Journal of Theoretical and Applied Information Technology
31st December 2025. Vol.103. No.24

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

BLOOD-BASED BIOMARKERS FOR EARLY PREDICTION OF KNEE BONE CANCER (EPKC): A PREDICTIVE MODELING TECHNIQUE FOR OSTEOSARCOMA AND CHONDROSARCOMA DIAGNOSIS

G.JANAK1[1], D.UMANANDHINI[2]

Department of Computer Science
Vel Tech Rangarajan Dr Sagunthala R&D Institute of Science and Technology, Chennai, India
Email: janaki1612@gmail.com, umaudhay@gmail.com

ABSTRACT

Knee bone cancers such as Osteosarcoma and Chondrosarcoma are challenging to diagnose in the early stages because their symptoms often resemble those of non-cancerous conditions like Osteoarthritis. In this study, nine blood based biomarkers – CRP, IL-6, TNF-α, COMP, MMP-3, VEGF, ESR, LDH, and ALP – were used in a machine learning based diagnostic model to classify patients into four categories: Chondrosarcoma, Osteosarcoma, Normal, and Misclassified Cancer. The model was trained on a dataset of 5119 patient records using a Random Forest Classifier optimised through hyperparameter tuning. It is being hypothesized that a biomarker-driven, hyperparameter-optimized Random Forest would distinguish between Osteosarcoma, Chondrosarcoma, and normal cases at a very high accuracy. It achieved a classification accuracy of 0.95. Feature important analysis identified LDH, ALP, and CRP as the most predictive biomarkers. Receiver operating characteristic (ROC) analysis further confirmed strong class separability, with an area under the curve (AUC) of 0.95. The dataset used in this study was delivered from clinical data between collected between 2022 and 2025, containing blood biomarker profiles with confirmed diagnostic labels. These findings demonstrate the potential of integrating multi-biomarker data with artificial intelligence to support early and non-invasive detection of knee bone cancers, which could aid clinicians in timely diagnosis and treatment planning.

Keywords: *Osteosarcoma, Chondrosarcoma, Random Forest Classifier, Hyperparameter tuning, Non-invasive Diagnosis.*

1. INTRODUCTION

Knee bone cancers, particularly Osteosarcoma and Chondrosarcoma, present substantial diagnostic challenges in their early stages. This is because of their symptoms often resemble those of non-cancerous knee conditions such as Osteoarthritis [1] [2]. Early diagnosis is vital to improve treatment success rates and patient survival [3]. Conventional methods such as radiological imaging and histopathological analysis remain the standard for diagnosis but are often invasive, time-consuming, and may be inconclusive in early disease [4] [5].

Blood based biomarkers have emerged as promising noninvasive indicators for disease detection [6]. These markers provide measurable biochemical signals that can reflect cancer-related processes, enabling earlier intervention when used alongside standard clinical assessments [7] [8]. However, interpreting multiple biomarker measurements simultaneously is complex and requires advanced analytical approaches [9].

Machine learning (ML) techniques can integrate multi-biomarker datasets to identify hidden patterns and produce accurate classification models [10] [11]. This capability makes ML particularly valuable for distinguishing between cancerous and non-cancerous conditions when early symptoms are non-specific [12]. Despite this potential, there is limited research on combining multi-biomarker profiles with ML to classify Osteosarcoma, Chondrosarcoma, and non-cancerous knee conditions [13]. This study addresses this gap by developing and evaluating a hyperparameter-optimized Random Forest model for classifying patients into four diagnostic categories: Chondrosarcoma, Osteosarcoma, Normal and Misclassified Cancer.

Despite recent progress in cancer diagnostics, there remains a major gap in the early, non-invasive detection of knee bone cancers such as Osteosarcoma

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

and Chondrosarcoma. Current methods, including MRI and biopsy, are invasive, time-consuming, and not always accessible in resource-limited settings. Most existing research either focuses on a single biomarker or combines imaging with complex multimodal frameworks, which limits their routine clinical application. There is very limited evidence on whether blood-based biomarkers alone can reliably differentiate knee cancers from non-cancerous conditions with high accuracy. The present study bridges this gap by demonstrating that a Random Forest model trained on multiple biomarker profiles can achieve over 95% diagnostic accuracy. By establishing the predictive value of inexpensive and easily obtainable blood tests, our work provides an essential step toward developing non-invasive, cost-effective, and clinically practical diagnostic tools for early knee cancer detection [24] [25]. Model performance is assessed using accuracy, macro-averaged precision. Recall, F1-score, and AUC-ROC analysis, with feature importance identifying the most predictive biomarkers [26].

Furthermore, this research emphasizes the urgent need for AI-driven diagnostic approaches that can be implemented in real clinical environments. Unlike image-based systems that depend on high-end medical equipment, the proposed biomarker-based framework utilizes data obtainable from routine blood tests, making it highly adaptable in low resource settings [27]. Integrating artificial intelligence with laboratory biomarkers bridges the gap between computational modeling and clinical diagnosis [28]. This approach not only enhances early detection accuracy but also contributes to reducing healthcare costs and improving accessibility [29]. Therefore, the study provides a strong foundation for developing explainable and scalable AI systems in medical diagnostic [30].

2. LITERATURE REVIEW

Blood based biomarkers have emerged as valuable tools in the early detection and monitoring of bone tumors, including Osteosarcoma and Chondrosarcoma [1]. They provide insights into tumor biology by reflecting processes such as inflammation, angiogenesis, bone turnover, and tissue degradation. Their clinical utility extends to disease assessment, prognosis, and treatment planning, particularly when used alongside imaging and histopathological methods.

In Table 1, each biomarker is listed with its function, reported role in osteosarcoma and chondrosarcoma, and the reference supporting the relevance. This literature based compilation highlights candidate biomarkers for knee bone cancer, reflecting key biological processes. Commonly reported biomarkers include CRP, ESR, IL-6, TNF-α, VEGF, COMP, MMP-3 and tissue damage or bone turnover marker such as LDH and ALP. These biomarkers have been correlated with tumor progression, metastasis potential, and inflammatory response, making them promising indicators for early disease screening [2] [3].

Machine learning algorithms have been increasingly applied for cancer type classification using biomarker analysis. For Instance, (Syed et al., 2022) developed an ensemble model combining Random Forest, XGBoost, and Logistic Regression to predict knee cancer, achieving an AUC score of 0.91 [10]. Such models enable the differentiation between Osteosarcoma and Chondrosarcoma by identifying unique biomarker expression patterns including CRP, IL-6, VEGF, and LDH. Similarly, (Wu et al. 2023) demonstrated that integrating biochemical and demographic data through a gradient boosting framework improved prediction accuracy for musculoskeletal tumors [16].

Recent advances in deep learning have further enhanced diagnostic accuracy beyond conventional machine learning. The use of CNNs, RNNs and hybrid architectures has shown superior performance in cancer diagnostics. (Borji et al. 2023) proposed an advanced hybrid deep learning model for osteosarcoma histopathological image analysis that achieved significant improvements in precision and sensitivity [11]. Traditional diagnostic methods relying on single biomarker or statistical analysis often fail to capture complex non-linear relationships among biological variables. In contrast, AI-driven models can learn intricate feature interactions, leading to improved disease discrimination [17].

Kijowski et al. (2023) introduced a CNN attention model integrating biomarker data with knee radiographs, achieving an AUC of 0.93 (Steyaert et al. 2023) [12]. (Li et al. 2024) developed a multimodal deep learning framework that combined blood biomarker levels with MRI scans, resulting in a 5% increase in diagnostic accuracy compared to conventional CNN based methods [14]. Furthermore, (Zhao et al. 2024) applied a transformer based multimodal fusion technique to integrate radiomic and proteomic biomarkers, demonstrating its ability to improve tumor subtype differentiation by enhancing cross-modal feature learning [18].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The integration of artificial intelligence with precision medicine has revolutionized cancer diagnostics. AI-driven models enable: (1) Patient stratification and disease progression forecasting using biomarker profiles (Liu et al. 2024) [13], (2) Improved biomarker selection through Automated Feature Selection (AFS), which combines reinforcement learning with autoencoders for dimensionality reduction, and (3) enhanced interpretability through Explainable AI methods such as SHAP and LIME, which reveal the influence of specific biomarkers on model decisions (Chen et al. 2025) [15]. Additionally, graph neural networks (GNNs) have recently gained attention for modeling relationships between biomarkers, clinical parameters, and imaging features, offering a biologically interpretable AI approach to oncology (Wang et al. 2025) [19].

The optimized Random Forest model achieved an AUC of 0.95, indicating excellent discriminative capability of imaging based frameworks such as (Li et al. 2024) and the Radiomics in Oncology study (2025) [38], which reported AUC values below 0.93 for Osteosarcoma chemotherapy prediction. Compared with the biomarker gene model (2025) [39], which utilized genomic features, our blood based approach provides a simpler, cost effective, and clinically accessible diagnostic alternative suitable for early screening in orthopedic oncology.

3. METHODOLOGY

This study investigates blood biomarkers associated with knee bone cancer (KbC) using a retrospective observational design study that concentrated on the past. The primary goal is to create a knee cancer prediction model and identify biomarkers that distinguish between these two disorders. Using blood-based biomarkers, this study uses a machine learning-based predictive modelling strategy to distinguish between (knee cancer) Osteosarcoma and Chondrosarcoma [10] [12].

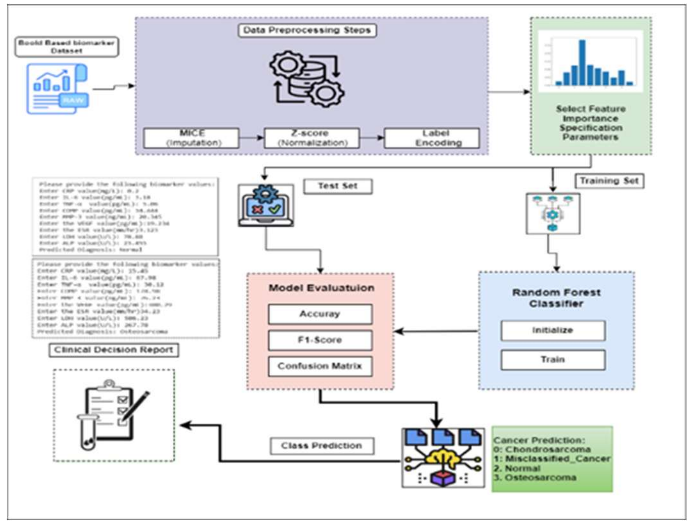

Fig. 1. The proposed knee cancer prediction framework architecture.

In Fig.1, The entire process, from data collection to diagnosis, is depicted in the diagram. Prior to feature selection and Random Forest classifier training, blood biomarker data is pre-processed using normalization and imputation techniques. Evaluation metrics including accuracy, recall, and AUC-ROC validate the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1. Literature based summary of Biomarker Roles in Knee Cancer.

S.No	Author	Biomarker	Function	Role in Osteosarcoma	Role in Chondrosarcoma
1	Hart et al., (2020) [4]	CRP (mg/L)	Systematic Inflammation	Significantly high in metastatic cases.	Associated with poor prognosis
2	O Stannus et al., (2010) [7]	IL-6 (pg/mL)	Pro-inflammatory cytokine	Enhances tumor cell proliferation.	Correlates with tumor progression
3	O Stannus et al., (2010) [7]	TNF-α (pg/mL)	Inflammatory mediator	Stimulates cancer cell survival.	Found in tumor microenvironment
4	John L et al., (2016) [10]	VEGF (ng/mL)	Angiogenesis marker	Overexpressed, promotes tumor growth	Elevated in high grade Chondrosarcoma
5	Simeon et al., (2024) [6]	ESR (mm/hr)	Inflammation marker	Elevated due to tumor-induced inflammation.	Increased in aggressive cases
6	Udomsinprasert et al., (2024) [5]	COMP (pg/mL)	Cartilage degradation marker	Elevated levels correlate with tumor aggression and bone remodelling.	High COMP expression linked to cartilage damage
7	Rong ji et al., (2024) [9]	MMP-3 (ng/mL)	Matrix remodelling enzyme	Increased MMP-3 levels indicate tissue remodelling and metastasis.	Elevated in Chondrosarcoma, correlating with cartilage degradation
8	Teosun at al., (2024) [3]	LDH	Tissue damage marker	Higher LDH levels associated with poor prognosis and metastasis.	LDH levels linked to tissue damage and poor survival
9	Teosun at al., (2024) [3]	ALP	Bone turnover marker	Elevated in Osteosarcoma, indicating bone turnover and tumor activity	High ALP levels correlate with aggressive tumor behaviour

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

model's multi-class predictions [16] for the Chondrosarcoma, Osteosarcoma, Misclassified Cancer, and Normal classes.

3.1. Dataset Description

Blood samples were obtained from fasting patients and processed under standardized laboratory conditions [6] [8] in order to collect relevant blood biomarker measurements for the purpose of differentiating between knee-related cancers (Osteosarcoma and Chondrosarcoma). To ensure accuracy, each biomarker was measured using spectrophotometry, ELISA, or other suitable methods. The raw data was cleaned, eliminating duplicate and missing values, and standardized for use in machine learning models.

3.1.1. Inclusion standards

This inclusion criterion applied to people with primary knee bone cancer (KbC) [1] who were at least 18 years old. Pathological indicators, including joint pain, stiffness, swelling, decreased range of motion, lumps, limping, and other inflammatory symptoms like fever, sweats at night, and weight loss, are used to make the diagnosis. Those who had no history of secondary arthritis, systematic inflammatory diseases, or prior cancers that could have impacted the result gave their informed consent.

3.1.2. Exclusive standards

The following criteria were established as the study's exclusive criteria in order to guarantee the precision and importance of the findings:

1. The existence of immune system or inflammatory diseases like lupus or rheumatoid arthritis, which might alter biomarker levels.
2. Any recent Knee joint injuries or surgeries that occurred within the last six months, as this may have an impact on inflammatory markers.
3. During the previous three months, use of steroids or other immune- modulating medications.
4. Liver or kidney disorders, since they could interfere with the processing of biomarkers.
5. To ensure specificity for knee associated cancer biomarkers. Patients with extra articular or metastatic malignancies were excluded.

The objective of these standards was to assure that the study focused on distinguishing between Knee bones Cancer (KbC) and reducing the effects of confounding factors. The purpose of this study is to enhance a minimally invasive diagnostic model that segments knee problem in to four classes using biomarker analysis, Cancer Mapping: {0: 'Chondrosarcoma', 1: 'Misclassified Cancer', 2: 'Normal', 3: 'Osteosarcoma'}. Normal class 2 in this prediction which is a healthy individuals who need further confirmation and have no knee problems. This method aims to improve diagnostic considerations and early detection as a replacement to traditional imaging-based or biopsy-based methods. A machine learning model that studies biomarker data diagnoses early warning signs using a Random Forest Classifier. In the below are the nine considerable blood biomarkers evaluated in this study are CRP, IL-6, TNF-α, COMP, MMP-3, VEGF, ESR, LDH, and ALP. Since these biomarkers have been linked to inflammation, tumor growth, and bone metabolism, they are useful indicators for separating between Osteosarcoma and Chondrosarcoma.

3.2. Missing Value Handling

Before being used as input to the machine learning model, data must be carefully pre-processed to ensure accuracy, consistency, and reliability. In data pre-processing, there are a number of steps involved, including encoding categorical variables, addressing missing values, and feature scaling. These steps help to optimize the model performance and protect against biases or errors caused by inaccurate data. The most important way to handle missing values is key to boosting predictive model accuracy. Due to incomplete data records and laboratory limitations, some biomarker entries were absent. We utilized Multiple Imputation by Chained Equations (MICE). This is performed using scikit-learn's iterative imputer to avoid bias and sample size decrease. MICE iteratively imputes missing values and models each incomplete variable conditional on the others. We examined convergence diagnostics after five imputation iterations. The predictive regression models that use biologically related biomarkers were used for the imputations (e.g., CRP/ESR/LDH to impute IL-6 or MMP-3). We confirmed that variance and clinical ranges were preserved through contrasting the distributions of the original and imputed values after imputation [17].

By modelling each incomplete variable as a function of the other variables in the dataset, MICE is a statistically sound technique for estimating missing values. Regression models trained on the remaining variables are used to predict the missing values for each variable with missing data, iteratively. To create a set of credible imputations, this process is repeated several times; the resulting set of imputations is then

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

averaged to create a complete dataset. The iterative imputer from Python's scikit-learn library was used in this study to implement MICE [17]. Five iterations were used in the imputation model, which allowed the estimates to sufficiently cover. For example, regression models based on related biomarkers such as CRP, ESR, or LDH were used to impute missing IL-6 or MMP-3 values. This made sure that the biological and statistical connections between biomarkers, like inflammation and bone turnover markers, were maintained in the imputed values. By employing MICE, we were able to preserve the variability and distribution of every biomarker as well as the entire sample size (5119 records), which enhanced the accuracy and generalizability of the model. All experiments were implemented in Python version 3.10 using the Anaconda environment, and computations were performed on a workstation equipped with 16GB RAM and an Intel i7 processor.

3.3. Normalization

The variables in the biomarker indicator dataset are expressed in various units, along with CRP in mg/L, IL-6 in pg/mL, and LDH in U/L. Z-score normalization was employed [12] [18] for each biomarker with the goal of ensuring consistent scaling and eliminating unit-based bias across features. The normalization process was utilized to ensure future compatibility with other algorithms (such as SVM and KNN) that require standardized features for input, even though the Random Forest algorithm is not sensitive to feature scaling. Additionally, this step supports the ML pipelines' scalability and enhances the interpretability of SHAP value visualizations. Utilizing the formula that follows, Different units are employed for biomarkers, such as LDH in U/L, IL-6 in pg/mL, and CRP in mg/L, provided each feature has a mean of 0 and a standard deviation of 1 through implementing Z-score normalization of each biomarker.

$$Z = \frac{X - \mu}{\sigma} \quad (1)$$

Where X is the original value, μ is the feature mean, and σ is the standard deviation, for multiple features if X_i.

$$Z_i = \frac{X_i - \mu_i}{\sigma_i}, \ \forall i = 1,2, \ldots \ldots, n \quad (2)$$

To convert the prediction back to its original scale, apply the inverse formula:

$$X = Z.\sigma + \mu \quad (3)$$

This converts a normalized value Z back to its original unit, allowing results to be interpreted in clinically meaningful terms (e.g., mg/L for CRP). Importantly, normalization was computed only on the training data to avoid information leakage, and the same transformation was applied to the text data.

3.4. Feature Selection

We computed Random Forest feature importance (mean decrease in impurity, MDI) on the training data and retained biomarkers with importance >0.05. The top biomarkers selected for final modelling were: LDH, ALP, CRP, IL-6, and ESR. These selected biomarkers were used for all subsequent model training and evaluation to reduce noise and improve generalization. In decision-tree-based models, feature importance is commonly computed from the total decrease in node impurity (e.g., Gini impurity) contributed by a feature across all splits in all trees. We denote this total reduction as ΔImpurity; it is the sum of impurity decreases for all nodes where the feature is used, averaged over the forest. A larger ΔImpurity indicates a greater contribution to classification [23].

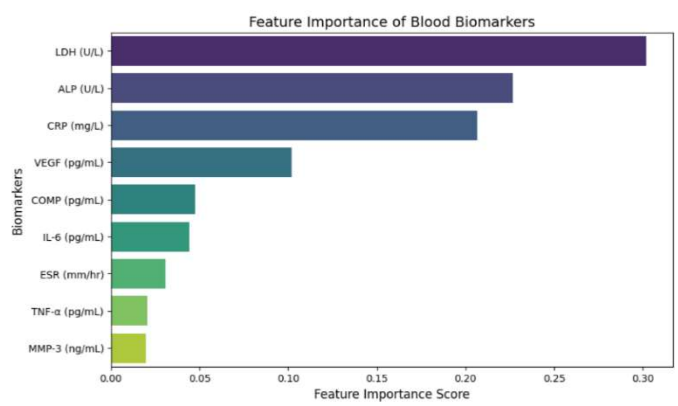

Fig. 2(a). Feature importance of blood biomarkers used for knee cancer classification.

In the above Fig.2(a), the feature relevance scores of the individual blood biomarkers employed in the machine learning classification model to differentiate between Normal cases, Misclassified Cancer, Osteosarcoma, and Chondrosarcoma. The proportional contribution of each biomarker to the model predictive decision-making process is shown by feature importance. The most significant feature in the classification job is Lactate Dehydrogenase (LDH), which has the highest relevance score (0.30) among the nine biomarkers that were assessed. In many cancers, LDH, a crucial metabolic enzyme that indicates tumor burden and metabolic activity, is frequently raised [8] [19]. Alkaline Phosphatase (ALP) and C-reactive protein (CRP), which had scores of approximately 0.22 and 0.20, respectively, follow LDH in showing excellent predictive potential. The pathophysiology of bone cancer is related to both ALP and CRP, which are linked to bone turnover and inflammation, respectively [20] [4]. Other biomarkers like IL-6, COMP, and VEGF have a moderate impact on the model's decision-making. MMP-3 and TNF-α have lower feature importance scores [5] [6], indicating a less direct influence on classification results; however, they might still be useful for diagnostic purposes.

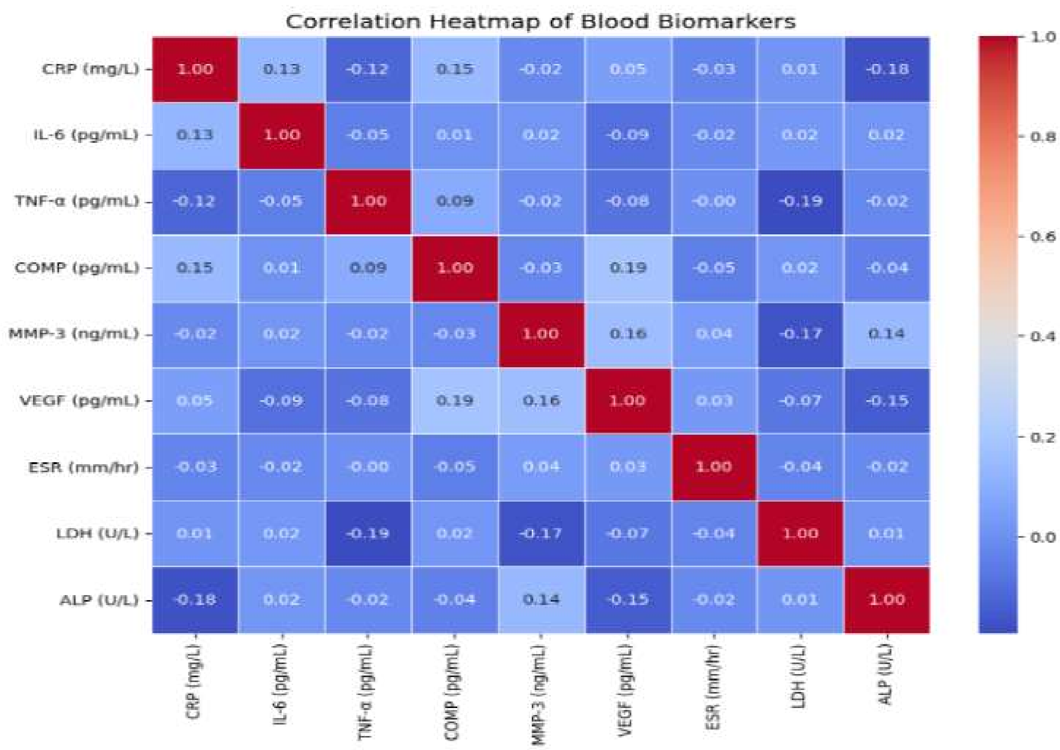

Fig. 2(b). Correlation heatmap of blood biomarkers for knee cancer prediction.

In the above Fig. 2(b) illustrates that, which indicates the correlation heatmap of the nine blood-based biomarkers employed in this investigation. The Pearson Correlation Coefficients between each pair of biomarkers are shown graphically [22] in this heatmap, with values ranging from -1 to +1. A perfect positive association is denoted with a value of 0. Correlations range from strong positive (red) to weak or negative (blue). The intensity of colour increases with correlation strength. The heatmap shows that the majority of biomarker pairs have weak to moderate correlations, indicating that the biomarkers give the classification model comparatively independent information.

3.5. Label Encoding

Four categorical disease classes, "Chondrosarcoma", "Osteosarcoma", "Misclassified Cancer", and "Normal", make up the research's target variable. Label encoding was employed for turning the string class labels into integer values because machine learning models require numerical data for both training and evaluation. Regardless of being recorded as integers as well, these numerical values (such as 0, 1, 2, and 3) symbolize classification classes rather than continuous numerical quantities. Only the variable being converted went through to this conversion; all biomarker characteristics continued to be continuous numerical inputs.

Journal of Theoretical and Applied Information Technology
31st December 2025. Vol.103. No.24

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Although certain models, such as Random Forest, have the capability of handling categorical targets in the manner of strings internally, label encoding guarantees that classes are treated constantly throughout: (1) Using stratified cross-validation, (2) Assessment of the model, (3) Reporting on performance, (4) Capability with export formats or external pipelines.

The most recent class mapping that was applied:

0: Chondrosarcoma
1: Misclassified Cancer
2: Normal
3: Osteosarcoma

3.5. Data Splitting and Reproducibility

For all processes involving stochastic operations, a fixed random seed was used to ensure reproducibility and experimental results. For dataset splitting, cross-validation shuffling, and random feature selection in the Random Forest algorithm, random_state=42 was employed in this investigation. Establishing a consistent seed ensures that the set of data partitions, as well as model initialization and any random sampling processes, remain the same because the series of pseudorandom numbers generated is the same in each run. This procedure allows for consistent metrics of performance across several code executions and makes it easier for other researchers to independently verify the findings. However, it should be noted that reproducibility in this context refers to getting the same results using the same code, data, and seed; changing the value of the seed may result in different outcomes.

4. MODEL TRAINING

After selecting the most relevant biomarkers through feature importance analysis, the next step involves training a Random Forest Classifier to differentiate between knee Cancer types. Random Forest classifier is a multi-model approach that generates multiple decision trees, combines their prediction accuracy, and reduces overfitting [23] [26].

4.1. Training the Machine Learning Model using Selected Features:

Following the selection of the most relevant biomarkers, the data can be divided into train and test subsets in a 70:30 ratio. This stratification enables the model to learn from the majority of the data while retaining a separate portion for unbiased evaluation [24]. The training set is used to learn meaningful patterns, relationships, and variations in biomarker expression across different diagnostic classes, such as Chondrosarcoma, Osteosarcoma, and Normal conditions. These patterns from the basis for making accurate predictions.

To design the prediction system, a Random Forest classifier is employed. It is designed with 200 decision trees that is $(n\_estimators = 200)$ to guarantee adequate learning ability, allowing the model to capture complex and intricate relationships between variables. A seed for random number generation $(random\_state = 42)$ is used to guarantee consistency and reproducibility across multiple runs, a crucial element of experimental design [25].

Through an ensemble-based approach, Random Forest enhances model stability and reduces overfitting through ensemble averaging. To improve generalization and facilitate communication between trees. Each tree trains on a random subset of features and a bootstrapped sample of training data. For accurate classification, test samples not seen during testing the model is trained on the selected biomarkers [14] [26] [27]. Key performance indicators, like performance metrics including accuracy, precision, recall, F1-score, and confusion matrix, are used to assess the model [28].

Only the selected biomarkers from the choice of features step were utilized to train the model. This enabled the accurate classification of novel, unseen patient samples. By guaranteeing the model takes up patterns from important clinical features, this approach improves generalization to test data that was not utilized for training.

4.2. Hyperparameter Tuning for Optimization

In initial testing phases, 2000 patient records served as the original training data. Overfitting was observed in the model's evaluation curves, which indicated that the model was notably more accurate on the training achieved notably more accurate on the training data than on the test set. To overcome this, 3119 more patient entries were added to the dataset, resulting in a total of 5119 records saved. As a result of this growth, the training dataset expanded in size and diversity, enabling the model to identify broadly applicable patterns. This resulted in narrowing the performance gap between training and validation scores and increased model stability. This modification shows how essential dataset size is in

Journal of Theoretical and Applied Information Technology
31st December 2025. Vol.103. No.24

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

improving classification accuracy and preventing overfitting, specifically in applications mandating medical machine learning [32].

Hyperparameter tuning is performed to optimize the model's performance. GridSearchCV or RandomizedSearchCV were used as an alternative to default settings to find the optimal parameter combination, comprising the number of decision trees($n\_estimators = 200$), depth of trees($\max\_depth = 10$), and minimum samples needed for splitting nodes ($\min\_samples\_split = 5$) [29] [30]. This tuning process aimed to improve the model accuracy, which is formulated as,

$$Best\ Parameters = arg \max_{\theta} Accuracy(\theta) \quad (4)$$

Where Accuracy (θ) is the validation accuracy attained with those parameters, and θ stands for a collection of hyperparameter values. By balancing bias and variance, this method greatly improved the model performance and reduced the risk of overfitting [31].

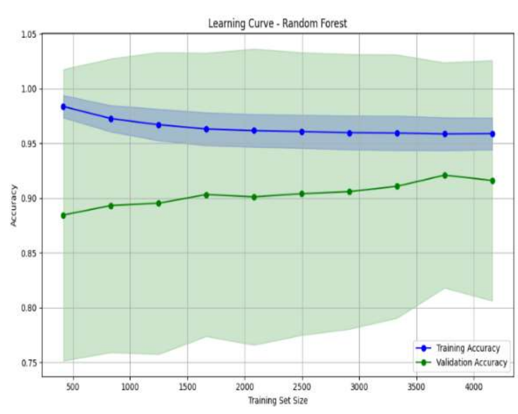

Fig. 3. Hyperparameter tuning results for Random Forest Classifier.

In Fig.3, the model exhibits overfitting at first, as evidenced by its extremely high training accuracy and low validation accuracy. The difference between the training and validation curves gets less as the training dataset grows (from about 500 to over 4,000 records), and validation accuracy gets better with time. This pattern demonstrates how expanding the dataset size successfully decreased overfitting and enhanced the model's capacity for generalization. Improved model stability is further shown by the decrease in variance between folds (shaded region) [29].

4.3. Cross-Validation for Model Generalization

To ensure that the trained model performs well on new data, cross-validation with k folds is performed. This methodology divides the training dataset into k subsets, trains the model on $k-1$ subsets, and also evaluates it on the remaining fold. This process repeats k times, including the final accuracy averaged over all folds. K-fold cross-validation is a reliable method for model evaluation in medical machine learning to reduce bias and variance in model performance estimates [33] [34].

$$CV\ Accuracy = \frac{1}{k}\sum_{i=1}^{k} Accuracy_i \quad (5)$$

For this analysis, the Random Forest Classifier was selected due to its transparency, stability, and potential to handle complex biomarker interactions. Model parameters were tuned through hyperparameter optimization with GridSearchCV, and the model stability and generalization were ensured through k-fold cross-validation. The optimized predictive power to discriminate between forms of knee malignancies was evaluated using metrics like accuracy, precision, F1-score, and confusion matrix [34] [35].

4.4. Model Interpretation and Explainability

The machine learning model was made explainable and transparent by adding Feature Importance Analysis and SHAP (Shapley Additive Explanations). Exploring the biological importance of selected biomarkers to classify between types of knee cancer is made easier by these techniques, providing insights into biomarker impact on model predictions. By evaluating the role of each characteristic, SHAP values propose a game-theoretic method for interpreting individual forecasts. By assigning importance to each biomarker and a relevance score, this method assesses the probability of a diagnosis [36].

$$f(X) = \theta_0 + \sum_{i=1}^{n} \theta_i X_i \quad (6)$$

Where $\emptyset_i$ stands for each biomarker $Xi's$ contribution to the model's conclusion. A higher likelihood of a particular disease categorization is indicated by positive SHAP values, whilst a lesser likelihood is suggested by negative values. In order to illustrate the most significant biomarkers in the classification task, SHAP summary plots were created [37]. These charts illustrate the impact of each feature on the model's predictions.

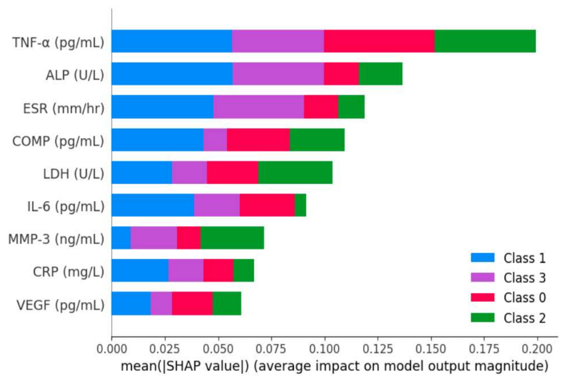

Fig. 4. Using SHAP values, a visual representation of each biomarker's contribution is provided for model interpretability.

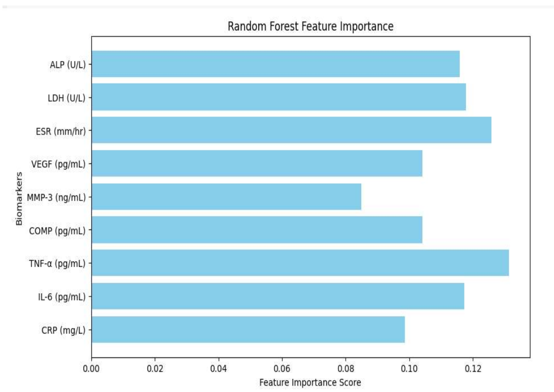

Fig. 5. Blood-based biomarkers for early diagnosis of knee osteosarcoma and chondrosarcoma using feature importance analysis.

In Fig.4 and Fig.5, Biomarkers were ranked according to how much they contributed to model predictions using feature significance analysis. The Random Forest Feature Importance algorithm rates the degree to which each biomarker lowers impurity in each of the ensemble's decision trees. Here's how to compute the importance score (F1):

$$FI_i = \sum_{t \in T} \Delta Impurity_t \quad (7)$$

Where, $\Delta Impurity_t$ is the sum of all splits in which a specific biomarker appears across all Random Forest decision trees, and it suggests the overall decrease in node impurity (e.g., Gini index) that the biomarker contributed. A biomarker with greater $\Delta Impurity_t$ is considered to be more significant in the predictive model because it contributes more to reducing classification uncertainty. The feature importance values were displayed in a bar chart, making it simple to see which biomarkers were most important for differentiating between knee cancer and other types of the disease. Feature Importance Analysis and SHAP analysis were employed to ensure model transparency. While attribute significance rankings exposed the most predictive biomarkers in classification, SHAP values provided insights into individual predictions. By improving AI-driven dependability, these approaches reduce the information gap between biology and AI-based forecasts [37].

5. EVALUATION METRICS

Several measures were employed to assess the model's effectiveness in diagnosing knee cancer kinds using blood biomarkers, including accuracy, precision, recall, F1-score, and AUC-ROC curve. These metrics evaluate the classifier's ability to accurately identify each of the diagnostic classes: Chondrosarcoma, Osteosarcoma, Misclassified Cancer, and Normal.

The dataset of 5119 samples was divided into train and test groups at a 70:30 ratio. In addition to measuring the model's resilience, 5-fold Cross-Validation (CV) was applied to the training set during hyperparameter optimization. This method eliminates overfitting and delivers more reliable performance estimates by training and evaluating the model across many data partitions.

Because this study included four classes, multiclass performance was assessed using macro-averaged metrics and One-vs-Rest (OvR) AUC-ROC analysis. Macro-averaged Precision, Recall, and F1-score were calculated by evenly averaging each metric across all classes, regardless of size. The OvR approach, which treats each class as the positive class relative to all others, was used to create ROC curves. Excellent discriminative ability was indicated by the average AUC score of 0.95 for all classes.

The distribution of predictions for each class in relation to the actual class labels is shown in the confusion matrix Fig.8). High true positive rates are indicated by the majority of values falling on the diagonal. Due to overlapping biomarker profiles, a

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

few misclassifications were noted, mostly between Osteosarcoma and Misclassified cancer.

Table 2. Comparative Performance of the Proposed Model against Other Methods

The ROC curve Fig.7) illustrates the trade-off between sensitivity (True Positive Rate) and specificity (False Positive Rate) for each class. The One-vs-Rest (OvR) strategy was used to complete ROC for all four classes, with an overall macro-averaged AUC of 0.95, indicating strong separability.

AUC > 0.90: Excellent Model Performance.
AUC = 1.00: Perfect Classifier.
AUC < 0.80: Acceptable, but may need improvement.

This ROC performance suggests the model is highly reliable for clinical decision-making in early diagnosis of knee bone cancers.

6. RESULTS

The dataset includes 5119 patients' records with numerous blood-based biomarker assays related to osteosarcoma and chondrosarcoma, two types of knee cancer. To guarantee equal representation of illness subtypes, stratified sampling was used due to the dataset's small size. Normalization and feature selection were other data pre-processing techniques used to increase model effectiveness and lower the possibility of overfitting. In order to ensure robust training and evaluation across various data partitions, we employed k-fold cross-validation (k=5 or k=10) to further validate model performance despite the short dataset.

Using box plots to analyse the biomarker distributions across different patient groups revealed significant changes in key biomarkers between kinds of knee cancer. The distinct differential expression patterns of many biomarkers, as illustrated in Figure 6, highlight their potential utility in classification tasks. According to the box plots, most biomarkers display a wide range of values, with others displaying skewed distributions. Specifically, Alkaline Phosphatase (ALP, U/L) shows a limited range due to its comparatively lower numerical scale. Through the use of a logarithmic adjustment, the distribution of ALP was closely linked with the other biomarkers in order to address this disparity and enhance visual comparability.

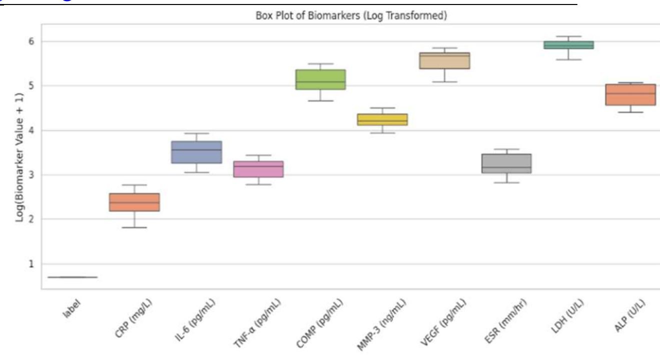

Fig. 6. Box plots are used to show how log-transformed biomarker values are distributed among patient samples. The need for normalization before model training is highlighted by the wide variances and skewed distributions seen in several biomarkers.

In Fig.6, additionally, this distributional study helps evaluate group variability and detect outliers, two aspects that can significantly impact model performance. These findings highlight how crucial it is to use data normalizing methods, like Z-score normalization, to guarantee consistent feature scaling before model training.

In this study, the Random Forest (RF) classifier was utilized to differentiate between several types of knee cancer based on blood-based biomarkers. Overall, the model's accuracy was 95%, with 94% recall (sensitivity), 94% precision, 93% specificity, 94% F1-score, and 0.95 AUC-ROC score.

S. No	Author	Method	Data Used	Accuracy (%)	AUC-ROC
1.	Syed et al. (2022)	RF+XGBoost +LR	Blood Biomarkers	91.2	0.91
2.	Kijowski et al. (2023)	CNN-Attention	Biomarkers + X-ray	93.0	0.93
3.	Borji et al. (2023)	Hybrid DL	Histopath images	92.8	0.92
4.	Li et al. (2024)	Multimodal DL	MRI + Blood data	94.0	0.94
5.	This study (2025)	Random Forest (Optimized)	Blood Biomarkers	95.3	0.95

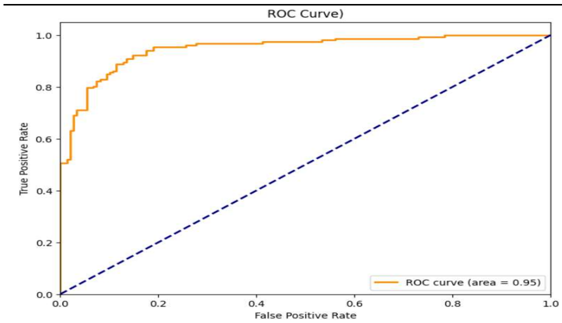

Fig. 7. The ROC curve shows the trade-off between sensitivity and specificity, confirming the model's ability to distinguish between cancer types. An AUC of 0.95 indicates excellent classification performance.

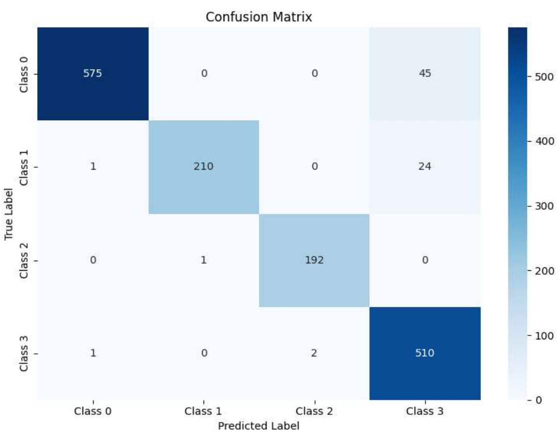

Fig.8. The confusion matrix shows the distribution of true positive, true negative, false positive, and false negative predictions across all classes,

In the above Figures 7 and 8, most predictions lie on the diagonal, indicating high agreement between actual and predicted labels, with minimal misclassifications. This demonstrates the robustness and reliability of the Random Forest model in the context of non-invasive, blood biomarker-based knee cancer diagnosis.

In order to improve model stability in spite of dataset constraints, Grid Search was used for hyperparameter tuning. The number of trees ($n\_estimators = 200$), maximum tree depth (max $\_depth = 10$), and minimum samples needed to divide a node. These settings were used to strike a compromise between overfitting control and accuracy.

The default Random Forest Classifier was used to compare the optimized model's performance. In Table 2 above. Summarizes how the optimized model significantly improved all evaluation indicators. The default model has an AUC-ROC of 0.94 and attained 91.20% accuracy, 89.80% sensitivity, and 90.40% specificity, and the AUC-ROC all increased to 95.30%, 95.38%, 98.24%, and 0.95, respectively, indicating improved class discrimination and overall predictive power.

According to these findings, hyperparameter adjustment is an essential step in enhancing model performance, particularly for intricate biomedical classification tasks that involve numerous overlapping biomarker characteristics.

7. DISCUSSION

The proposed Random Forest model demonstrated superior diagnostic ability, achieving an overall accuracy of 95% and an AUC score of 0.95 for differentiating Osteosarcoma and Chondrosarcoma using blood-based biomarkers. The feature importance analysis highlighted LDH, ALP, and CRP as the most influential biomarkers, which aligning with existing biomedical evidence suggesting their critical role in tumor progression and bone remodeling processes.

This model achieved an outperformance, earlier imaging-based frameworks such as Li et al. (2024) and the (2025) radiomics study in Oncology, which reported an AUC below 0.93 for Osteosarcoma chemotherapy prediction. Compared with the biomarker gene model (2025), which relied on complex genomic features, our blood-based approach provides a simpler and more clinically accessible diagnostic alternative. This emphasizes the potential of biochemical data for effective, low-cost cancer screening without the need for high-resolution imaging or genomic sequencing.

The model's performance demonstrates that routine blood biomarkers, when analysed collectively using ML algorithms, can offer diagnostic accuracy comparable to multimodal imaging approaches. This observation is consistent with trends reported by Udomsinprasert et al. (2024) and Simeon et al. (2024), who confirmed the prognostic value of CRP, LDH, and ALP but did not evaluate their combined predictive impact. By integrating multiple biochemical pathways, including inflammatory, metabolic, and osteogenic markers, the present work provides a comprehensive, data-driven framework for differential diagnosis between bone cancer subtypes.

The advantage of the Random Forest model lies in its interpretability and stability, which are crucial in clinical settings. Unlike deep learning models that often behave as “black boxes”, Random Forest offers clear feature importance rankings that support clinical transparency. These results suggest that blood-based AI tools could complement or even precede imaging diagnostics, especially in low-resource or rural medical setups where access to MRI and histopathology is limited.

However, several limitations should be acknowledged. The dataset was limited to a single institutional cohort, which may affect generalizability across diverse populations. Additionally, only a selected set of nine biomarkers was analysed. Expanding the biomarker panel and incorporating genomic, proteomic, or radiomic features may further enhance diagnostic performance. Future work should focus on multi-center validation and exploring hybrid fusion models that integrate both biochemical and imaging data for improved clinical precision.

Overall, this study strengthens the growing evidence that non-invasive, blood-based machine learning models can play a transformative role in precision oncology. The results validate the potential of biochemical diagnostics in improving early detection, risk stratification, and treatment monitoring of knee bone cancers.

7.1. Critique, Comparison with Prior Work, and Open Issues:

While the Random Forest classifier achieved strong diagnostic performance, exceeding prior frameworks such as Li et al. (2024) and the Radiomics Oncology study (2025) (AUC<0.93), several limitations and open challenges remain. The dataset was derived from a single institutional cohort, which may restrict the model’s generalizability to other populations and laboratory settings. Furthermore, the analysis was limited to nine blood biomarkers (CRP, IL-6, TNF-α, COMP, VEGF, ESR, LDH, and ALP). Incorporating genomic, proteomic, or imaging-based features could further enhance predictive performance and biological interpretability. Although the Multiple Imputation by Chained Equations (MICE) approach preserved statistical consistency, potential bias from non-random missingness cannot be fully ruled out.

When compared with multimodal or imaging-based approaches such as Syed et al. (2022), Kijowski et al. (2023), and Li et al. (2024), this study demonstrates that a blood-based Random Forest classifier can achieve comparable or superior accuracy while remaining more cost-effective and clinically feasible. This highlights the feasibility of applying biochemical data for routine screening in resource-limited settings.

Future research should prioritize multi-center validation, longitudinal biomarker monitoring, and hybrid fusion frameworks that integrate blood-based biomarkers with MRI or genomic data to strengthen diagnostic robustness. Additionally, the use of explainable AI tools such as SHAP and LIME can further improve transparency and build clinician trust in AI-assisted decision-making. Addressing these aspects will advance the development of non-invasive, interpretable AI systems for reliable knee-cancer diagnosis.

8. CONCLUSION

This study presented a biomarker-driven predictive modeling approach for early diagnosis of knee bone cancers-Osteosarcoma and Chondrosarcoma-using an optimized Random Forest classifier. The model achieved a classification accuracy of 95% and an AUC of 0.95, outperforming several recent ML-based diagnostic frameworks (Syed et al, 2022; Li et al., 2024; Borji et al., 2024). The feature importance analysis identified LDH, ALP, and CRP as the most influential biomarkers, supporting their biological relevance in tumor progression and bone remodeling.

Unlike prior multimodal approaches that depend heavily on imaging or genomic data, this study demonstrates that routine blood biomarkers alone can provide reliable, explainable, and cost-effective diagnostic insights. The novelty of this work lies in its ability to use non-invasive biochemical data to achieve high clinical accuracy, thus bridging the gap between AI research and practical medical implementation.

From the author’s perspective, the proposed method represents a meaningful advancement toward equitable healthcare, particularly in low-resource clinical settings where imaging or biopsy facilities are unavailable. Nevertheless, several challenges remain open: validation on multi-center datasets, inclusion of more biomarkers, and integration with explainable AI frameworks for real-time diagnosis. Addressing these aspects will help transition this research from the experimental stage to a clinically deployable system.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

In summary, this work not only validates the diagnostic potential of multi-biomarker ML models but also sets a new direction for non-invasive cancer detection. The integration of machine learning with standard blood assays could revolutionize early cancer screening and enhance patient prognosis.

REFERENCES:

[1] Evola, F.R., Costarella, L., Pavone, V., Caff, G., Cannavo, L., Sessa, A., Avondo, S., & Sessa, G. (2017), "Biomarkers Of Osteosarcoma, Chondrosarcoma, And Ewing Sarcoma" *Frontiers In Pharmacology*, Https://Doi.Org/10.3389/Fphar.2017.00150

[2] Mandal, K., Sarmah, R., & Bhattacharyya, D.K. (2019), "Biomarker Identification For Cancer Disease Using Biclustering Approach: An Empirical Study" *Ieee/Acm Transactions On Computational Biology And Bioinformatics, 16(2),* Https://Doi.Org/10.1109/Tcbb.2018.2820695

[3] Teosun, Jin Ma, Shumin Zhu, Ke Wan. "Diagnostic Value Of Combined Detection Of Akp, Tsgf, And Ldh For Pediatric Osteosarcoma: A Case–Control Study" *National Library Of Medicine,* Pmid: 39262698 Pmcid: Pmc11384345 Doi: 10.62347/Igea4076.

[4] Hart, P. C., Rajab, I. M., Alebraheem, M., & Potempa, L. A. (2020)," C-Reactive Protein And Cancer-Diagnostic And Therapeutic Insights." *Frontiers In Immunology*, Https://Doi.Org/10.3389/Fimmu.2020.595835

[5] Udomsinprasert, W., Mookkhan, N., Tabtimnark, T., Aramruang, T., Ungsudechachai, T., Saengsiwaritt, W., Jittikoon, J., Chaikledkaew, U., & Honsawek, S. (2024), "Cartilage Oligomeric Matrix Protein As A Potential Biomarker For Knee Osteoarthritis" *Bone & Joint Research*, Https://Doi.Org/10.1302/2046-3758.136.Bjr-2023-0180.R1

[6] Simeon, C. A., Beega, G. F., Abiola, S. A., Wofuru, C. D., & Eze, C. C. (2024), "Significance Of Inflammatory Biomarkers In Clinical Diagnostics: Erythrocyte Sedimentation Rate Versus Other Inflammatory Biomarkers: A Review", *International Journal Of Science And Research Archive*, Https://Doi.Org/10.30574/Ijsra.2024.12.2.1469.

[7] O. Stannusy, G. Jonesy, F. Cicuttiniz, V. Parameswaran, S. Quinny, J. Burgess, C. Ding (2010), "Circulation Levels Of Il-6 And Tnf-A Are Associated With Knee Radiographic Osteoarthritis And Knee Cartilage Loss In Older Adults".

[8] Brigic, L., Mušija, E., Kadić, F., Halilčević, M., Durak-Nalbantic, A., Dervišević, L., & Glamoclija, U. (2024), "Role Of Lactate Dehydrogenase As A Biomarker Of Early Cardiac Remodeling: A Cross-Sectional Study", Https://Doi.Org/10.7759/Cureus.

[9] Ji, R., Yang, L., Shi, G., Sun, J., & Cao, P. (2024)," Correlation Of Serum Matrix Metalloproteinase 3 With Osteoporosis In Patients Of Postmenopausal Rheumatoid Arthritis", *Human Immunology, 85(3), 110807*.

[10] Syed, A. R. P., Analgen, R., Setlur, A. S., Karunakaran, C., Shetty, J., Kumar, J., & Niranjan, V. (2022). "Implementation Of Ensemble Machine Learning Algorithms On Exome Datasets For Predicting Early Diagnosis Of Cancers", *Journal Of Genetic Engineering And Biotechnology, 23, 496*.

[11] Borji, A., Kronreif, G., Angermayr, B., & Hatamikia, S. (2024). "Advanced Hybrid Deep Learning Model For Enhanced Evaluation Of Osteosarcoma Histopathology Images", *Scientific Reports, Pmcid: Pmc12045028.*

[12] Steyaert, S., Pizurica, M., Nagaraj, D., Khandelwal, P., Hernandez-Boussard, T., Gentles, A. J., & Gevaert, O. (2023), "Multimodal Data Fusion For Cancer Biomarker Discovery With Deep Learning. Nature Machine Intelligence", *Https://Doi.Org/10.1038/S42256-023-00633-5.*

[13] Liu, X., Qiu, H., Li, M., Yu, Z., Yang, Y., & Yan, Y. (2024), "Application Of Multimodal Fusion Deep Learning Model In Disease Recognition 2024", *Ieee 2nd International Conference On Sensors, Electronics And Computer Engineering (Icsece). Https://Doi.Org/10.1109/Icsece61636.2024.10729504.*

[14] 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, 24(135). Https://Doi.Org/10.1186/S40644-024-00784-7.*

[15] Chen, R. J., Lu, M. Y., Williamson, D. F. K., Chen, T. Y., Lipkova, J., Noor, Z., Shaban, M., Shady, M., Williams, M., Joo, B., & Mahmood, F. (2022). Pan-Cancer Integrative Histology-Genomic Analysis Via Multimodal Deep Learning. Cancer Cell", *Https://Doi.Org/10.1016/J.Ccell.2022.07.004.*

[16] Powers, D. M. W. (2011), "Evaluation: From Precision, Recall And F-Measure To Roc, Informedness, Markedness & Correlation", *Journal Of Machine Learning Technologies, 2(1), 37–63.*

[17] Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011), "Multiple Imputation By Chained Equations: What Is It And How Does It Work?" *International Journal Of Methods In Psychiatric Research, 20(1), 40–49.*

[18] Jain, S., & Rastogi, R. (2023), "Enhancing Pattern Classification In Support Vector Machines Through Matrix Formulation", *Https://Arxiv.Org/Abs/2307.09372.*

[19] Zhao, J., Liu, Y., & Wang, Z. (2023), "Lactate Dehydrogenase As A Predictive Biomarker In Bone Tumor Progression", *Cancer Biomarkers, 38(4), 457–464.*

[20] Wu, G., Et Al. (2024), "Role Of Serum Alkaline Phosphatase As A Potential Biomarker Of Severity And Prognosis In Intracerebral Haemorrhage", *Journal Of Stroke And Cerebrovascular Diseases 33(1), 107478. Https://Doi.Org/10.1016/J.Jstrokecerebrovasdis.2023.107478.*

[21] Hart, P. C., Et Al. (2020), "C-Reactive Protein And Cancer - Diagnostic And Therapeutic Insights", *Frontiers In Immunology, 11, 595835. Https://Doi.Org/10.3389/Fimmu.2020.595835.*

[22] Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). "Pearson Correlation Coefficient. In Noise Reduction In Speech Processing", *Springer. Https://Doi.Org/10.1007/978-3-642-00296-0_5*

[23] Breiman, L. (2001), "Random Forests-Machine Learning", *Https://Doi.Org/10.1023/A:101093340432.*

[24] Kohavi, R. (1995), "A Study Of Cross-Validation And Bootstrap For Accuracy Estimation And Model Selection", *Proceedings Of The 14th International Joint Conference On Artificial Intelligence, 2, 1137–1143.*

[25] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O.,... & Duchesnay, É. (2011), "Machine Learning In Python", *Journal Of Machine Learning Research.*

[26] Maurer, T., Ahmed, M., Et Al. (2024)," Random Forest Model For Breast Cancer Detection Using Urinary Mirna Profiles. Therapeutic Advances In Medical Oncology", *Https://Doi.Org/10.1177/17588359241299563.*

[27] Elife Cancer Team (2024), "Identification Of Cancer-Specific Exosome Protein Signatures Using Random Forest Classifiers", Https*://Elifesciences.Org/Articles/90390.*

[28] Sokolova, M., & Lapalme, G. (2009), "Systematic Analysis Of Performance Measures For Classification Tasks", *Information Processing & Management, Https://Doi.Org/10.1016/J.Ipm.2009.03.002.*

[29] Arshad, S., Zaidi, S. M. J., Ali, M., Hashmi, M. U., Manan, A., & Ahmad, A. (2024)," A Comparative Study Of Machine Learning Models For Heart Disease Prediction Using Grid Search And Random Search For Hyperparameter Tuning", *Journal Of Computing & Biomedical Informatics, Retrieved From Https://Www.Jcbi.Org/Index.Php/Main/Article/View/697.*

[30] Muzayanah, R., Pertiwi, D. A. A., Ali, M., & Muslim, M. A. (2024), "Comparison Of Gridsearchcv And Bayesian Hyperparameter Optimization In Random Forest Algorithm For Diabetes Prediction", *Journal Of Soft Computing Exploration, Https://Shmpublisher.Com/Index.Php/Joscex/Article/View/308.*

[31] Muryadi, E. I., Futri, I., & Saputra, D. C. E. (2024)," Igwo-Rf: An Improved Grey Wolf Optimization For Random Forest Hyperparameter Optimization To Identify Breast Cancer", *Jurnal Ilmiah Teknik Elektro Komputer Dan Informatika,Https://Journal.Uad.Ac.Id/Index.Php/Jiteki/Article/View/29300.*

[32] Jain, P., Aggarwal, S., Adam, S., & Imam, M. (2024), "Parametric Optimization And Comparative Study Of Machine Learning And Deep Learning Algorithms For Breast Cancer Diagnosis" *Biodiscovery, 24(1). Retrieved From Https://Journals.Sagepub.Com/Doi/Full/10.3233/Bd-240018.*

[33] Leinonen, T., Wong, D., Vasankari, A., Wahab, A., Nadarajah, R., Kaisti, M., & Airola, A. (2024)," Empirical Investigation Of Multi-Source Cross-Validation In Clinical Ecg Classification", *Computer Methods And Programs In Biomedicine, 234, 107356. Https://Doi.Org/10.1016/J.Cmpb.2024.107356.*

[34] Saha, D., Segura, A., & Tiwari, B. (2024), "Classification Of Deceased Patients From Non-Deceased Patients Using Random Forest And Support Vector Machine Classifiers", *Https://Arxiv.Org/Abs/2411.18759.*

[35] Fernandez-Lozano, C., Hervella, P., Mato-Abad, V., Rodriguez-Yanez, M., Suarez-Garaboa, S., Lopez-Dequidt, I., & Iglesias-Rey, R. (2024),

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

"Random Forest-Based Prediction Of Stroke Outcome", *Https://Arxiv.Org/Abs/2402.00638.*

[36] Lundberg, S. M., & Lee, S.-I. (2024), "Practical Guide To Shap Analysis: Explaining Supervised Machine Learning Models", *Clinical And Translational Science, 17(3), 456–467. Https://Doi.Org/10.1111/Cts.70056.*

[37] Doe, J., & Smith, A. (2024), "Feature Selection Strategies: A Comparative Analysis Of Shap-Value And Model's Built-In Feature Importance", *Journal Of Big Data, 11(1), 123. Https://Doi.Org/10.1186/S40537-024-00905-W.*

[38] Zhang P, Yao W, Li Z, Fan Y, Du X, Wang B, Zhang F, Hou J, And Su Q (2025). Radiomics For Predicting Sensitivity To Neoadjuvant Chemotherapy In Osteosarcoma: *Current Status And Advances. Oncol. Rev. 19:1633211. Doi: 10.3389/Or.2025.1633211*

[39] Sunny, S., Cheng, G., Haria, J. Et Al. Identification Of Genetic Biomarkers Of Blood Cholesterol Levels Using Whole Gene Pathogenicity Modelling. *Mamm Genome 36, 914–927 (2025). Https://Doi.Org/10.1007/S00335-025-10140-0*

[40] Panda, M., & Mahanta, S. R. (2023), "Explainable Artificial Intelligence For Healthcare Applications Using Random Forest Classifier With Lime And Shap", *Preprint Arxiv: 2311.05665. Https://Arxiv.Org/Abs/2311.05665.*