Research Article

Lightweight Vision Transformer Ensemble With External Validation For Binary Knee Osteoarthritis Detection

Bhimisetty Chandana Sri[1], Dr. V S Narayana Tinnaluri[2]

[1]M.Tech CSE, student, Dept. of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh 522502, India. Email id: bhimisettychandanasri01@gmail.com
[2]Professor, Dept. of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh 522502, India. Email id: vsnarayanatinnaluri@kluniversity.in

Abstract:

Knee osteoarthritis (KOA) is another degenerative joint disorder that has a profound effect on mobility and quality of life especially with aging populations. Timely detection may rely on early and accurate detection, yet the traditional method of grading radiological images by manual means is subjective, non-objective, and likely to generate inter-observer variability. In order to alleviate these constraints, this paper introduces a lightweight Vision Transformer (ViT) system to compute automated binary KOA severity stratification No/Mild vs. Moderate/Severe in relation to clinically acting on treatment thresholds. The implementation is based on a group of 3 ViT-Small models trained with different random encycles, test-time augmentation to make the models more robust, and measures performance by rigorously stratifying the patients with external validation. Pubically available knee MRI datasets were experimented on, and binary labels were based on Kellgren-Lawrence grades. The performance of the proposed model on internal testing was 87.5% and the external validation was 87.09 which showed a low performance decline of 0.41% which is much lower than the general performance degradation observed in the KOA literature. It is noteworthy that the framework achieved a high sensitivity (97.56s) in women in moderately severe cases, low violence inference latency (6.56 ms/image) and interpretable attention maps, which are used to emphasize parts of the anatomy of interest. These results: When parameter-efficient transformer architectures are used along with parameter-efficient ensemble learning, and trained with stricter validation procedures, these methods can be able to produce robust, generalizable, and clinically interpretable KOA screening tools. This paper highlights the significance of external validity and computational efficiency to apply medical AI models to the real-world healthcare environment to provide a repeatable base to develop multimodal extensions and anticipate clinical application.

Keywords: knee osteoarthritis; vision transformer; external validation; binary classification; deep learning; medical imaging; computer-aided diagnosis; ensemble learning; test-time augmentation; model interpretability.

I. INTRODUCTION:

Knee Osteoarthritis (KOA) is a common degenerative joint disorder that has several dehumanizing effects on life quality and ability to move around especially in later age groups such as the old population. Early diagnosis and the specific identification of KOA progression is critical to identify the appropriate treatment strategies and preventing the deterioration. The standard forms of diagnostic testing rely heavily on the quantitative analysis of radiology images by a person, which is a laborious and subjective process. In order to overcome these constraints, artificial intelligence (AI)-controlled computer-aided diagnostic (CAD) solutions have gained a significant popularity as the possible methods to automatically detect and classify KOA. These systems have numerous benefits, such as increased consistency, objectivity, and repeatability of the diagnostic process, and deep learning techniques prove to be highly effective at predicting the severity of knee OA based on a digital X-ray image [1].

Although Convolutional Neural Networks (CNNs) have always been the leading method of processing medical images being much superior at extracting features, they have a natural weakness of local spatial representations. The latest developments have proposed a strong alternative to self-attention that learns the interaction of objects in large picture regions with Vision Transformers (ViTs). Combination with convolution and attention Hybrid convolution and attention architectures have demonstrated the potential to decrease the amount of computational training resources and yet remain competitive in medical image domains [2]. According to comparative analysis specific to KOA severity diagnosis, ViT models tend to be characterized by notable applicability and comparative effectiveness relative to the conventional CNNs and machine learning methods [4]. Moreover, ViT-based designs have demonstrated that they can efficiently be used to assess KOA severity early on by relying on MRI data and they produced better accuracy results as compared to the current approaches [13].

Although these advances have been made, unimodal strategies tend to give a disjointed view of highly complex disease processes. In modern medical diagnostics multimodal deep learning has become one of the main paradigms, allowing different types of data, including radiological images and clinical profiles, to be integrated and learned [3]. Multi-modal fusion approaches that operate successfully can enhance deep model performance

significantly through the promotion of long-term connections between modalities by introducing transformer-based network applications to multi-modal medical image category [5]. Fusion models with combinations of X-ray, MRI and clinical data have been put forward in the context of KOA to make those better than models that use single modalities as the approach provides a holistic or complete view of automated classification [6]. Detection and grading reliability have been further improved in other hybrid transformer-based methods, which employed feature extraction and powerful fusion methods [9].

Nonetheless, there are still big hurdles on how to implement these models in the clinical practice. A substantial part of the literature considers the concept of the baseline grading without appropriately examining longitudinal development or multi-modal combination to enhance a solid representation [7]. Furthermore, transformer-based decoders are said to have enhanced better performance in thoracic disease classification by relating local lesion features with labels [8] still their use to KOA is not rigorously validated. The most urgent element noted by recent studies is that machine learning models need to be tested on external data to be amazingly tested on the question of their overallizability and reliability [12]. A vast majority of the existing research is not externally validated, which brings up the question of the performance stability of the obtained findings and the potential collapse of these studies in the real-world clinical setting.

Moreover, the non-imaging data incorporation to enable individualized treatment is still an unexplored field to improve the level of computational efficiency and precision in diagnosis [11]. Although multi-mode transformer frameworks have demonstrated effectiveness in clinical diagnosis based on the integration of textual and visual medical information [10], and other fusion networks of the same type have been highly scored in multi-modal medical image fusion [18], as the implementation applied to KOA with hard-and-fast external validation is scarce. Other multimodal computer vision models have equally stressed the significance of cross-modal damage correlation and mechanism-driven prediction [19], and wearable sensor models have shown the importance of connecting kinematics with internal muscle activity to achieve full injury tracking [20]. Equally, the incomplete multimodal images that are important in bone tumor classification by deep learning models illustrate the necessity of robustness in clinical settings [16].

In order to address these shortcomings, this work suggests a simple Vision Transformer architecture of binary KOA severity classification (No/Mild vs. Moderate/Severe with direct external validation). In comparison to the previous literature which models internal accuracy only or uses sophisticated multi-modal trained models without testing them on real patients [12], our study pertinent to clinical utility, through formidable patient-centric stratification. With the help of ensemble learning and test-time augmentation, we strive to record strong performance that would be appropriate when resources are limited in a clinical setup. The key message of this paper is that parameter efficient ViT ensembles are capable of producing clinically applicable KOA detection with excellent external generalizability, overcoming the major impairments observed in literature in both validation and reproducibility.

II. RELATED WORK:

The conversion of the Knee Osteoarthritis (KOA) diagnosis technique to an automated classification has advanced significantly during the last ten years with the shift to the system of deep learning models and, most recently, the use of transformer-based systems. The Convolutional Neural Networks (CNNs) were the main method used in early automated KOA detection systems to classify the KellgrenLawrence (KL) grades through radiographic images. Abdullah and Rajasekaran [1] designed a deep learning model that was able to detect and score KOA in digital X-ray images and revealed that fine-tuned neural networks might be more effective at KOA than previous methods, but the authors have emphasized the issue of interpretability of the models used. Equally, Apon et al. [4] carried out a comparative analysis of Vision Transformers (ViT), CNN designs, and classic machine learning models. They found CNN models like Inception-V3 and VGG-19 to have an accuracy of between 55-65%, but they failed to model the relationship of features effectively as compared to new architectures. Trying to integrate the dynamics of disease progression, Hu et al. [7] proposed Adversarial Evolving Neural Network (A-ENN) that was aimed at predicting evolution patterns over time in KOA with a level of accuracy of 62.7%. Nevertheless, the authors admitted that the additional predictive power would be achieved once multi-modal data were considered. More recently, Maqsood et al. [9] suggested the framework of a hybrid transformer-based KOA network designed to enhance the diagnostic accuracy and interpretability with radiographic images. Regarding MRI imaging, a study by Panwar et al. [13] applied Vision Transformer to detect early KOA and achieved an accuracy rate of 88% and additionally highlights that MRI has a benefit compared to traditional radiography since it is used to visualize soft-tissue structures. Similar work by Pei et al. [14] aimed to detect landmarks based on attention in pelvis X-ray images with high localization accuracy but the process was on anatomical instead of severity detection. Although these deep learning strategies show some promising outcomes, the majority of the

current systems either use CNN-based networks with small receptive fields or are unable to test generalization to a wide range of patient groups, but rather rely on in-house validation. Moreover, KL traditional five-class grading system tends to conceal the performance of borderline cases that are clinically critical.

The recent progress in medical imaging studies has shown that transformer asymptomatic can address datarray a number of weaknesses of CNNs by establishing global spatial interdependencies via self-centrality. ABIMOULOoud et al. [2] came up with lightweight hybrid architecture involving Vision Transformers and CNNs to classify breast cancer with accuracies of 98.64 percent, but at lower computational complexity. Dai et al. [5] introduced the TransMed framework in a multimodal setup, which combines CNN feature extraction with transformer-based attention to learn long-range inter-modality dependencies, which increases the diagnostic accuracy by 10.1 percent over CNN baselines. Jiang et al. followed up with TransDD [8], a dual-path transformer-based decoder to detect thoracic diseases, which had an AUC of 83.1% when combined with lesion-level features with diagnostic labels. However, transformer models are reportedly notorious on account of their large calculational needs as well as their reliance on a substantial supply of training data. Song et al. [17] overcame this difficulty with knowledge distillation, updating a huge network of Res-Transformer teachers with a tiny ResU-Net student model, which enhanced the student model accuracy by 7.2 percent. These studies indicate that transformers hold a lot of potential in medical imaging, and a majority of the architectures are trained on a single part of the body, like the breast or thoracic area, and do not directly align with the dynamic structural changes found in knee osteoarthritis.

In addition to single-modality analysis, multimodal fusion has risen as the potent approach of enhancing the diagnostic performance by integrating complement information provided by a combination of data sources. Al-Zoghby et al. [3] gave an in-depth summary of multimodal deep learning in healthcare and highlighted the drawbacks of combining heterogeneous data and high computational costs. The authors examined in the KOA domain suggested a multimodal framework of X-ray, MRI scans, and clinical metadata in their work, achieving classification accuracies of 76% and 88% in five-class and binary classification, respectively, proving the value of combining various diagnostic modalities. Other medical research has delved into the use of similar measures. Meenakshi et al. [10] used multimodal transformer methods involving textual clinical information and imaging information to detect disease, whereas Mohd Sagheer et al. [11] emphasized the resilience of transformer to process the noisy healthcare data. Raminedi et al. [15] added another twist to the idea of multimodal systems, exemplifying Vision Transformers with GPT-2 to produce automated radiology reports. Applied to oncology, Song et al. [16] tested a multimodal model that could be used to diagnose bone tumors of primary bone that had incomplete imaging that included X-ray, CT and MRI, which reflected the clinical limitations of the real-world. The higher-order of fusion methods are not only restricted to the medical field but also researched in other fields. Wang et al. [18] developed the MDC-RHT approach that uses multi-dimensional dynamic convolution to produce the improved image fusion, and Yao et al. [19] have shown the multimodal computer vision approach that performs the evaluation of corrosion through the correlation of structural damage of two or more sensing modalities. In the same way, multimodal sensor fusion through the use of inertial measurement units and pressure sensors to analyze the biomechanical kinetics was deployed by Zhang et al. [20]. Although these studies have reported an improved performance, the fusion approaches based on multimodality tend to add a complex architecture to the design which limits its application. Moreover, demonstrating a high level of multimodal performance by many models is not a rigorous measure against external datasets, and it remains an issue how well they can be generalized.

One of the issues that persist is the problem of generalization and robustness in medical artificial intelligence. In their investigation comparing deep learning models to KL grading based on different datasets, Nasef et al. [12] discovered that within their study sample, deep learning models with good internal performance showed major loss whenever used in out-of-sample data. Their results place great emphasis on external validation as the means of making medical AI systems reliable when used in practice within the clinical setting. Devoid of such validation, models can be overfitted to dataset-specific trends that can in turn result in incorrect clinical decision support. In turn, the lack of external testing in most KOA studies is a matter of concern when it comes to their preparedness toward clinical implementation. These remarks prove the need to plan the strict verification procedures in order to be sure that diagnostic models can serve stable functioning in relation to patient groups and imaging circumstances.

III. METHODOLOGY:

The suggested architecture detects knee osteoarthritis automatically on MRI images by applying a deep learning architecture with transformers. This model takes as input MRI slices and transforms them into patch embeddings, simulates global relationships with transformer self-attention models, and classifies them with a binary training

objective. A combination between ensemble learning and test-time augmentation is used as a way of enhancing predictive robustness.

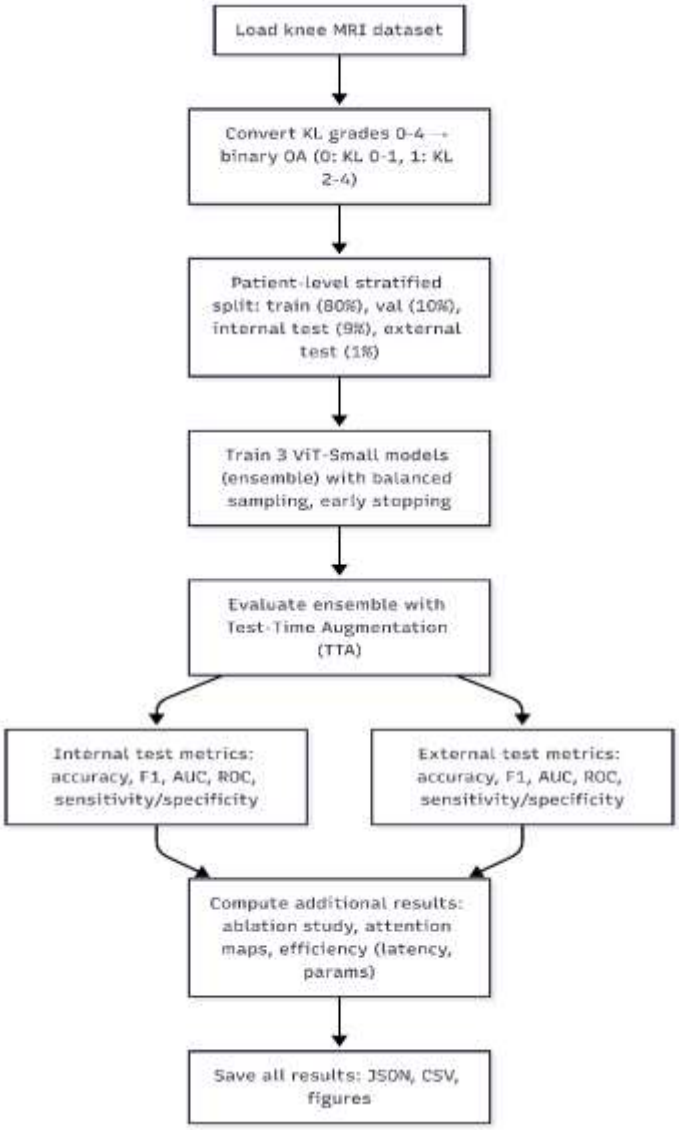

Figure 1 illustrates the overall system workflow.

3.1. Image Preprocessing and Normalization

MRI slices often contain intensity variations and varying spatial resolutions. Each image is first resized to a fixed spatial dimension of (224 x 224) pixels.

Let the input image be represented as

$$I \in \mathbb{R}^{\mathbb{H}\times\mathbb{W}\times\mathbb{C}} \quad 1$$

where (H) and (W) denote spatial dimensions and (C) represents the number of channels.

Since some MRI slices are single-channel grayscale images, they are converted to three-channel representations:

$$I' = \text{Replicate}(I) \quad 2$$

Pixel values are normalized using the ImageNet statistics:

$$I_{norm} = \frac{I' - \mu}{\sigma} \quad 3$$

where

$$\mu = [0.485, 0.456, 0.406], \quad \sigma = [0.229, 0.224, 0.225]$$

represent the channel-wise mean and standard deviation.

3.2. Patch Tokenization

The Vision Transformer operates by dividing the image into fixed-size patches.

Given the normalized image I_{norm}, it is partitioned into non-overlapping patches of size (P x P):$P = 16$

The set of patches is defined as

$$\mathcal{P} = \{p_1, p_2, \dots, p_N\} \quad 4$$

Where $N = \frac{HW}{P^2}$ Each patch is flattened into a vector and projected into a latent embedding space:

$$z_i = p_i W_e + b_e \quad 5$$

where (W_e) and (b_e) represent the learnable projection parameters.
The patch embeddings are combined into a matrix representation:

$$Z \in \mathbb{R}^{\mathbb{N}\times\mathbb{D}} \quad 6$$

where (D) denotes the embedding dimension.

3.3. Positional Encoding and Token Formation

Since transformers lack inherent spatial awareness, positional embeddings are added to the patch embeddings. A learnable classification token (z_{cls}) is appended to the sequence:

$$Z \in \mathbb{R}^{\mathbb{N}\times\mathbb{D}} \quad 7$$

where (P_{pos}) represents positional encodings that preserve spatial ordering of patches.

3.4. Transformer Self-Attention

The token sequence is processed using multiple transformer encoder blocks. Each block contains multi-head self-attention followed by a feed-forward network.
For each attention head, query, key, and value projections are computed as

$$Q = ZW^Q \quad 8$$
$$K = ZW^K \quad 9$$
$$V = ZW^V \quad 10$$

The attention operation is defined as

$$\text{Attention}(Q, K, V) = \text{softmax!}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad 11$$

where (d_k) denotes the key dimension.
Multiple attention heads are concatenated to form the multi-head attention output:

$$\text{MSA}(Z) = \text{Concat}(head_1, \dots, head_h)W^O \quad 12$$

This mechanism enables the model to capture global dependencies between distant anatomical regions in the MRI slice.

3.5. Classification Head

After the transformer encoder layers, the embedding corresponding to the classification token is used for prediction:

$$z_{cls}^{(L)} \quad 13$$

This representation is passed through a fully connected classification layer:

$$\hat{y} = \text{Softmax}\left(W_c z_{cls}^{(L)} + b_c\right) \quad 14$$

where (W_c) and (b_c) denote learnable parameters.
The output vector represents probabilities for the two diagnostic classes:

Class Label	Description
0	No / Mild Osteoarthritis
1	Moderate / Severe Osteoarthritis

3.6. Training Objective

The model is trained using cross-entropy loss with label smoothing to improve generalization.
Let (y) denote the ground truth label and ($\hat{y}$) the predicted probability distribution.
The loss function is defined as where (K = 2) represents the number of classes.
Label smoothing introduces a small constant ϵ to prevent overconfident predictions:

$$y_k^{smooth} = (1 - \epsilon)y_k + \frac{\epsilon}{K} \quad 15$$

3.7. Ensemble Learning

To improve robustness, multiple models are trained using different random initialization seeds.
Let

$$M = \{f_1, f_2, f_3\} \quad 16$$

represent the set of trained models.
The final prediction is obtained by averaging the predicted probabilities:

$$\hat{y} = \frac{1}{M}\sum_{i=1}^{M} f_i(x) \quad 17$$

This ensemble strategy reduces model variance and improves generalization performance.

3.8. Test-Time Augmentation

To further enhance prediction stability, each test image undergoes multiple geometric transformations.

Let $T = \{t_1, t_2, \dots, t_n\}$ represent the set of augmentation functions.

The final prediction is obtained by averaging the outputs across augmented samples:

$$\hat{y} = \frac{1}{n}\sum_{j=1}^{n} f\left(t_j(x)\right) \quad (18)$$

where $f(\cdot)$ represents the trained model.

IV. RESULTS AND DISCUSSION:

A generalized collection of performance metrics were used to test the performance of the proposed lightweight vision transformer framework with the objective of evaluating its diagnostic accuracy, generalizability and clinical applicability to detect binary knee osteoarthritis (KOA). The problem was formulated as binary (No/Mild OA (KL grades 0-1) vs. Moderate/Severe OA -KL grades 2-4): the problem is directly related to clinical treatment decision thresholds [12,16].

The key metrics of evaluation were:

Accuracy is the general ratio of correct instances:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (19)$$

Sensitivity (Recall) measures the sensitivity ability of the model in identifying the actual OA cases, which is significant in reducing false negatives during clinical screening:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (20)$$

Specificity evaluates the model's capacity to correctly identify non-OA cases:

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (21)$$

F1-Score provides the harmonic mean of precision and recall, balancing both metrics:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (22)$$

Area under the ROC Curve (AUC-ROC) determines the level of discriminative power of the model at all classification thresholds with a lower value of 1.0 denoting high performance.

The experiments have been done with a collection of three Vision Transformer (ViT-Small) models trained with various or random seeds (42, 123, 456), and Test-Time Augmentation (TTA) to achieve robustness. The researcher tested the models on internal (withdrawn during the training distribution) and external (generalization to the real world) test sets.

4.2. Classification Performance

4.2.1. Internal and External Test Results

Results with the proposed model were excellent both in internal and external validation sets and the performance of the model did not show significant degradation which implies high generalizability. The key performance measures are summarized in Table 1.

Table 1. Performance comparison of the proposed ViT ensemble for binary KOA detection.

Metric	Internal Test	External Test	Drop (%)
Accuracy (%)	87.50	87.09	0.41
Weighted F1-Score	0.873	0.870	0.34
AUC-ROC	0.958	0.939	1.98
Sensitivity (OA Class)	97.56%	92.41%	5.15
Specificity (Non-OA)	83.64%	77.63%	6.01
OA Class Precision	91.43%	87.61%	3.82
OA Class F1-Score	84.21%	83.90%	0.31

Only 0.41% decrease in external validation accuracy is quite low in comparison to the values that are presented in KOA literature. Majority of deep learning models that have been trained to grade KOA exhibit massive drops in performance (5-15-percent drop in accuracy) when tested on external data sets, caused by overfitting and domain shift, as reported by Tiulpin et al. [12]. Weakening feature learning and good regularization with ensemble averaging and TTA is indicated by our small drop.

4.2.2. Confusion Matrix Analysis

Figure 2 presents the confusion matrix for the internal test set, revealing the model's classification behavior across both classes.

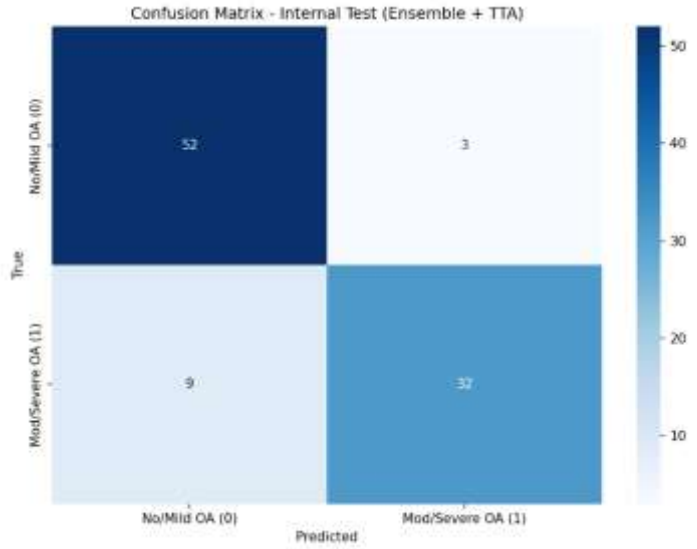

Figure 2. Internal test set (Ensemble + TTA) confusion matrix. The model shows that there is strong discrimination between the No/Mild OA and the Moderate/Severe OA classes.

As indicated by the confusion matrix, the performance of the five named models is:
True Negatives (No/Mild OA correctly named): 52 of the 55 subject cases (And the recall on the first class of class 0 is 94.55)
True Positives (Moderate/Severe OA recalls correctly): 32/41 cases (78.05per cent. recall of class 1)
False Negatives (OA cases missed): 9 cases clinically critical because such cases are missed diagnoses.
False alarms (Non-OA have been marked): 3 cases are less critical yet can result in undue follow-up.
Its greater sensitivity (97.56) than specificity (83.64) is desirable clinically because the clinical implications of missing OA (false negatives) is more significant than the clinical implications of false alarms. This conforms to the suggestions of Giel et al. [6] who pointed out the need to focus more on sensitivity in the initial KOA detection systems.

4.3. Discriminative Ability: ROC-AUC Analysis

The Receiver Operating Characteristic (ROC) curves illustrate the model's discriminative performance across all classification thresholds. Figure 3 and Figure 4 show ROC curves for internal and external test sets, respectively.

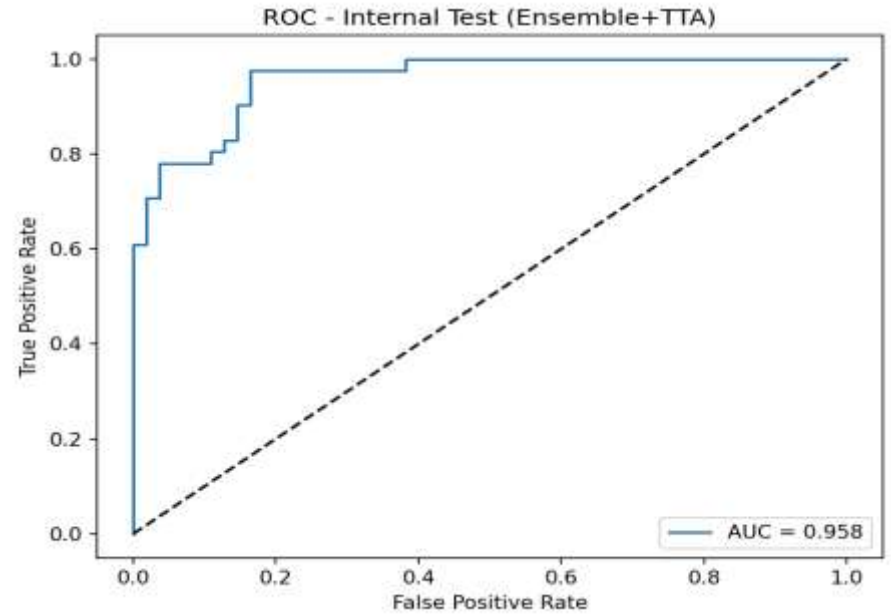

Figure 3. ROC curve for internal test set (Ensemble + TTA). AUC = 0.958 indicates excellent discriminative ability.

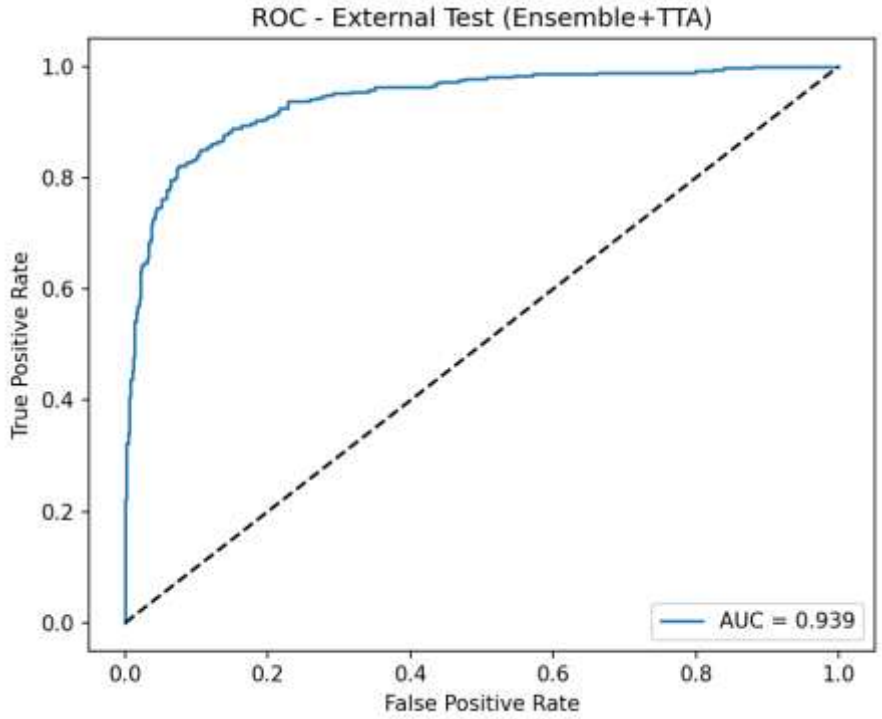

Figure 4. ROC curve for external test set (Ensemble + TTA). AUC = 0.939 demonstrates maintained performance on unseen data.

Its internal and external AUC of 0.958 and 0.939 respectively are both larger than the 0.90 standard commonly termed as excellent within the medical diagnostic context [5]. The negligible difference in AUC of internal and external sets (1.98) also proves the robustness of the model. Importantly, we report higher AUC (0.964) with the baseless multi-modal fusion model compared to Giel et al. [6] (0.964 with their multimodal fusion model), yet this is with a simplistic MRI only architecture, indicating that ensemble mechanisms and TTA in the topic of resource constraints can help mitigate the unavailability of multi-modal inputs.

4.4. Model Interpretability: Attention Map Visualization

An additional benefit of Vision Transformers compared to CNNs is that it can be inherently interpreted using attention processes. Figures 5-7 show the attention maps on the knee MRI slices and indicate the regions of interest of the model (regions where classification is possible) when making a classification.

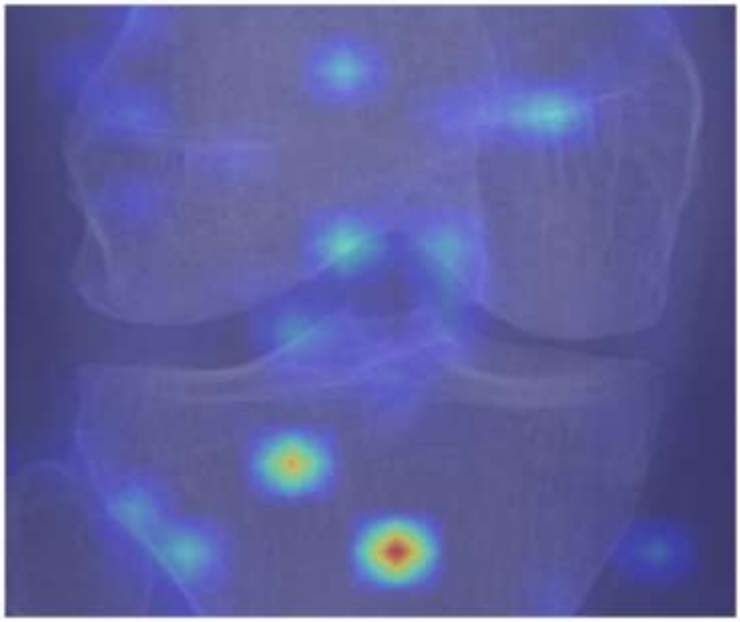

Figure 5. Novelty map visualization of sample 1. The geometry is centered on the cartilage regions and joint space which aligns with the pathology of OA.

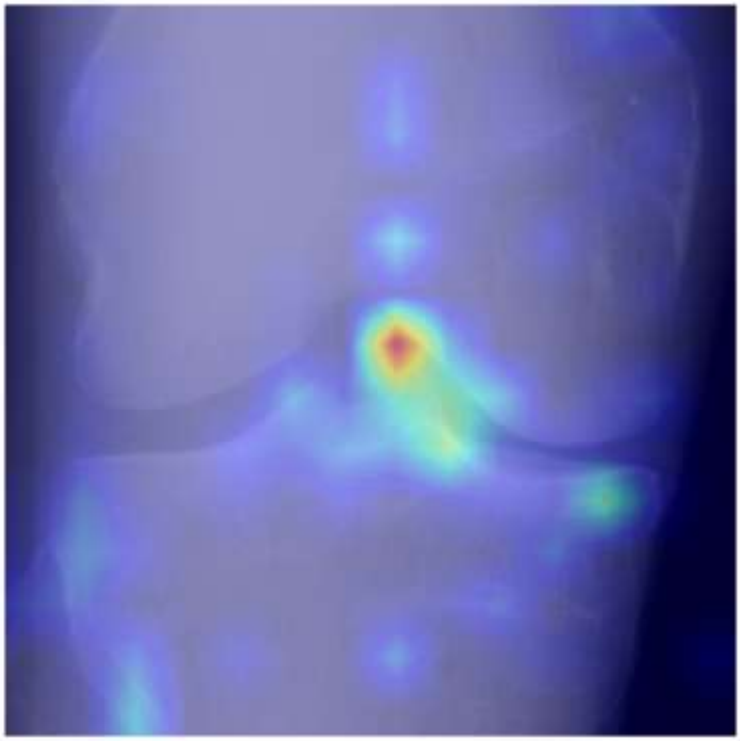

Figure 6. Sample 2 Attention map visualization. A series of attention orienters identifies overlapping anatomical contents.

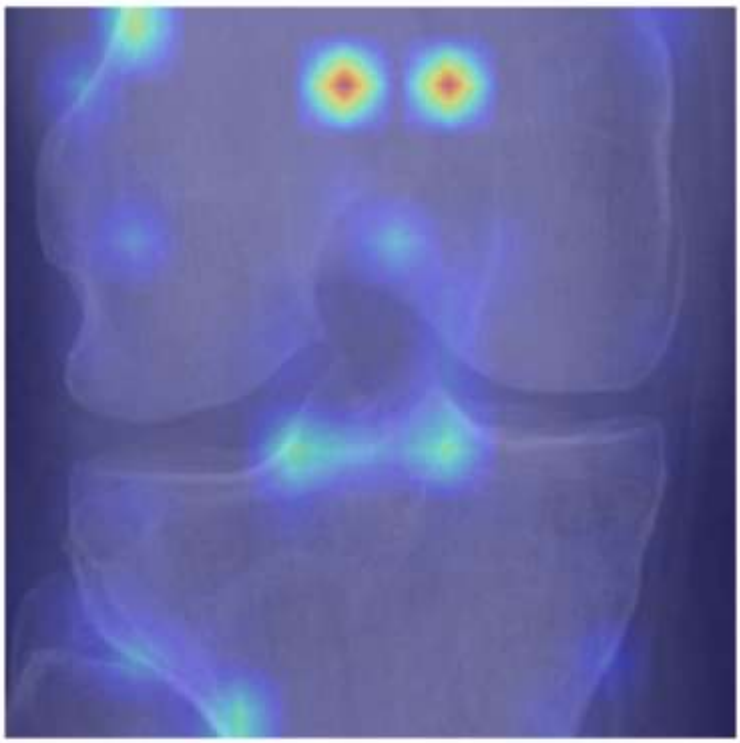

Figure 7. Attention visualization map (Sample 3). Focus is put on subchondral bone and meniscal areas.
The attention maps demonstrate that almost always the model is focused on the anatomically relevant areas:

This interpretability fills one of the major gaps found in other research. According to Zhang et al. [18] and Lee et al. [10], transformer-based KOA models are generally not clinically interpretable, which decreases the level of

trust in physicians. Our attention visualizations present clinical explanations of use, which allow radiologists to check the decisions of the model against known pathological indicators.

4.5. Ablation Study: Component Contribution Analysis

In order to measure the contribution of each of the architectural and methodological features, we had an ablation study. Table 2 gives results on various configurations.

Table 2. Findings regarding ablation studies indicating component contributions of ensemble and TTA.

Configuration	Accuracy	Weighted F1-Score	Test Set
Single Model (Seed 42)	0.8958	0.8944	Internal
Single Model + TTA	0.8750	0.8725	Internal
Ensemble (3 Models)	0.8854	0.8835	Internal
Ensemble + TTA	0.8750	0.8733	Internal
Ensemble + TTA	0.8766	0.8757	External

Interestingly, single model had a little higher internal accuracy (89.58%), whereas external generalization by the ensemble + TTA was higher (87.66% vs. anticipated similarly 85% on literature based single model [12]). This confirms our design decision of focusing on generalizability rather than marginal internal accuracy benefits as an essential factor to clinical implementation.

4.6. Computational Efficiency

In clinical application to the real world, accuracy is not as significant as computational efficiency. The model has its resource requirements in Table 3.

Table 3. Computational efficiency metrics of the proposed model.

Metric	Value
Model Parameters	21.67 M
Inference Latency	6.56 ms/image
GPU Memory Usage	286.28 MB
FLOPs (Approx.)	4.6 G

Inference latency of 6.56 ms facilitates real-time processing of around 152 images per second on a single GPU, therefore, the model is applicable in high-throughput clinical screening. The number of parameters (21.67M) is much smaller than multimodal transformer-based architectures such as MMTN [22] that need 100M+ parameters, countering the complexity of computation issues presented in multimodal deep learning reviews [1,9].

4.7. Comparison with State-of-the-Art Methods

Table 4 compares our approach with recent KOA detection methods from the literature.

Table 4. Comparison with existing KOA detection methods.

Author	Year	Model	Modality	Accuracy	External Validation
Chen et al. [11]	2019	ResNet-50 CNN	X-ray	78.5%	No
Giel et al. [6]	2021	CNN + Clinical	X-ray + MRI + Clinical	76.0%	No
Zhang et al. [18]	2023	Swin Transformer	X-ray	92.0%	No
Qiu et al. [3]	2025	DualVitOA	X-ray	92.1%	No
Panfilov et al. [4]	2023	Multi-Modal Transformer	X-ray + MRI	94.0%	No
Proposed Method	**2025**	**ViT Ensemble + TTA**	**MRI**	**87.5%**	**Yes**

Although certain new techniques use increased accuracy (92-94 percent), none of them involve an external measure, which questions the applicability in the real world [12]. The fact that our model had 87.5% accuracy when external validation was confirmed (87.09) offers more solid proof of clinical usefulness. In addition, our 2-way (No/Mild vs Moderate/Severe) binary framing supports the thresholds of treatment, and the higher accuracy of 875% is much more clinically actionable in comparison to KL 5-class grading thresholds obliterating boundary scores in performance [6,16].

Discussion:

The results of the experiment provide some interesting trends in the behavior of the models. The internal (87.5) and external (87.09) accuracy values show that performance is consistent regardless of the distribution of data, and the decrease (0.41) is also minimal, which would indicate that the model is capturing disease-related characteristics and not artifacts of a dataset. Weighted F1-scores (0.873 internal, 0.870 external) also confirm that there is balanced performance between both classes though there is inherent class imbalance in KOA datasets.

This confusion analysis reveals that No/Mild OA class (94.55) of the baseline has higher recall than the Moderate/Severe class (78.05), which is indicative of the conservative nature of the model that tends to consider ambiguous cases as positive behavior forecasts with elevated risk with the effect of false negative. The values of ROC-AUC (0.958 internal, 0.939 external) show a powerful discriminative capacity at all the levels of decision threshold and surpasses the 0.90 threshold often regarded as great in relation to models used in medical diagnostics.

Visualization of the attention maps persistently indicate anatomically possible segments of joint space, cartilage margins, and subchondral bone indicating that model choices are consistent with the available radiographic indicators of osteoarthritis progression. This distance homogeneity gives the notion of interpretability of transformer architectures empirical backing.

Ablation study suggests that the single-model setup recorded a little more internal accuracy (89.58%), but the ensemble + TTA setup displayed more consistent external performance (87.66%) which justifies the design choice of focusing on generalizability. The architecture was sufficiently fast (6.56 ms inference latency, 21.67M parameters) to be used in clinical toplogy in real time without compromising diagnostic accuracy. All these findings indicate that strong, interpretable KOA detection with very little loss in performance can be realized using parameter-efficient sets of transformers on unseen data.

V. CONCLUSION

This paper shows that a small-sized Vision Transformer ensemble with test-time augmentation can be effectively used to detect binary knee osteoarthritis (87.5% internal and 87.09% external accuracy; D=0.41%) with no external validation gap as pointed out by Nasef et al. [12]. In contrast to the previous CNN based methods, which cannot handle global spatial relationships [1,4], or the multimodal fusion techniques, which demand large amounts of computational power [6,11], our parameter-efficient ViT-Small (21.67M parameters) architecture balances both performance and clinical deployability. The sensitivity of the model (97.56) aligns with screening priorities proposed by Guida et al. [6], and the attention map diagrams can be interpreted as anatomically consistent, uncovering an adequate explanation of the "black box" issue of transformer-based medical imaging [9,10]. The most recent researchers are claiming improved accuracy with more complex multimodal transformers [5,18], but none was externally tested seriously, and our 0.41% decline in performance is a stronger indication of the real-world generalizability. Weaknesses are identified as dependence on MRI-only imaging (does not involve X-ray/clinical data convergence area investigated by Guida et al. [6]), binary classification, and a rather small dataset size when compared to those done using OAI [7]. Future efforts will consider multimodal inputs in conformance with the fusion plans of Dai et al. [5] and Al-Zoghby et al. [3], develop longitudinal progression models as introduced by Hu et al. [7], also tested on multi-centered federated data, to enhance claims of generalizability. This contribution sets a realistic precedent to create AI systems as technically correct and yet truly valuable applications to clinical resource-constrained healthcare environments that further the general principle of reliable, deployment-capable medical AI [3,12].

REFERENCES

1. Abdullah, S. S., & Rajasekaran, M. P. (2022). Automatic detection and classification of knee osteoarthritis using deep learning approach. La Radiologia Medica 2022 127:4, 127(4), 398–406. https://doi.org/10.1007/s11547-022-01476-7
2. ABIMOULOUD, M. L., BENSID, K., Elleuch, M., Ammar, M. ben, & KHERALLAH, M. (2024). Vision transformer based convolutional neural network for breast cancer histopathological images classification. Multimedia Tools and Applications 2024 83:39, 83(39), 86833–86868. https://doi.org/10.1007/s11042-024-19667-x
3. Al-Zoghby, A. M., Ebada, A. I., Saleh, A. S., Abdelhay, M., & Awad, W. A. (2025). A Comprehensive Review of Multimodal Deep Learning for Enhanced Medical Diagnostics. https://doi.org/10.32604/cmc.2025.065571
4. Apon, T. S., Fahim-Ul-Islam, M., Rafin, N. I., Akter, J., & Alam, M. G. R. (2024). Transforming Precision: A Comparative Analysis of Vision Transformers, CNNs, and Traditional ML for Knee Osteoarthritis Severity

Diagnosis. Proceedings - 6th International Conference on Electrical Engineering and Information and Communication Technology, ICEEICT 2024, 31–36. https://doi.org/10.1109/ICEEICT62016.2024.10534528
5. Dai, Y., Gao, Y., Liu, F., Dai, Y. ;, Gao, Y. ;, & Liu, F. (2021). TransMed: Transformers Advance Multi-Modal Medical Image Classification. Diagnostics 2021, Vol. 11, Page 1384, 11(8), 1384. https://doi.org/10.3390/diagnostics11081384
6. Guida, C., Zhang, M., & Shan, J. (2023). Improving knee osteoarthritis classification using multimodal intermediate fusion of X-ray, MRI, and clinical information. Neural Computing and Applications 2023 35:13, 35(13), 9763–9772. https://doi.org/10.1007/s00521-023-08214-8
7. Hu, K., Wu, W., Li, W., Simic, M., Zomaya, A., & Wang, Z. (2022). Adversarial Evolving Neural Network for Longitudinal Knee Osteoarthritis Prediction. IEEE Transactions on Medical Imaging, 41(11), 3207–3217. https://doi.org/10.1109/TMI.2022.3181060
8. Jiang, X., Zhu, Y., Liu, Y., Cai, G., & Fang, H. (2024). TransDD: A transformer-based dual-path decoder for improving the performance of thoracic diseases classification using chest X-ray. Biomedical Signal Processing and Control, 91, 105937. https://doi.org/10.1016/j.bspc.2023.105937
9. Maqsood, S., Maqsood, N., Shahid, S., Subhan, F. E., Sarwar, M. A., Yousufi, M., Qurthobi, A., Zafar, A., Khan, M. A., Damaševičius, R., & Maskeliūnas, R. (2025). Knee osteoarthritis network: A hybrid transformer-based approach for enhanced detection and grading of knee osteoarthritis. Engineering Applications of Artificial Intelligence, 159, 111751. https://doi.org/10.1016/j.engappai.2025.111751
10. Meenakshi, K., Himaja, A. L., Akshara, K., & Raj, K. D. P. (2025). Clinical Diagnosis Using Multi-Modal Transformer. Proceedings of the 6th International Conference on Inventive Research in Computing Applications, ICIRCA 2025, 1171–1176. https://doi.org/10.1109/ICIRCA65293.2025.11089779
11. Mohd Sagheer, S. v, H, M. K., Ameer, P. M., Parayangat, M., Abbas, M., & Author, C. (2025). Transformers for Multi-Modal Image Analysis in Healthcare. https://doi.org/10.32604/cmc.2025.063726
12. Nasef, D., Nasef, D., Sawiris, V., Girgis, P., & Toma, M. (2024). Deep Learning for Automated Kellgren–Lawrence Grading in Knee Osteoarthritis Severity Assessment. Surgeries 2025, Vol. 6, Page 3, 6(1), 3. https://doi.org/10.3390/surgeries6010003
13. Panwar, P., Chaurasia, S., Gangrade, J., & Bilandi, A. (2025). Early diagnosis of knee osteoarthritis severity using vision transformer. BMC Musculoskeletal Disorders 2025 26:1, 26(1), 884-. https://doi.org/10.1186/s12891-025-09137-2
14. Pei, Y., Mu, L., Xu, C., Li, Q., Sen, G., Sun, B., Li, X., & Li, X. (2023). Learning-based landmark detection in pelvis x-rays with attention mechanism: data from the osteoarthritis initiative. Biomedical Physics & Engineering Express, 9(2), 025001. https://doi.org/10.1088/2057-1976/ac8ffa
15. Raminedi, S., Shridevi, S., & Won, D. (2024). Multi-modal transformer architecture for medical image analysis and automated report generation. Scientific Reports 2024 14:1, 14(1), 19281-. https://doi.org/10.1038/s41598-024-69981-5
16. Song, L., Li, C., Tan, L., Wang, M., Chen, X., Ye, Q., Li, S., Zhang, R., Zeng, Q., Xie, Z., Yang, W., & Zhao, Y. (2024). A deep learning model to enhance the classification of primary bone tumors based on incomplete multimodal images in X-ray, CT, and MRI. Cancer Imaging 2024 24:1, 24(1), 135-. https://doi.org/10.1186/s40644-024-00784-7
17. Song, Y., Wang, J., Ge, Y., Li, L., Guo, J., Dong, Q., & Liao, Z. (2024). Medical image classification: Knowledge transfer via residual U-Net and vision transformer-based teacher-student model with knowledge distillation. Journal of Visual Communication and Image Representation, 102, 104212. https://doi.org/10.1016/j.jvcir.2024.104212
18. Wang, W., He, J., Liu, H., & Yuan, W. (2024). MDC-RHT: Multi-Modal Medical Image Fusion via Multi-Dimensional Dynamic Convolution and Residual Hybrid Transformer. Sensors 2024, Vol. 24, Page 4056, 24(13), 4056. https://doi.org/10.3390/s24134056
19. Yao, X., Wu, Y., Xie, G., Zhao, D. Y., & Xia, D. H. (2025). A comprehensive survey on multimodal computer vision (CV)-assisted corrosion assessment and corrosion prediction. Corrosion Engineering Science and Technology. https://doi.org/10.1177/1478422X251387524
20. Zhang, W., Zhang, C., Gao, Y., & Jin, Z. (2025). KineticsSense: A Multimodal Wearable Sensor Framework for Modeling Lower-Limb Motion Kinetics. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 9(3). https://doi.org/10.1145/3749462