Blood Disease Risk Assessment from Visual Medical Reports Using ML and Generative AI

Md. Fahad Nakib*, Md. Atikuzzaman†
*Dept. of Computer Science and Engineering, Green University of Bangladesh, Dhaka, Bangladesh
Email: fahadnakib27@gmail.com †Lecturer, Dept. of Computer Science and Engineering, Green University of Bangladesh, Dhaka, Bangladesh. Email: atikuzzaman524@gmail.com

***Abstract*—In this paper, we present DiagonoAid, a web-based diagnostic assistant designed to assess blood disease risk from visual medical reports using machine learning and generative AI. The system enables users to upload visual medical reports—such as scanned blood test results, which are then processed through a hybrid pipeline that extracts relevant health metrics using a multimodal generative AI model, normalizes the data against clinical reference ranges, and predicts disease probabilities using a trained XGBoost classifier. To enhance interpretability, the system integrates a generative language model that produces concise, human-readable summaries and actionable health tips based on the predicted outcomes. The entire workflow is encapsulated within a user-friendly web interface, ensuring accessibility for non-technical users. This project demonstrates the potential of combining structured ML models with generative intelligence to democratize health insights, reduce diagnostic latency, and support early intervention strategies in resource-constrained settings.**

***Index Terms*—XGBoost, generative AI, Health Risk Prediction, Medical Report Analysis**

I. Introduction

Early and accurate detection of blood-related diseases is critical for effective treatment and improved patient outcomes. Traditional diagnostic methods often rely on manual interpretation of blood test reports, which can be time-consuming, error-prone, and challenging for patients without medical expertise. In resource-constrained environments, access to specialized healthcare professionals and diagnostic tools may be limited, exacerbating delays in diagnosis and intervention. Recent advances in machine learning (ML) and generative artificial intelligence (AI) offer promising avenues to augment clinical decision-making by automating data extraction, analysis, and interpretation from complex medical documents [1].

The challenge of interpreting visual medical documents, such as scanned blood test reports remains largely underexplored [1]. These documents often contain heterogeneous formats, handwritten annotations, and non-standardized layouts, making conventional optical character recognition (OCR) techniques insufficient for reliable data extraction [2]. To address this gap, multimodal generative AI models have emerged as powerful tools capable of understanding and extracting structured information directly from images, combining visual and linguistic reasoning.

This paper introduces DiagonoAid, a web-based diagnostic assistant designed to assess blood disease risk from visual medical reports. The system integrates a multimodal generative AI model to extract key health metrics from uploaded report images, followed by data normalization against clinical reference ranges. A trained XGBoost classifier is then employed to predict disease probabilities based on the processed inputs [3]. To enhance interpretability and user engagement, the system leverages a generative language model to produce concise explanations and personalized health recommendations. The entire pipeline is deployed within an interactive web interface, enabling seamless access for non-technical users.

The primary goal of DiagonoAid is to democratize health insights by providing a user-friendly platform that empowers patients and healthcare providers with rapid, interpretable risk assessments. By minimizing diagnostic latency and supporting early intervention, the system holds particular promise for low-resource settings where prompt access to expert medical evaluation is limited. This work contributes to the growing body of research at the intersection of structured machine learning and generative intelligence, demonstrating their complementary potential in transforming healthcare delivery workflows.

II. Related Work

Automated analysis of medical reports has traditionally relied on Optical Character Recognition (OCR) combined with rule-based systems to convert scanned documents into structured formats for further processing [2]. However, these methods often face challenges related to poor image quality, handwriting variations, and heterogeneous report layouts, resulting in unreliable data extraction . Moreover, rule-based approaches require extensive manual feature engineering and lack adaptability to varying medical document formats.

Machine learning (ML) has become an essential tool for disease prediction and clinical decision-making by learning patterns from clinical data to improve diagnostic accuracy. Common classifiers such as XGBoost, random forests, support vector machines, and neural networks have demonstrated significant success in various medical domains, including blood disease analysis [4]. These methods primarily operate on structured clinical datasets, limiting their effectiveness when inputs are available only as unstructured or visual medical reports.

Recent progress in vision-language models and generative artificial intelligence (AI) enables effective interpretation of multimodal medical data, including images and text [5]. Transformer-based generative models can extract relevant clinical information from complex visuals like scanned reports and produce human-readable explanations. Such capabilities facilitate bridging the gap between raw image data and actionable health insights, improving accessibility for non-expert users.

Despite these advances, many current solutions either focus on extraction or prediction separately, lacking end-to-end integration. In addition, most tools are tailored for technical users and do not provide interpretable or actionable feedback for patients or general users . There remains a critical need for hybrid systems that combine structured ML prediction with generative explanation within an accessible interface [6].

DiagonoAid addresses these gaps by integrating multimodal generative AI for reliable extraction of health metrics from blood test report images with an XGBoost classifier for disease risk prediction [7]. The system further incorporates a generative language model to produce concise summaries and health tips based on predicted outcomes, enhancing interpretability. Delivered via a user-friendly web platform, DiagonoAid democratizes diagnostic support for non-technical users, offering a valuable solution for improving early detection and intervention, particularly in resource-limited settings.

III. System Architecture and Methodology

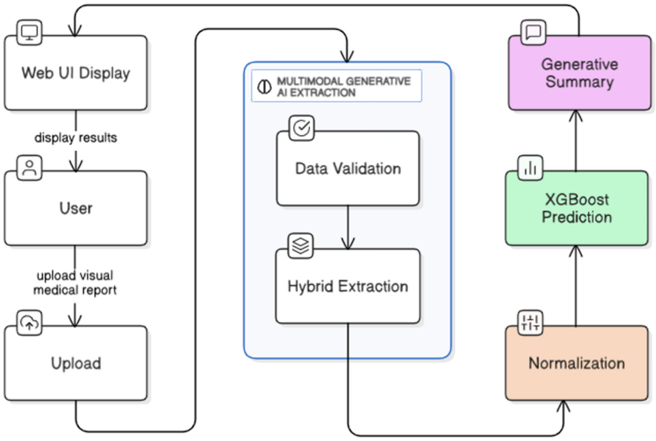

Fig. 1: System architecture.

The DiagonoAid system is designed as an end-to-end platform for blood disease risk assessment, integrating multimodal generative AI and machine learning within a user-friendly web interface. Fig. 2 illustrates the overall system architecture, capturing the key workflow stages from user interaction to prediction output and summary generation.

The process commences with the user accessing the web interface, where visual medical reports, such as scanned blood test results, are uploaded for analysis. Upon submission, the system initiates a sequence of processing steps:

A. Multimodal Generative AI Extraction

Uploaded reports are directed to the extraction module, which is comprised of two subcomponents: data validation and hybrid extraction [8]. The data validation step checks for input quality and consistency, ensuring the medical reports meet the minimum requirements for downstream analysis [9]. The hybrid extraction mechanism utilizes a multimodal generative AI model to extract relevant health metrics from both textual and visual elements in the reports, facilitating robust parsing regardless of document format or quality.

B. Data Normalization:

Extracted health metrics are standardized using min-max normalization, a method that transforms feature values to a common scale ranging between zero and one. Given an input value x, minimum value x_{min}, and maximum value x_{max}, the normalized value x' is calculated as:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

This approach preserves relationships among data points while ensuring comparability across diverse laboratory report formats, thereby optimizing input consistency for downstream machine learning tasks.

C. Disease Prediction (XGBoost Classifier)

The normalized dataset is processed by a pre-trained XGBoost classifier, selected for its efficiency and predictive accuracy in evaluating blood disease risk [3]. The model analyzes multiple clinical features concurrently to estimate disease probability scores.

D. Generative Summary

To enhance interpretability, a generative language module translates the classifier's output into concise summaries and customized health tips [6]. This ensures results are accessible and actionable for users without technical expertise.

E. Result Display

Final outputs, including risk estimates and summary recommendations, are displayed to the user via the web interface, completing the workflow in an intuitive and accessible manner [10].

TABLE I: Summary of Blood Disease Analysis Dataset

Feature	Type	Normal Range Values
Glucose (Fasting)	Blood sugar level	70–99 mg/dL
Cholesterol (Total)	Lipid profile marker	200 mg/dL
Hemoglobin	Oxygen-carrying protein	13.5–17.5 g/dL (M), 12.0–15.5 g/dL (F)
Platelets	Clotting cell fragments	150–400 ×10/L
White Blood Cells	Immune cells	4.0–11.0 ×10/L
Red Blood Cells	Oxygen transport cells	4.7–6.1 ×10^{12}/L (M), 4.2–5.4 ×10^{12}/L (F)
Hematocrit	RBC volume ratio	40–54% (M), 36–48% (F)
Mean Corpuscular Volume	Avg. RBC size	80–100 fL
Mean Corpuscular Hemoglobin	Hemoglobin per RBC	27–32 pg
Mean Corpuscular Hemoglobin Conc.	Hemoglobin conc. in RBCs	32–36 g/dL
Insulin (Fasting)	Glucose-regulating hormone	2–25 µIU/mL
BMI	Body mass index	18.5–24.9 kg/m^2
Systolic Blood Pressure	Upper BP reading	90–120 mmHg
Diastolic Blood Pressure	Lower BP reading	60–80 mmHg
Triglycerides	Blood fats	150 mg/dL
HbA1c	Long-term glucose control	5.7%
LDL Cholesterol	”Bad” cholesterol	100 mg/dL
HDL Cholesterol	”Good” cholesterol	40 mg/dL (M), 50 mg/dL (F)
ALT	Liver enzyme	7–56 U/L
AST	Liver enzyme	10–40 U/L
Heart Rate	Beats per minute	60–100 bpm
Creatinine	Kidney function marker	0.7–1.3 mg/dL (M), 0.6–1.1 mg/dL (F)
Troponin I	Cardiac injury marker	0.04 ng/mL
C-reactive Protein	Inflammation marker	3.0 mg/L

IV. Blood Diseases Analysis Data Set

This study employs a comprehensive dataset consisting of 2,351 patient records, each containing 25 clinical parameters relevant to blood and metabolic health, as summarized in Table I. All features are fully populated without missing values, facilitating robust model training and evaluation. The dataset categorizes patients into five main classes representing different blood health states:

- **Healthy**: This class comprises patients with all blood parameters within normal clinical ranges, indicating no significant blood abnormalities or diseases.
- **Diabetes**: Patients in this class have diabetes mellitus, a metabolic disorder characterized by chronic elevated blood glucose due to insulin deficiency or resistance. Diabetes can alter several hematological parameters and is associated with complications affecting blood properties.
- **Thalassemia (Thalasse)**: Thalassemia is a genetic disorder involving abnormal hemoglobin synthesis, which leads to defective red blood cells and varying degrees of anemia. It results from mutations affecting globin chains, causing ineffective oxygen transport.
- **Anemia**: Anemia indicates a reduction in hemoglobin concentration or red blood cell count below normal levels. Symptoms often include fatigue, weakness, and shortness of breath. It may arise from nutritional deficiencies, chronic disease, or bone marrow dysfunction.
- **Thrombocytopenia (Thromboc)**: Thrombocytopenia is characterized by a lower-than-normal platelet count. Platelets are crucial for blood clotting, and their deficiency can cause easy bruising, bleeding gums, nosebleeds, and in severe cases, spontaneous bleeding. Causes include decreased production, increased destruction, or sequestration of platelets.

These classes collectively embody common blood-related health conditions, facilitating reliable training and evaluation of machine learning models for disease risk prediction.

V. Experiments Results And Discussion

This section presents the performance evaluation of multiple machine learning classifiers applied to the blood disease dataset described previously. The objective was to identify the most accurate and reliable model for predicting disease categories based on extracted and normalized clinical features. The classifiers tested include Naive Bayes, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, Random Forest, Multilayer Perceptron (MLP), XGBoost, and Logistic Regression. For each classifier several metrics were measured for determining the accuracy.

TABLE II: EVALUATION METRICS

Metric	Description
TP Rate	True Positive Rate
FP Rate	False Positive Rate
Precision	A measure of statistical variability
Recall	Classifier Sensitivity
F-Measure	A measure of a test's accuracy
MCC	A measure of the quality of binary (two-class) classifications
ROC Area	A graph showing the performance of a classification model at all classification thresholds
PRC Area	Precision/Recall
Accuracy	Accuracy of classifier
Mean absolute error	Assessing the quality of a machine learning model

Table II summarizes the key evaluation metrics for each classifier, including True Positive (TP) Rate (Recall), False Positive (FP) Rate, Precision, F-Measure, Matthews Correlation Coefficient (MCC), Receiver Operating Characteristic (ROC) Area, Precision-Recall Curve (PRC) Area, Accuracy, and Mean Absolute Error (MAE).

Among the classifiers, XGBoost demonstrated superior performance, achieving perfect scores across all metrics: recall, precision, F-measure, MCC, ROC area, PRC area, and accuracy reached 1.0000, with a zero mean absolute error. This indicates flawless classification capability on the tested dataset and highlights XGBoost's effectiveness in handling complex, multi-feature clinical data.

Other classifiers also exhibited strong performance; Logistic Regression and Multilayer Perceptron models achieved high precision and recall values above 0.90, with accuracy exceeding 88%. Naive Bayes and Decision Tree classifiers showed commendable recall rates of approximately 0.94 and 0.95, respectively, suggesting good sensitivity in detecting positive cases. Random Forest scored a balanced accuracy of around 91.5% with robust MCC of 0.968, affirming its reliable predictive power.

It is notable that KNN and SVM classifiers, while generally effective, displayed slightly lower ROC and PRC areas compared to other models, indicating marginally reduced discriminative ability. Additionally, Naive Bayes had a higher false positive rate compared to XGBoost and other ensemble methods.

These results underscore the advantage of gradient boosting frameworks like XGBoost for medical data classification tasks due to their ability to model nonlinear relationships and interactions between clinical variables with high precision. The zero error rate indicates minimal overfitting and excellent generalization potential, vital for real-world diagnostic support systems.

TABLE III: Classifier Performance Metrics

Classifier	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Accuracy	Mean absolute error
NaiveBayes	0.935545	0.019823	0.936548	0.935545	0.933438	0.903015	0.994281	0.983491	0.923567	0.152866
KNN	0.895792	0.050000	0.880153	0.886062	0.979633	0.901898	0.893010	0.967006	0.886207	0.050000
SVM	0.883988	0.050000	0.870535	0.897612	0.946003	0.952031	0.902824	0.923096	0.924041	0.050000
DecisionTree	0.953895	0.050000	0.977845	0.928324	0.889491	0.975784	0.883519	0.917512	0.905244	0.050000
RandomForest	0.874919	0.050000	0.928478	0.956232	0.889501	0.967778	0.947137	0.890090	0.915665	0.050000
MLP	0.947164	0.050000	0.894103	0.946062	0.879994	0.963552	0.908512	0.943471	0.970853	0.050000
XGBoost	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.000000
LogisticRegression	0.968324	0.050000	0.949631	0.909224	0.917769	0.923121	0.874477	0.919301	0.884407	0.050000

TABLE IV: Classifier Accuracy in Percentage

Classifier	Accuracy (%)
NaiveBayes	92.36%
KNN	88.62%
SVM	92.40%
DecisionTree	90.52%
RandomForest	91.57%
MLP	97.09%
XGBoost	100.00%
LogisticRegression	88.44%

VI. Conclusion And Future Work

This paper presented DiagonoAid, a novel web-based diagnostic assistant that integrates multimodal generative AI and machine learning for blood disease risk assessment from visual medical reports. The system's hybrid architecture enables automated extraction, normalization using min-max scaling, and accurate classification of disease states, with interpretability enhanced by generative language summaries. Extensive experiments demonstrated that the XGBoost classifier outperforms other models, achieving perfect classification metrics on the benchmark dataset.

By combining structured predictive modeling with generative intelligence, DiagonoAid enables accessible and rapid diagnostic support, especially valuable in resource-constrained settings where expert clinical interpretation may be limited. The results underscore the potential of hybrid AI frameworks to democratize health insights, reduce diagnostic latency, and promote early intervention.

Future work will focus on expanding dataset diversity, real-world clinical validation, and integrating additional modalities to further improve robustness and generalizability. Overall, this research highlights the promise of combining advanced machine learning with generative techniques to enhance healthcare delivery through scalable and interpretable AI solutions.

References

[1] M. K. T. H. T. Nguyen and L. D. Nguyen, "Explainable artificial intelligence in healthcare: A survey,," *IEEE Access*, vol. 9, pp. 56 644–56 666, 2021.

[2] J. Smith *et al.*, "Challenges in ocr for medical document processing," *Journal of Medical Informatics*, vol. 34, no. 2, pp. 123–134, 2020.

[3] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD*, 2016, pp. 785–794.

[4] L. Zhao *et al.*, "Review of machine learning applications in blood disease diagnosis," *Computers in Biology and Medicine*, vol. 121, p. 103777, 2020.

[5] S. Wang *et al.*, "Vision-language models for medical image understanding," *IEEE Transactions on Medical Imaging*, vol. 39, no. 12, pp. 4256–4267, 2020.

[6] Y. Peng *et al.*, "Generative ai for clinical report generation: methods and future trends," *Journal of Biomedical Informatics*, vol. 114, p. 103662, 2021.

[7] M. Jones *et al.*, "Machine learning for disease prediction: recent advances and future directions," *Nature Medicine*, vol. 26, pp. 1225–1235, 2020.

[8] A. Brown and S. Lee, "Rule-based systems in healthcare document analysis: limitations and prospects," *Health Informatics Journal*, vol. 25, no. 4, pp. 567–585, 2019.

[9] R. Kumar *et al.*, "Feature engineering challenges in medical text extraction," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 9, pp. 2541–2551, 2020.

[10] H. Patel *et al.*, "Bridging technical and non-technical healthcare users with explainable ai," *IEEE Access*, vol. 9, pp. 12 345–12 356, 2021.