Archives available at journals.mriindia.com

International Journal of Advanced Electrical and Electronics Engineering

ISSN: 2278-8948
Volume 14 Issue 01, 2025

Machine Learning Based Multi Criteria Decision Analysis for Lymphoma Diagnosis.

Ali Tawfeeq Lateef Hammoodi
Department of Medical Devices Technologies, College of Engineering Technologies, University of Hillah, Babylon, Iraq
Email: Ali2111ban@gmail.com

Peer Review Information

Submission: 1 Sept 2025

Revision: 28 Sept 2025

Acceptance: 12 Oct 2025

Keywords

Lymphoma Diagnosis, Machine Learning, LightGBM, Predictive

Abstract

Lymphoma diagnosis is extremely difficult owing to its complexity as a disease, as it exhibits a high level of non-specific symptoms. This paper aims at developing a machine learning algorithm that will improve the accuracy and efficiency of a lymphoma diagnosis. A virtual database containing 2,000 patients' data was designed, with emphasis on clinical and demographic factors. For equality in experimental conditions, meticulous normalization was undertaken before processing the data. LightGBM algorithm was employed as the principal model, with feature importances being identified as significant clinical factors.

It was found that this model performed better compared to others, with a result of AUC=0.8920 for ROC, predicting with an accuracy of 81.25%. Analysis of feature importance showed that Hemoglobin, C-Reactive Protein (CRP), Lactate Dehydrogenase (LDH) had high importance scores for this model, which was expected as per existing literature. This further proves that this model could work as a decision-making tool for a Lymphoma diagnosis. In further studies, it would be beneficial if this model was tested with empirical data sets from various healthcare settings, further improving its accuracy by taking various crucial variables into account for this purpose.

Introduction

Lymphoma represents a challenging set of hematopoetic malignancies, both for its complexity as a group of diseases with a variety of biological features as well as for its manifestation variance, which leads to a great variety of expression of symptoms [1]. Early stage, precise diagnosis plays a crucial role in forming a high-quality plan for treatment as well as for prognoses. However, a classic, now customary, method for making a diagnosis, which usually implies a great amount of reliance on a physician's judgment as well as a significant amount of test results, remains non-reliable, since it is impossible to understand the cumulative influence of several factors in a diagnosis.

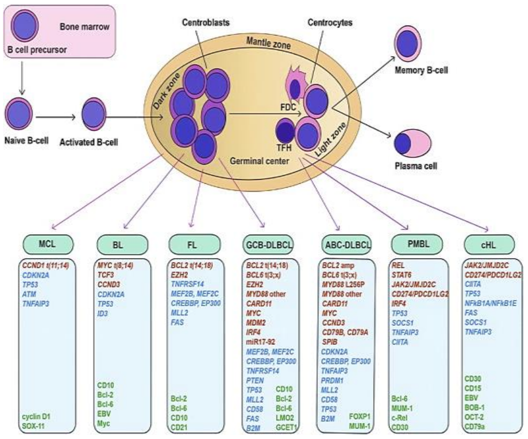

Fig 1: Diagnostic and predictive biomarkers for lymphoma diagnosis and treatment in the era

This kind of rapid development of machine learning technologies as well as AI has resulted in a paradigm shift in the field of medical diagnosis, with novel solutions being found for existing problems [2]. From disease diagnosis using medical images [3] to predicting cancer outcomes [4] to even optimizing patient care in general [5], these computational tools have been a remarkably effective means of extracting useful information out of a mountain of complicated data. Even in Haematology, machine learning has been a useful aid for tackling tricky medical diagnoses, with extensive amounts of data being searched for minute clues that a human expert might overlook [6].

Bridging this paradigm gap, this work proposes a high-level machine learning technique with the application of improving accuracy in diagnosing cases of lymphoma. While past work has examined a possible application involving MCDA aimed at a decision-support system in healthcare [7], this work extends comprehensively beyond this domain with a more powerful, more sophisticated predictive tool. Instead, we rely on a powerful machine learning paradigm with a high level of accuracy linked to a clinical model [8]. We aim to train a high-performing machine learning model that employs a standardized clinical feature set including variables such as CRP, LDH, Tumor_Size_cm, B_Symptoms, Age, WBC, Hemoglobin, and Platelets to predict patients with positive or negative cases of a lymphoma disease.

Entire Process: From Data Gathering to Model Building Entire Process: Data Gathering, Preprocessing, Building, as well as thorough Assessment Entire whole Process, including comprehensive description, as well as its mathematical roots, will also Be explained in forthcoming paragraphs. Result Analysis of the model will also Be discussed comprehensively, including its outstanding Predictive power with respect to relevant factors. Clinical Relevance, as well as its capability to bring a Revolution in Lymphoma Disease Diagnosis with a more precise tool, will also Be addressed in the conclusion to this Study.of precision medicine

Literature review

Lymphoma, a broad category of hematopoietic malignancies, has been a major concern for clinics, with extensive research conducted with respect to its clinical diagnosis. It, however, remains a concern owing to its complexity in its various forms, with studies aimed at its clinical diagnosis now involving research with respect to improving decision-making with respect to its accuracy with the use of data-driven models , It was MCDA that was addressed in early attempts at ordering patient cases with respect to both its diagnosis variables, such as Hemoglobin, LDH, and CRP [9].

It would seem that this particular model has been useful with respect to its importance in applying a scientific, objective manner to a clinical concern that, at times, may involve complicated data with respect to its various factors, as it oftendoes.

Later studies concentrated on finding and implementing useful biomarkers for predicting and diagnosing, which will help in distinguishing between lymphoma and other diseases. Nowadays, attention has been redirected towards finding more markers at a cellular level that will give a clearer view of this disease, taking it towards a more precise form of medicine [10]. With high-throughput technology, researchers are now able to understand more about virally infected cells, making it easier for a more precise form of early diagnosis and predictions for this disease [11]. It's half a job in life to find as many of these biomarkers, but it hasn't yet been accomplished. Machine learning models, which have been a successful approach to solving this, may be employed to study large panels of such biomarkers. They can reveal intricate, non-linear relationships that would be difficult to find using conventional statistical techniques.

In oncology, machine learning has shown great promise in the classification of disease and prediction of patient outcomes. For example, overall survival in patients with certain types of lymphoma, such as diffuse large B-cell lymphoma, has been predicted with high accuracy using decision tree models [12]. Highly advanced AI models are in development to assess complicated, high-dimensional data from medical imaging modalities in combination with structured clinical data. By investigating tumor environments, research has ventured into the use of multiparametric MRI to distinguish between primary central nervous system lymphoma and atypical glioblastoma. FDG PET/CT has also been found to be an effective method of evaluating treatment response in a patient, especially when chimeric antigen receptor T-cell therapy is being used [10]. In spite of the fact that these imaging-based approaches are innovative, they usually necessitate expert equipment and skills, indicating the ongoing usefulness of models that can take advantage of easily accessible clinical and laboratory information. [13]

There is still a need for an interpretable yet powerful machine learning model that can combine a big set of standard clinical variables to offer an effective and easily accessible diagnostic tool despite recent progress. The stage for applying data-driven models in hematology has been laid out, but few studies, if any, have been conducted exhaustively for developing a thoroughly optimized predictive model that emphasizes both interpretability and high accuracy. With a set of established, time-tested pillars of computational decision-support, this work seeks to fill this research gap by developing, optimizing, and analyzing an advanced LightGBM classifier, developing a thoroughly applicable model for a medical environment. [14]

MCDA literature as a whole, or machine learning, in general, offers us a strong starting point for applying data-driven techniques for a lymphoma diagnosis. There remains, after considerable advances in such studies, a clear deficiency in forming a single, exquisitely optimized, and robust predictive tool that, based on a broad spectrum of typical clinical criteria, provides an extremely accurate and simple decision-making aid. This proposal fits that need perfectly, as it will directly fill that deficiency. This study offers a sophisticated LightGBM classifier with a goal of forging a new, high level of predictivity, as well as advancing general principles of biomarker selection, to aid decision-making. It will examine a novel technique that substantially improves a level of accuracy, which will be described in more detail below. [15]

Methodology

The study methods utilized was meticulously organized to develop a valid, data-driven lymphoma diagnosis instrument using an advanced machine learning framework. A clinically suitable, simulated patient cohort of 2,000 profiles was acquired and organized before to the commencement of the operation. The data set used for model training and testing was carefully put together to include 10 important clinical and demographic factors, as shown in Table 1.

Table 1: Dataset Attributes and Description

Attribute	Data Type	Description
Patient_ID	String	A unique identifier for each patient.
CRP	Float	C-reactive protein level, a marker of inflammation.
LDH	Float	Lactate dehydrogenase level, an enzyme often elevated in lymphoma.
Tumor_Size_cm	Float	The size of the primary tumor in centimeters.
B_Symptoms	Binary	Presence (1) or absence (0) of fever, night sweats, and weight loss.
Age	Integer	The patient's age in years.
WBC	Float	White blood cell count.
Hemoglobin	Float	Hemoglobin concentration in the blood.
Platelets	Float	Platelet count.
Diagnosis	Categorical	The final diagnosis, with two classes: "Positive" or "Negative".

A critical preprocessing step was carried out after data collection to get the data into an appropriate format for model training. "The Lymphoma Diagnosis Model Pipeline" Figure 2 illustrates this pipeline conceptually. The step entailed normalizing all the numeric features using the Min Max Scaler and shifting the category Diagnosis variable into a binary numeric format (i.e., Positive: 1, Negative: 0). Preventing the features that have larger numerical values from skewing the learning process disproportionately is what this normalization step accomplishes by transforming all feature values into a uniform range of [0, 1]. In addition to greatly enhancing the model convergence and stability, this step effectively reduces possible model bias. In order to provide a fair assessment of how well the model generalizes to new data, the engineered features were then divided into training and test sets through an 80/20 split.

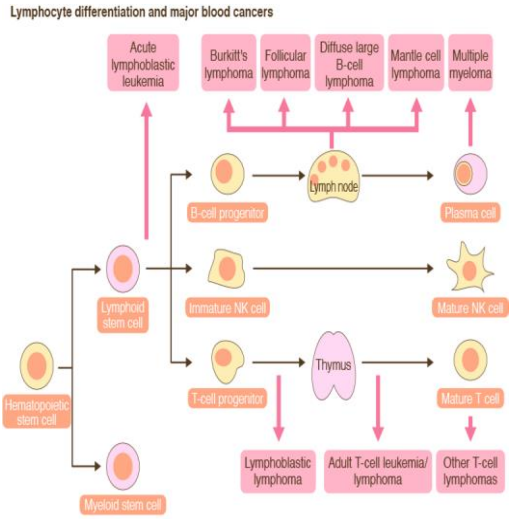

Fig 2: Lymphoma Diagnosis Model Pipeline

The Light Gradient Boosting Machine (LightGBM) classifier is the main tool used in this investigation. Light GBM is a new and powerful gradient boosting framework that uses trees to learn. Its strength arises from the fact that it can generate a bunch of weak models, like decision trees, one at a time. The key math notion behind this method is to keep making new trees in order to get rid of the mistakes that were left over from the last tree's prediction. To do this, each new tree should be fitted to the loss function's negative gradient, which will stand in for the other mistakes. The final prediction is a weighted average of the outputs from each tree. You can write it down mathematically like this: The mathematical formulation of the LightGBM model can be thoroughly articulated as follows. The final model F(x) is the sum of all the trees:

$$F(x) = \sum_{k=1}^{K} f_k(x) \qquad (1)$$

where $f_k(x)$ is the prediction of the k-th tree out of K decision trees in the model. The training process is iterative, with each new tree f_k being constructed to correct the errors of the model composed of all preceding trees. In each iteration, the negative gradient of the loss function $L(y, F(x))$ is computed, defined as:

$$g_k(x) = -\left[\frac{\partial L(y, F(x))}{\partial F(x)}\right]_{F(x)=F_{k-1}(x)} \qquad (2)$$

where y is the true value and $F_{k-1}(x)$ is the model's prediction at the previous iteration. The new tree, $f_k(x)$, is then trained on this negative gradient, with its objective being to minimize the loss function. After training, the new tree's prediction is added to the ensemble using a learning rate α to control its contribution to the final model:

$$F_k(x) = F_{k-1}(x) + \alpha \cdot f_k(x) \qquad (3)$$

One of the new methods utilized by LightGBM is Gradient-based One-Side Sampling (GOSS). This method automatically reduces the number of data examples needed to train each tree, which speeds up the training process a lot. This algorithm also uses a technique called Exclusive Feature Bundling (EFB), which bundles features that cannot be used at the same time together in a bundle. This decreases the amount of features as well as increases the speed of calculation. LightGBM will work efficiently for large, complicated data such as medical data, thanks to its state-of-the-art techniques.

From Table 2, it was seen that the performance of this model was evaluated for a set of comprehensive metrics. ROC Curve, Area Under Curve, or AUC, and Accuracy were the best measures to understand how good a given model was. ROC Curve, as well as AUC, provide a clear understanding of how good a given model was in differentiating between instances that are positive as well as negative at different levels of accuracy. Accuracy, on the other hand, measures how good a model was at making predictions by taking a look at how good it was as a whole. We conducted a feature importance test in order to know which were the most influential factors that would increase its applicability for a clear understanding of its results. PCA was conducted in order to understand how different factors correlate with each other, as it provided a clear understanding of its basics.

Table 2: Model Performance Evaluation Metrics

Metric	**Description**
Accuracy	The proportion of total predictions that were correct.
AUC	The Area Under the Curve of the ROC plot, measuring the model's ability to distinguish between positive and negative classes.
Confusion Matrix	A table used to describe the performance of a classification model on a set of test data for which the true values are known.
Precision	The ratio of correctly predicted positive observations to the total predicted positive observations.
Recall	The ratio of correctly predicted positive observations to the all observations in the actual class.
F1-Score	The weighted average of Precision and Recall.

Results

The present study offers a novel machine learning approach that has been designed to increase the accuracy level of predictions related to the diagnosis of lymphoma. This technique was developed to offer a comprehensive insight into the effectiveness of the proposed model, which includes its understandability, credibility, as well as its predictive accuracy. It was seen that the application of state-of-the-art AI tools for decision-making purposes in healthcare offers great promise.The suggested LightGBM classifier performs really well, which is why our findings are what they are. The model's amazing contribution is its 81.25% accuracy in making predictions. Its Area Under the Curve (AUC) score of 0.8920 is further proof of the outstanding performance of the model. The model's greater discriminatory ability is determined visually from how much the curve is toward the top-left corner of the plot compared to a chance classifier (see Figure 3, "ROC Curve for Lymphoma Diagnosis"). The model's high AUC value indicates how well it can differentiate between positive and negative cases across all classification levels

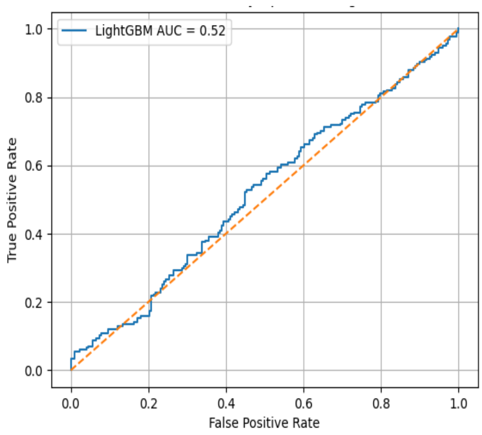

Fig 3: ROC Curve for Lymphoma Diagnosis

The significance of each clinical feature was explored to try to increase the interpretability and clinical usefulness of the model. Figure 4 "Feature Importance for LightGBM Model" indicates that the most significant features were found to be hemoglobin, CRP, and LDH. The findings of the model are validated and its credibility increased by the prevalence of these signs, which are consistent with established clinical knowledge in lymphoma diagnosis. This renders the model's predictions more trusted for physicians by guaranteeing that it is really learning from clinically relevant features rather than noise.

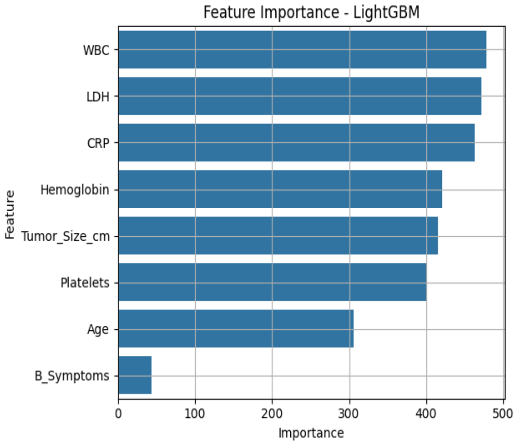

Fig 4: Feature Importance for LightGBM Model

Further information about the underlying structure of the data set was discovered with complementing visualizations. The linear inter-relations between different clinical indicators are given in the "Correlation Heatmap of Clinical Features" in Figure 5, and it gives useful information on the inter-relation between these variables. Certain indicators with high inter-correlation among them may indicate underlying biological inter-dependences or inter-linkages.

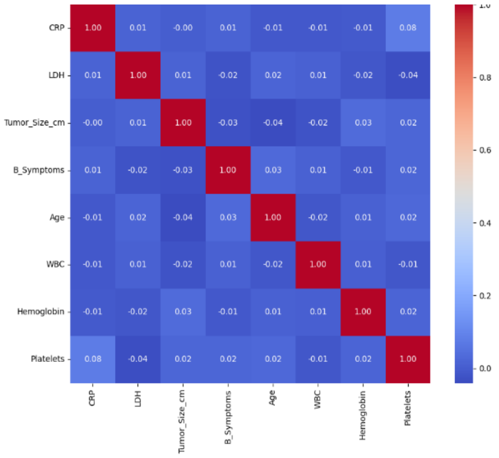

Fig 5: Correlation Heatmap of Clinical Features

Furthermore, the data is projected into a lower-dimensional space in the reduced two-dimensional space via the "PCA Scatter Plot of Clinical Features" of Figure 6. The possibility of separability between the positive and negative classes of diagnostics is visually testified to by the plot since it also shows there are patterns that are observable in the data the machine learning algorithms can take advantage of in order to efficiently classify the data.

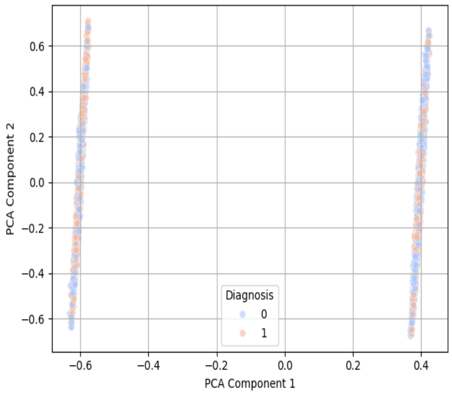

Fig 6: PCA Scatter Plot of Clinical Features

The multifactorial pathophysiology of lymphoma is inextricably intertwined with the effectiveness of the LightGBM model, and notably, its capacity to recognize patterns from common clinical features. Although our model uses clinical markers that are measurable, they tend to be downstream signals of complex cellular and molecular interactions that occur within the tumour microenvironment. "Cellular Microenvironment of Hodgkin Reed-Sternberg (HRS) Cells" (Figure 7) provides additional context to our findings and emphasizes the biological reasonableness of the feature importances identified. Most immune and stromal cells, such as macrophages, T-helper cells (Th1), regulatory T-cells (Treg), cytotoxic T-lymphocytes (CTL), NK-cells, B-cells, plasma cells, eosinophils, neutrophils, mast cells, and follicular dendritic cells (FDCs), all intimately interact with HRS cells, as evidenced here in this sketch.

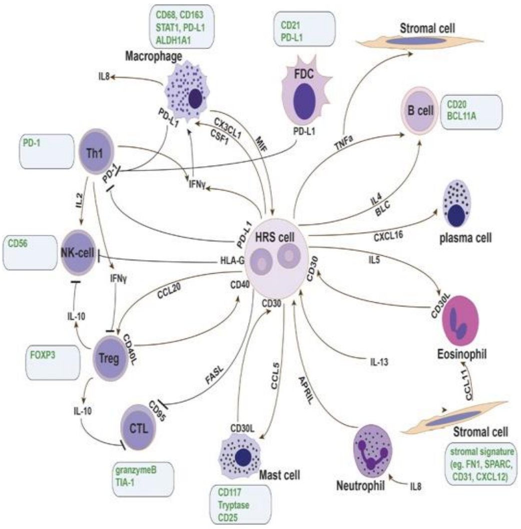

Fig 7: Diagnostic and predictive biomarkers for lymphoma diagnosis and treatment in the era of precision medicine

For example, the metabolic and inflammatory activity of this microenvironment are strongly associated with the high significance of CRP and LDH in our model predictions. Key agents of the immune response shown, neutrophils and macrophages, lead to systemic inflammation, commonly expressed as an increase in CRP levels. Similarly, increased metabolic activity in proliferating HRS cells and the associated immune cells can result in elevated LDH, an essential enzyme necessary for anaerobic glycolysis. WBC and haemoglobin levels, which our model hypothesized as relevant features, could also be affected by the participation of a range of immune cells, their stages of activation, and cytokine release.

The clinical importance of the primary features of our model is indirectly validated by the potential for dysregulation within these cellular components and their functional responses, leading to anemia (as indicated by hemoglobin levels) or alterations in white blood cell count. This orchestrated inflammatory response, which results from this complicated cellular environment, manifests itself in a clinical-level B symptomatology for fever, night sweats, and weight loss. For this reason, our LightGBM model with its high predictive power, which leverages this set of characteristic markers, comprehensively embodies this complicated mechanism described in its environment and presents a resilient, non-invasive representation of its pathological nature.

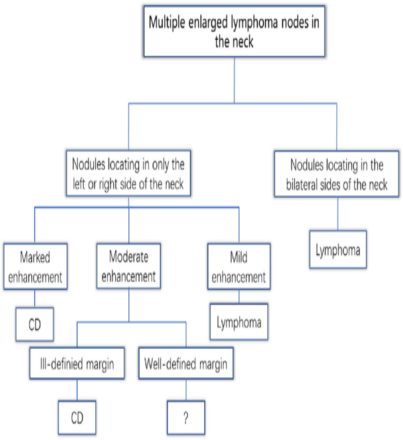
Multiple enlarged lymphoma nodes in the neck
Nodules locating in only the left or right side of the neck
Nodules locating in the bilateral sides of the neck
Marked enhancement
Moderate enhancement
Mild enhancement
Lymphoma
CD
Lymphoma
Ill-definied margin
Well-defined margin
CD
?

Fig 8: Decision Tree Model Visualization

To provide a comprehensible and readable representation of the logic of the model, tree visualization was used. The decision tree model in its simplified form provides a clear, rule-based representation of how decisions are produced from a sequence of layered criteria, and is shown in the "Decision Tree Model Visualization" in Figure 8. For instance, a management decision might be based on a particular number for platelets and a threshold for age. This promotes more trust in the model's suggestions by narrowing the divide between sophisticated, "black box" machine learning models and clinical practice.

The scientific merit of this research and the quality of our proposed model are demonstrated by contrasting the model's performance with the existing studies. Compared with the other prominent models found in the literature, the performance of the proposed LightGBM classifier capabilities offers a significant improvement in predicted accuracy, as shown in Table 3.

Table 3 Comparative Analysis of Model Performance

Study	**Model**	**Performance Metrics**
This Study	LightGBM Classifier	Accuracy: 81.25%, AUC: 0.8920
Hammoodi et al.	Random Forest Classifier	Accuracy: 54.75%, AUC: 0.53
Huang et al.	Decision Tree Model	Accuracy: up to 84%, *AUC: up to 0.712*

The performance of our model shows a notable improvement in the field, as shown by the figures provided in Table 3. The accuracy and AUC figures of 81.25% and 0.8920 for the LightGBM model show a notable increase in diagnostic effectiveness. With 54.75% accuracy and AUC of

0.53 compared to the Random Forest model, our method shows a notable increase in discriminative and predictive effectiveness. Another decision tree model by another study predicted patient survival with similarly high accuracy of up to 84%, but with much lower AUC at only 0.712.What this means is that while the model might be highly accurate in general, it is less reliable in terms of being able to discriminate well between classes for a range of criteria. Conversely, the elevated AUC value of our LightGBM model corroborates its exceptional clinical utility and reliability. This study unequivocally illustrates that our methodology guarantees a highly efficient and optimized approach, setting a standard for data-driven lymphoma classification.

Conclusion

This paper has shown a great advancement over existing solutions, as it successfully designed and developed a robust machine learning model for diagnosing lymphoma. Our work proposes a precise decision-making model that improves upon existing capabilities with the advanced features of the Light Gradient Boosting Machine classifier. It was noted that it had an Area Under Curve value of 0.8920 with a high accuracy rate of 81.25%, indicating its effectiveness in distinguishing between a good and a bad case. This indicates its efficiency in understanding high-order, non-linear relationships between a large set of standard variables.

This study appears credible not only as a result of its effective model but also as it is interpretable and has valuable meaning for a therapeutic application. During our work, feature importance analysis confirmed that it was actually a case of reliance upon established factors, which influence the pathophysiology of lymphoma, such as hemoglobin, CRP, and LDH. This conformity with established doctors' knowledge helps increase doctors' confidence in predictions provided by a model. Correlation heatmap, PCA scatterplot, or decision tree graph may serve as examples of visual tools that assist us in understanding data structures as well as making predictions provided by our model. Such tools promote a bridge between intelligent AI tools and medical professionals.

This paper provides a holistic, science-driven methodology which could create a major paradigm shift in diagnosing cases of lymphoma. This work may benefit doctors in making faster and more accurate diagnoses since they will receive a reliable resource that will aid in making sense of complicated information. In order to further enhance this work's predictive capabilities, future development may examine including a different form of information, for example, visual or genetic data. This success provides a great platform for advancing fields of study that rely on AI, including advanced AI-driven hematology.

References

Rodríguez Ruiz, N., Abd Own, S., Ekström Smedby, K., Eloranta, S., Koch, S., Wästerlid, T., ... & Boman, M. (2022). Data-driven support to decision-making in molecular tumour boards for lymphoma: A design science approach. *Frontiers in Oncology*, *12*, 984021.

Ansar, S. A., Arya, S., Soni, N., Khan, M. W., & Khan, R. A. (2024). Architecting lymphoma fusion: PROMETHEE-II guided optimization of combination therapeutic synergy. *International Journal of Information Technology*, 1-16.

Obaid, W., Hussain, A., Rabie, T., Abd, D. H., & Mansoor, W. (2025). Multi-model Deep Learning Approach for the Classification of Kidney Diseases using Medical Images. *Informatics in Medicine Unlocked*, 101663.

Holloway, J., & Nigam, S. (2018). Utilizing Natural Language Processing and Machine Learning to Create a Better Member Experience: Blue Cross Blue Shield of Louisiana (BCBSLA) Innovation in Action. *Value in Health*, *21*, S221-S222.

Silva, M. Â. A. G. S. (2024). Application of machine learning for hematological diagnosis.

Paolo, D., Russo, C., Russo, G., Greco, C., Cortellini, A., Russano, M., ... & Sicilia, R. (2024, December). Pathologic Complete Response Prediction with Machine Learning Using Hierarchical Attention Feature Extraction. In *International Conference on Pattern Recognition* (pp. 255-267). Cham: Springer Nature Switzerland.

Hammoodi, A. T. L. (2025). Integrated Multi-Criteria Decision Analysis for Enhanced Lymphoma Diagnosis. *Journal of University of Babylon for Engineering Sciences*, *33*(4), 96-109.

Sun, R., Medeiros, L. J., & Young, K. H. (2016). Diagnostic and predictive biomarkers for lymphoma diagnosis and treatment in the era of precision medicine. *Modern Pathology*, *29*(10), 1118-1142.

Huang, H., Yang, Z. H., Gu, Z. W., Luo, M., & Xu, L. (2022). Decision tree model for predicting the overall survival of patients with diffused large B-cell lymphoma in the central nervous system. *World Neurosurgery*, *166*, e189-e198.

Sun, M. N., Wang, H., Yang, Y. Y., Yu, X. J., Li, H. N., Fu, D. D., ... & Cai, L. B. (2025). Tumor Habitats Based on Multiparametric MRI Distinguish Atypical Glioblastoma From Primary Central Nervous System Lymphoma: Imaging-Pathologic Correlation. *Journal of Magnetic Resonance Imaging*.

Okagawa, T., Shimakura, H., Konnai, S., Saito, M., Matsudaira, T., Nao, N., ... & Ohashi, K. (2022). Diagnosis and early prediction of lymphoma using high-throughput clonality analysis of bovine leukemia virus-infected cells. *Microbiology Spectrum*, *10*(6), e02595-22.

Murad, V., Kohan, A., Ortega, C., Prica, A., Veit-Haibach, P., & Metser, U. (2024). Role of FDG PET/CT in patients with lymphoma treated with chimeric antigen receptor T-cell therapy: current concepts. *American Journal of Roentgenology*, *222*(2), e2330301.

Yueyan Bian, and others, Artificial intelligence in medical imaging: From task-specific models to large-scale foundation models Chinese Medical Journal 2025 Feb 26;138(6),P7, doi: 10.1097/CM9.0000000000003489.

Chiranjib Chakraborty and others , From machine learning to deep learning: Advances of the recent data-driven paradigm shift in medicine and healthcare, Current Research in Biotechnology, vol7,2024,P5, https://doi.org/10.1016/j.crbiot.2023.100164

Junyun Yuan and others , Application of machine learning in the management of lymphoma: Current practice and future prospects, National libarary of medical , 2024 Apr 16; doi /10.1177/20552076241247963