and objective method of assessing risk of VSF following SBRT. Ultimately this approach can be used to inform clinical decision making by identifying those subjects most likely to benefit from mechanical stabilization prior to or soon after SBRT.

REFERENCES

[1] Y. R. Kim, C. H. Lee, S. H. Yang *et al.*, "Accuracy and precision of the spinal instability neoplastic score (sins) for predicting vertebral compression fractures after radiotherapy in spinal metastases: A meta-analysis," *Scientific Reports*, vol. 11, p. 5553, 2021.

[2] M. Guckenberger, F. Mantel, P. C. Gerszten, J. C. Flickinger, A. Sahgal, D. Letourneau, I. S. Grills, M. Jawad, D. K. Fahim, J. H. Shin *et al.*, "Safety and efficacy of stereotactic body radiotherapy as primary treatment for vertebral metastases: a multi-institutional analysis," *Radiation Oncology*, vol. 9, no. 1, pp. 1–8, 2014.

[3] C. Gui, X. Chen, K. Sheikh, L. Mathews, S.-F. L. Lo, J. Lee, M. A. Khan, D. M. Sciubba, and K. J. Redmond, "Radiomic modeling to predict risk of vertebral compression fracture after stereotactic body radiation therapy for spinal metastases," *Journal of Neurosurgery: Spine*, vol. 36, pp. 294–302, 2022.

[4] S. Faruqi, C. L. Tseng, C. Whyne *et al.*, "Vertebral compression fracture after spine stereotactic body radiation therapy: A review of the pathophysiology and risk factors," *Clinical Neurosurgery*, vol. 83, p. 314–322, 2018.

[5] G. Klein, C. Whyne, and A. S. et al., "Metastatic vertebrae segmentation for use in a clinical pipeline," *Lecture Notes in Computer Science*, vol. 11963, pp. 15–28, 2020.

[6] A. Sahgal, E. G. Atenafu, S. Chao, A. Al-Omair, N. Boehling, E. H. Balagamwala, M. Cunha, I. Thibault, L. Angelov, P. Brown, J. Suh, L. D. Rhines, M. G. Fehlings, and E. Chang, "Vertebral compression fracture after spine stereotactic body radiotherapy: A multi-institutional analysis with a focus on radiation dose and the spinal instability neoplastic score," *Journal of Clinical Oncology*, vol. 31, no. 27, pp. 3426–3431, 2013. [Online]. Available: https://ascopubs.org/doi/10.1200/JCO.2013.50.1411

[7] C. Fisher, C. Dipaola, and T. e. a. Ryken, "A novel classification system for spinal instability in neoplastic disease: An evidence-based approach," *Spine*, vol. 35, no. 22, pp. E1221–E1229, 2010.

[8] D. D. Shi, L. M. Hertan, T. C. Lam, S. Skamene, J. H. Chi, M. Groff, C. H. Cho, M. L. Ferrone, M. Harris, Y.-H. Chen, and T. A. Balboni, "Assessing the utility of the spinal instability neoplastic score (sins) to predict fracture after conventional radiation therapy (rt) for spinal metastases," *Practical Radiation Oncology*, vol. 8, no. 5, pp. e285–e294, 2018.

[9] A. L. Versteeg, J. M. Velden, H. M. Verkooijen, M. Vulpen, F. C. Oner, C. G. Fisher, and J.-J. Verlaan, "The effect of introducing the spinal instability neoplastic score in routine clinical practice for patients with spinal metastases," *The Oncologist*, vol. 21, no. 1, pp. 95–101, 12 2015. [Online]. Available: https://doi.org/10.1634/theoncologist.2015-0266

[10] R. S. Gunninski, F. Cuccia, F. Alongi *et al.*, "Efficacy and safety of sbrt for spine metastases: A systematic review and meta-analysis for preparation of an estro practice guideline," *Radiotherapy and Oncology*, 2023.

[11] H. Kim, H. Pyo, H. C. Park, D. H. Lim, J. I. Yu, W. Park, Y. C. Ahn, D. H. Choi, D. Oh, J. M. Noh, W. kyung Cho, G. S. Yoo, S. H. Jung, E.-S. Kim, S.-H. Lee, S.-J. Park, and C.-S. Lee, "Clinical and dosimetric risk factors for vertebral compression fracture after single-fraction stereotactic body radiation therapy for spine metastases," *Journal of Bone Oncology*, vol. 28, p. 100368, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2212137421000221

[12] K. Liu, S. Qin, J. Ning, P. Xin, Q. Wang, Y. Chen, W. Zhao, E. Zhang, and N. Lang, "Prediction of primary tumor sites in spinal metastases using a resnet-50 convolutional neural network based on mri," *Cancers*, vol. 15, p. 2974, 2023.

[13] C. Whyne, M. Hardisty, F. Wu, and T. Skrinskas, "Quantitative characterization of metastatic disease in the spine: Part ii. histogram-based analyses," *Medical Physics*, vol. 34, no. 8, pp. 3279–3282, 2007.

[14] M. Hardisty, T. Wright, M. Campbell, M. Burke, E. Atenafu, M. Cawricz, C. Caldwell, A. Yee, A. Sahgal, and C. Whyne, "Ct based quantitative measures of the stability of fractured metastatically involved vertebrae treated with spine stereotactic body radiotherapy," *Clinical & Experimental Metastasis*, vol. 37, no. 5, pp. 575–584, Oct 2020, epub 2020 Jul 8. [Online]. Available: https://doi.org/10.1007/s10585-020-10049-9

[15] N. G. Nia, E. Kaplanoglu, and A. Nasab, "Evaluation of artificial intelligence techniques in disease diagnosis and prediction," *Discover Artificial Intelligence*, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9885935/

[16] X. Wang, D. Zhou, Y. Kong *et al.*, "Value of 18f-fdg-pet/ct radiomics combined with clinical variables in the differential diagnosis of malignant and benign vertebral compression fractures," *EJNMMI Research*, vol. 13, p. 89, 2023.

[17] S. Fox, M. Spiess, L. Hnenny, and D. R. Fourney, "Spinal instability neoplastic score (sins): Reliability among spine fellows and resident physicians in orthopedic surgery and neurosurgery," *Global Spine Journal*, vol. 7, no. 8, pp. 744–748, December 2017. [Online]. Available: https://journals.sagepub.com/doi/10.1177/2192568217697691

[18] G. Klein, M. Hardisty, C. Whyne, and A. L. Martel, "Vertdetect: Fully end-to-end 3d vertebral instance segmentation model," *arXiv preprint arXiv:2311.09958*, 2023.

[19] G. Klein, I. Carreno, J. Finkelstein, A. Sahgal, J. Detsky, C. M. Whyne, A. Martel, and M. Hardisty, "Advancements in metastatic spinal angle assessment: Clinical application of a novel 3d vertebrae detection algorithm," in *Orthopaedic Research Society Annual Meeting*, Long Beach, California, United States of America, Feb 2023, conference Date: 2024/2, Abstract Accepted.

[20] M. P. Abdel, W. S. Bodemer, and P. A. Anderson, "Supine thoracolumbar sagittal spine alignment: Comparing computerized tomography and plain radiographs," *Spine*, vol. 37, no. 4, pp. 340–345, Feb 2012.

[21] G. Fon, M. Pitt, and J. Thies, AC, "Thoracic kyphosis: range in normal subjects," *American journal of roentgenology (1976)*, vol. 134, no. 5, pp. 979–983, 1980.

[22] M. Hardisty, J. Finkelstein, J. Detsky, I. Thibault, and A. Sahgal, "Vertebral compression fracture after spine sbrt," in *Spine Radiosurgery*, P. C. Gerszten, A. Sahgal, and S. I. Ryu, Eds. Springer, 2017, pp. 241–257.