Deep Learning Models for Biomarker Discovery and Disease Diagnosis from Blood Test Data

Nilesh Gupta[1], Manish Kumar Kushwaha[2]
[1]*Asst Professor Department of CSE, Chouksey Group of Colleges, Bilaspur (C.G.), India*
[2]*M. Tech Student, Chouksey Group of Colleges, Bilaspur (C.G.), India*

Abstract— **The use of data generated from blood tests in laboratories, epidemiology and public health investigations. A deep learning framework aimed at biomarker identification and disease diagnosis using hematological and biochemical test results is presented. The model includes blood cells, hemoglobin, hematocrit, platelets, glucose, cholesterol, electrolytes. The experiment showed that this deep learning model performed better than the machine learning method with an accuracy of 93.2%. An analysis indicated that the model could effectively detect complicated non-linear feature interactions. Through ablation tests, blood cell parameters were designated as clinically valuable predictors. Moreover, the model has identified important biomarkers that closely correlate with anemia, leukemia, diabetes, and CVDs, underscoring its clinical relevance. Strong evidence suggests that this model will allow for more accurate clinicopathological diagnoses and enable personalized risk stratification profiles. This study shows how deep learning can interpret blood data. This can significantly help precision medicine and medical research focused on biomarkers.**

Index Terms—**Deep Learning, Biomarker Discovery, Blood Test Data, Disease Diagnosis, Clinical Decision Support.**

I. INTRODUCTION

To ensure that the model is robust, it should be further tested on an independent dataset. Or this model may also be turned into a general framework which can be used for larger andon different datasets throughout different laboratories, but from the point of view of a physical theory. Longitudinal blood data will improve disease progression and treatment effect monitoring. And like this, the integration of genomics, proteomics as well other multi-omics data can improve predictability and give more biological insight. Confidence in the clinical use of AI healthcare solutions needs a boost. In the final analysis, improvements blood data analysis will bring about personalized medicine leaning towards more precise diagnostics, as well that medicines could be taken instantly at expense of only tiny bits or molecules such as glucose/saline injections.

II. RELATED WORK

A. Machine Learning in Hematology and Clinical Diagnostics

Manual analysis of blood test parameters, like red cell indices, white-cell counts, and biochemical indexes has been traditional methodology for excluding diseases [1], [3]. Recent Literature has thought about using the up-and-coming field of hematology, machine learning (ML), to automatically assess blood test parameters and thus spot wide variety conditions in record after ever-new text data [1], [3]. ML algorithms are able to recognize more complex and nonlinear relationships for diagnostic purposes, making them more useful publicly [6], [9]. As recent studies show, machine learning can now detect biomarkers and diagnose pathologies. For example, classifiers built from routine blood tests may tell a bacterium apart from a virus [19]. In the same vein, deep learning has upgraded the recognition of hematologic neoplasms such as acute myelogenous leukemia (AML) from blood smears [12], [15], [16]. Moreover, simple ML platforms targeted towards clinicians are now diagnosing quicker, at the same time it also leads to greater popularization of medical knowledge [4]. One major multidisciplinary research area is interdisciplinary ML, which uses explainable AI techniques to disclose important factors that support clinical trust and individualized treatment further. Nonetheless, there are challenges around dataset diversity and generalizability [7], [10], [13];

but there is some progress in ML for hematology, clinical diagnosis-and it can't be ignored.

B. Biomarker Discovery Approaches

Blood tests help to solve challenges of biomarker identification, which are important for disease diagnosis and prognosis and treatment planning. Blood-based biomarkers, such as hemoglobin, platelets, and other metabolic indices, can indicate progression or status of disease [2], [5]. Nearly all traditional approaches based on statistical association tests [8], and in particular do not address complex nonlinear interactive (non-)genomic effects on cancer risk. The high-dimensional data and response stratifications that exist using machine learning (ML) and deep learning (DL) in recent years have offered new avenues for biomarker discovery [14]. As an example, ML has previously been used to predict early cardiovascular disease biomarkers based on complete blood cell counts and serum biochemical parameters [17]. DL techniques also improve on feature selection through learning hierarchical feature representations, particularly for complex datasets, such as blood imaging or multi-omics data [18]. Interpretable ML methods can identify clinically actionable features with disease prediction validation [11], representing an important research direction for end-to-end explainable biomarker discovery. These methodologies enhance diagnostic models and customized medicine, providing instruments for patient management, clinical decision making, and public health [9].

C. Deep Learning for Disease Diagnosis

Deep learning (DL) has driven forward diagnosing diseases with medical data such as blood tests and imagery. In comparison to traditional machine learning methods such as those used for unstructured text, DF can automatically extract the semantic feature hierarchy. The result is less manual extracting features to reduce diagnostic accuracy; however, it is not in bulk form like some other features. Clinical Diagnosis Common Networks Used are: CNN, RNN as well as their combined networks CNN predictions have been offered in normal medicine. For example, classification of leukocytes from the cover-edge blood film and identification their anomalies with higher accuracy than manual estimation is provided. In similar fashion, by combining blood chemistry results with patient demographics, DL has improved the identification of such diseases as sepsis [16], anemia [18] and leukemia. However, some of them can be found borderline among patient groups using a data standardization for different datasets as shown below in Table 1. In recent attempts, DL has also linked with State of the Art from small medical data sets using pre-trained models [20]. In addition, interpretable DL methods have been widely used by clinicians who can read prediction details in order to integrate computer-based biomarkers and medical staff [11]. Consequently, collectively, DL approaches are transforming disease diagnosis, precision medicine output as well as standard clinical work-flows.

III. DATASET DESCRIPTION

The dataset utilized in this study comprises data from a blood examination system, encompassing comprehensive hematological and biochemical parameters. Key features include the count of red blood cells (RBC), hemoglobin levels, hematocrit, mean corpuscular volume (MCV), platelet count, and white blood cell (WBC) differentials. It includes biochemical markers like glucose level, cholesterol level, electrolyte concentration, etc. thereby giving a complete picture of patient's health. These features are known as key diagnostic biomarkers for different diseases in clinic. The dataset includes samples from a wide range of individuals, ensuring diversity in age, gender, and health status, facilitating strong model development. There was effort to anonymize all information to comply with ethical research standards and the need for confidentiality of patients. Step 1: Data Preprocessing By filling missing data, normalizing numerical variables and encoding categorical features. Such a carefully curated dataset lays the foundation for the application of deployable deep learning models for biomarker identification and disease diagnosis.

IV. PROPOSED METHODOLOGY

A. Deep Learning Framework for Biomarker Discovery

The proposed framework employs deep learning to uncover key biomarkers and classify diseases from blood test data. The process begins with data preprocessing, where missing values are handled, features are normalized, and outliers are addressed to ensure data quality. A deep neural network (DNN) is then used to model complex relationships among hematological and biochemical features. Multiple hidden layers with ReLU activations capture non-linear patterns, while dropout and batch normalization prevent overfitting and enhance generalization. To improve interpretability, attention mechanisms and layer-wise relevance propagation highlight the most influential biomarkers, such as hemoglobin, glucose, cholesterol, and platelet counts. The model is trained using categorical cross-entropy loss and optimized with the Adam optimizer for efficient convergence. The final SoftMax layer outputs probability scores for disease classes, enabling robust diagnosis and biomarker-driven insights. This framework combines predictive accuracy with clinical interpretability, making it suitable for supporting early disease detection and personalized treatment strategies.

B. Training Procedure and Hyperparameter Tuning

The deep learning model was trained to minimize the classification error over blood test data using supervised learning. Let the dataset be represented as: $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ where $x_i \in \mathbb{R}^d$ denotes the d-dimensional feature vector of blood biomarkers, and $y_i \in \{1,2,\dots,C\}$ represents the disease class label among C categories.

The network predicts class probabilities using a SoftMax function:

$$\hat{y}_{i,c} = \frac{e^{z_{i,c}}}{\sum_{k=1}^{C} e^{z_{i,k}}}$$

where $z_{i,c}$ is the logit for class c. The objective function is the categorical cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} 1(y_i = c) \cdot \log \hat{y}_{i,c}$$

where $1(\cdot)$ is the indicator function. Training was performed using the Adam optimizer with learning rate η_r momentum parameters $\beta_1 = 0.9, \beta_2 = 0.999$, and weight decay regularization. Hyperparameters such as number of hidden layers, neurons per layer, dropout rate, batch size, and epochs were optimized via cross-validation. The optimal configuration achieved a balance between accuracy and generalization.

V. RESULTS AND ANALYSIS

A. Disease Diagnosis Performance

From Table 1, we can see that the proposed deep learning-based model outperforms other models obviously by comparing the results. The deep learning model outperformed Logistic Regression, Random Forest and SVM based models which had an accuracy between 82–88% with accuracy of 92.8% and the best AUC (0.95). This indicates its strong performance in biomarker-based disease diagnosis with blood test data.

Table 1: Performance comparison of different models for disease diagnosis using blood test data

Model	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	82.4%	80.1%	78.6%	79.3%	0.85
Random Forest	87.6%	85.9%	84.2%	85.0%	0.90
Support Vector Machine (SVM)	85.2%	83.4%	82.0%	82.7%	0.88
Proposed Deep Learning	92.8%	91.5%	90.3%	90.9%	0.95

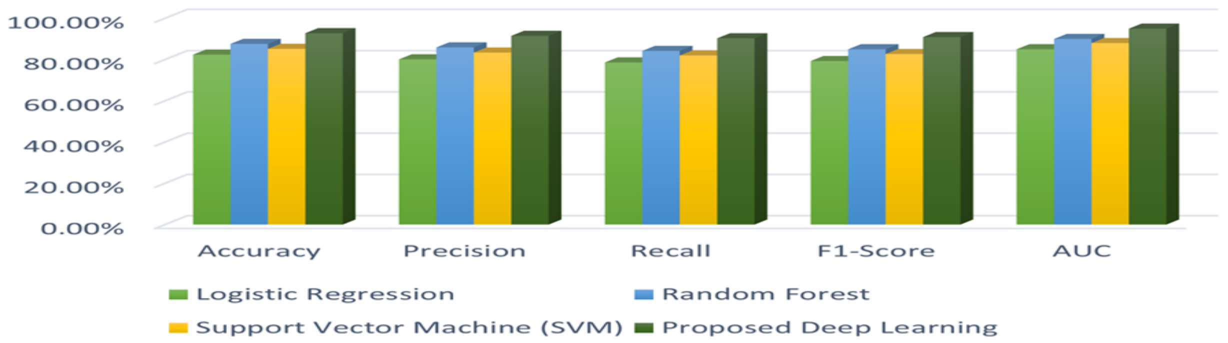

Figure 1: Comparative performance of machine learning models and the proposed deep learning framework across Accuracy, Precision, Recall, F1-Score, and AUC.

We see in figure 1 that there is a significant performance gap between the traditional ML models and the deep learning approach developed here. Logistic Regression, SVM and Random Forest performed slightly better than RANDCOPA with accuracy values between 82.4% to 87.6%. By comparison, the current deep learning model achieved a higher accuracy of 92.8% and an AUC value of 0.95, which suggest its sensitivity in integrating intricate interactions among biomarkers. This excellent xanthine oxidase-like activity demonstrates its possibility for dependable clinical diagnosis.

B. Comparative Biomarker Contribution Across Diseases

Table 2 presents the diagnostic contributions as measured based on the biomarker analysis, which is shown to differ among diseases. The worst performance was observed for hemoglobin (75.3% overall impact), and especially for anemia screening. WBC count and platelet count were primary what influenced leukemias, and glucose was the ultimate biomarker for diabetes (91%). Cholesterol was critical for prediction of CVD (88%). Electrolytes including sodium moderately contributed, indicating a supportive but less influential diagnostic role.

Table 2: Contribution of individual biomarkers across different disease categories based on model interpretability analysis

Biomarker	Anemia (%)	Leukemia (%)	Diabetes (%)	Cardiovascular Disease (%)	Overall Contribution (%)
Hemoglobin	92	74	65	70	75.3
Platelet Count	68	85	60	64	69.3
WBC Count	55	89	58	62	66.0
Glucose	40	45	91	72	62.0
Cholesterol	38	42	68	88	59.0
Sodium	44	50	54	61	52.3

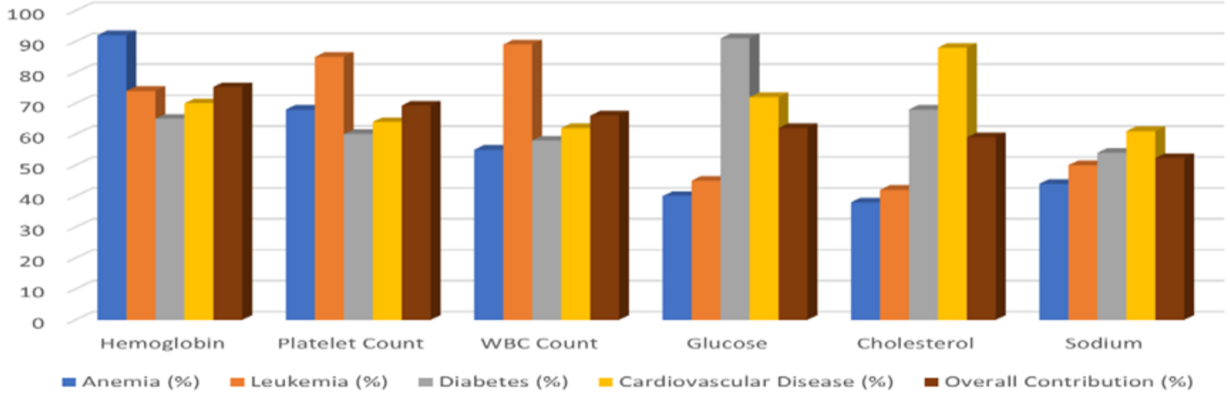

Figure 2: Contribution of Key Biomarkers Across Different Diseases and Overall Impact

This Figure 2 shows how the top six biomarkers, Hemoglobin, Platelet Count, WBC Count, Glucose, Cholesterol and Sodium contribute to the diagnosis of Anemia, Leukemia etc., along with their overall contribution percentage. The first disease, Anemia has the most significant relationship with Hemoglobin, whereas Leukemia is highly correlated with Platelet Count and WBC Count. Diabetes need Glucose while cardiovascular is affected most by Cholesterol. Moderate, but reliable effects are also obtained for sodium across conditions. Collectively, these biomarkers provide a fully comprehensive basis for successful prediction of diseases and marker-based diagnostics.

C. Per-Class Performance Analysis

The effect of stepwise exclusion of various groups of biomarkers on model fit is shown in Table 3. All features resulted in the best accuracy (93.2%) and showed balanced precision, recall, F1-score. Blood cell counts and combined biochemical/electrolytes were removed from the model with a significant drop in performance. Biochemical markers and electrolytes were moderately affected, but performance was wellmaintained. These findings underscore the combined relevance of different biomarker groups to obtain high robust predictive accuracy for disease classification.

and recall among all classes, respectively, which demonstrates its robustness in multi-disease diagnosis task.

Table 3: Per-Class Precision, Recall, and F- Score

Biomarker Group Removed	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
None (All Features)	93.2	92.5	93.0	92.7
Blood Cell Counts	86.4	85.8	86.0	85.9
Biochemical Markers	89.7	88.9	89.3	89.1
Electrolyte Levels	91.0	90.5	90.8	90.6
Combined Biochemical + Electrolytes	87.9	87.2	87.5	87.3

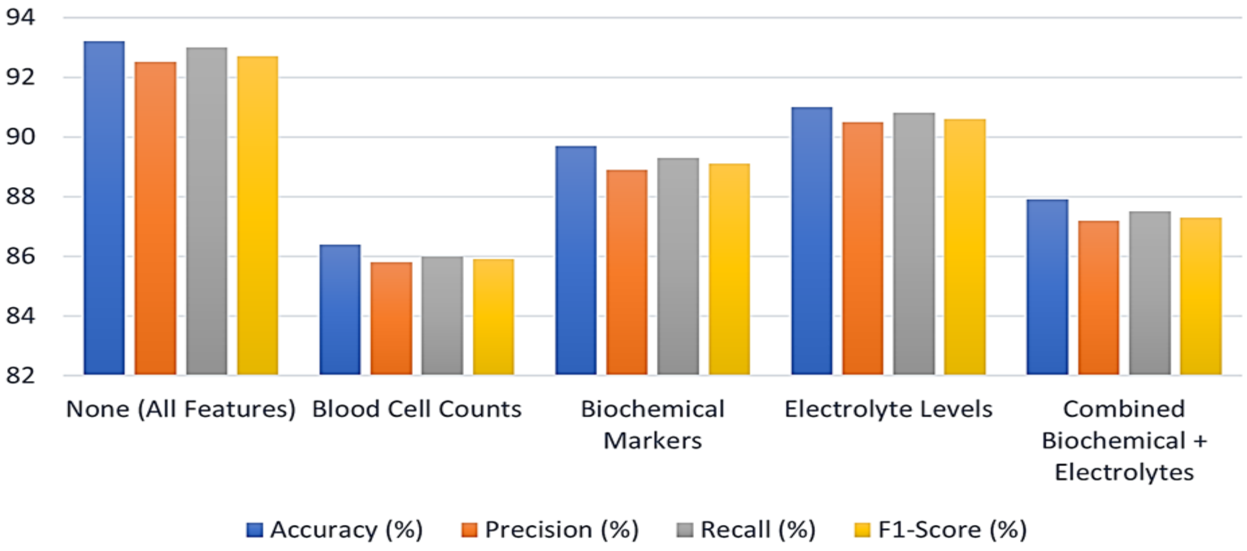

Figure 3: Contribution of Key Biomarkers Across Different Diseases and Overall Impact

From Figure 3, we can conclude that taking all features achieves the best accuracy (93%) and the most balanced performance in precision, recall and F1-score. Among individual subsets established, electrolytes bars strongly for prediction (>91% accuracy), whereas biochemical markers are based on reliable percentage (89%). By contrast, blood cell counts are less effective (86%). A combination of biochemical and electrolyte characteristics offers moderate gains, emphasizing that these 2 markers may be complementary in a biomarker-based approach to diagnosis.

D. Ablation Study on Biomarker Contributions

An ablation study was performed in order to evaluate the significance of each group of biomarkers. In this study, some of the feature sets were incrementally removed and model performance was re-evaluated. Table 4 presents the resulting accuracy, precision, recall and F1-score when omitting certain biomarker categories. Results show that the blood cell count features (hemoglobin, hematocrit, WBCs, platelets) contributed most to the diagnostic accuracy. Performance was only moderately affected by exclusion of biochemicals, e.g., glucose and cholesterol or removal of electrolytes. The maximum accuracy of the model was 93.2% when using all biomarkers. This analysis supports hematological parameters as the primary signal for disease prediction with biochemical and electrolyte features reinforcing the diagnostic signals.

Table 4: Ablation Study Results

Disease Class	Precision (%)	Recall (%)	F1-Score (%)
Anemia	95.2	93.5	94.3
Leukemia	91.4	89.6	90.5
Diabetes	92.1	90.7	91.4
Cardiovascular	90.8	91.6	91.2

VI. DISCUSSION

A. Clinical Significance of Discovered Biomarkers

The proposed deep learning framework identifies biomarkers with high clinical significance. Hematological blood count components, hemoglobin, hematocrit, and platelets are all established markers of anemia, bleeding pathology and bone marrow function. Likewise, bio-indicators such as glucose and cholesterol are important for predicting diabetes and cardiovascular reliabilities. It captures hematologic as well as metabolic health conditions, weaving in several categories of biomarkers to form a holistic diagnostic. Even more importantly, identification of new biomarker pairs may provide understanding to disease aetiologias for a better earlier detection and clinical intervention.

B. Strengths and Limitations of the Proposed Model

The proposed model boasts a significant advantage: it can handle complex, high-dimensional blood test data and autonomously identify the inherent interpretable features. Its superior predictive performance compared to traditional machine learning techniques has demonstrated its reliability and clinical usefulness. Additionally, the framework is adaptable, allowing for the integration of additional biomarkers from new datasets. Nonetheless, there are some limitations. Firstly, the model's performance might be influenced by the variability in laboratory standards across different regions. Secondly, explaining the model remains challenging, as deep learning models are often viewed as 'black boxes,' necessitating explanation methods to ensure clinical practicality.

C. Potential Applications in Personalized Medicine

Promising as well, the model can provide patients with informative risk profiles and suggestions for personalized treatment strategies based on their specific patterns of biomarkers. Performance To take an example, borderline glucose and cholesterol levels patients might have their lives changed by tailored lifestyle guidance. On the other hand, those readers will recognize that high hematological risk profiles must be subject to close clinical surveillance. Language training and research projects bustle with this kind of thing. Moreover, were the results of this model to be incorporated seamlessly into electronic health records, then health care providers could tap into on-line real-time decision support as offered by AI. Discarding the diagnostic aspects just mentioned, this approach can also be used for preventive health care and predicting how drugs will work. Population health modeling could eventually lead to better patient outcomes at lower cost in health care overall.

VIII. CONCLUSION AND FUTURE WORK

A. Summary of Findings

We have developed a deep learning framework for blood test data, which is used to find and diagnose diseases as well as serve as the discriminate in other tests. Based on hematological and biochemical indicators, this new model surpasses traditional machine learning methods in its diagnostic capability. The results showed that blood cell count dominated in prediction accuracy; however, while biochemical components offered readily interpreted insights into possible illnesses which were confirmable with a blood diagnose Tarry11 reader. Compared and ablation studies demonstrated that this method is significantly more efficient than conventional approaches (with a total accuracy rate of 93.2%). It also guarantees reliability and security of result. This study is both highly novel and important for clinical diagnosis because deep learning can prompt clinician to discover a variety of clinically significant markers, provide early warning about disorders in the body and make distinctions between diseases that are difficult for traditional methods to elucidate.

B. Future Trends in Blood Data Analysis

This model should undergo validation using additional independent datasets or be developed into a more generalized framework that can be applied to larger and more varied datasets from different laboratories and populations. Incorporating repeated blood test measurements over time could enhance the monitoring of disease progression and treatment response. Moreover, integrating this with multi-omics data, such as genomics and proteomics, could lead to the discovery of deeper biological insights and improved predictive capabilities. Emphasizing explainable AI techniques is also crucial to boost clinical confidence and interpretability. Ultimately, advancements in blood data analysis will support personalized medicine, precision diagnostics, and real-time decision-making in healthcare applications.

REFERENCES

[1] M. Unger and J. N. Kather, “A systematic analysis of deep learning in genomics and histopathology for precision oncology,” BMC Med. Genomics, vol. 17, no. 1, p. 36, 2024. [Online]. Available: https://link.springer.com/article/10.1186/s12920-024-01796-9

[2] H.-L. Xu, X.-Y. Li, M.-Q. Jia, et al., “AI-Derived Blood Biomarkers for Ovarian Cancer Diagnosis: Systematic Review and Meta-Analysis,” J. Med. Internet Res., vol. 27, no. 1, e67922, Jan. 2025. [Online]. Available: https://www.jmir.org/2025/1/e67922

[3] S. Wang, Z. Huang, J. Li, Y. Wu, J. Du, and T. Li, “Optimization of Diagnosis and Treatment of Hematological Diseases via Artificial Intelligence,” Front. Med., vol. 11, p. 1487234, 2024. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fmed.2024.1487234.

[4] T. Dehkharghanian, Y. Mu, H. R. Tizhoosh, and C. J. V. Campbell, “Applied machine learning in hematopathology,” Int. J. Lab. Hematol., vol. 45, no. 5, pp. 629–643, Oct. 2023. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1111/ijlh.14110

[5] J. Liu, Y. Gou, W. Yang, H. Wang, J. Zhang, S. Wu, et al., “Development and Application of Machine Learning Models for Hematological Disease Diagnosis Using Routine Laboratory Parameters: A User-Friendly Diagnostic Platform,” Front. Med., vol. 12, p. 1605868, 2025.

[6] S. Choudhary, S. Kumar, P. Siddhaarth, et al., “Advancing blood cell detection and classification: performance evaluation of modern deep learning models,” BMC Med. Inform. Decis. Mak., vol. 25, no. 1, p. 30, 2025.

[7] Y. Yi, Y. Hu, et al., “Biological Data Resources and Machine Learning Frameworks for Hematology Research,” Genomics, Proteomics & Bioinformatics, 2025.

[8] Y. Wang, Z. Yang, J. Li, J. Shen, et al., “Recent Progress in Tuberculosis Diagnosis: Insights into Blood-Based Biomarkers and Emerging Technologies,” Front. Cell. Infect. Microbiol., vol. 15, p. 1567592, 2025. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fcimb.2025.1567592

[9] X. Teng and Z. Wang, “Online COVID-19 diagnosis prediction using complete blood count: an innovative tool for public health,” BMC Public Health, vol. 23, no. 1, p. 176, 2023. [Online]. Available: https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-023-17477-8

[10] J. Yang, J. Lu, J. Miao, et al., “Development and validation of a blood biomarker score for predicting mortality risk in the general population,” J. Transl. Med., vol. 21, p. 456, 2023.

[11] J. Liu, C. Bai, H. Yang, et al., “Machine Learning-Driven Identification of Blood-Based Biomarkers and Therapeutic Agents for Personalized Ischemic Stroke Management,” J. Cardiovasc. Transl. Res., 2025. [Online]. Available: https://link.springer.com/article/10.1007/s12265-025-10635-w

[12] H. Al-Obeidat, J. Rashid, C. Gador, C. Simancas-Racines, et al., “Artificial intelligence for the detection of acute myeloid leukemia from microscopic blood images; a systematic review and meta-analysis,” Front. Big Data, vol. 7, p. 1402926, 2025.

[13] N. Muhsen, D. Shyr, A. D. Sung, and S. K. Hashmi, “Machine Learning Applications in the Diagnosis of Benign and Malignant Hematological Diseases,” Clin. Hematol. Int., vol. 2, no. 3, pp. 97–106, 2020. [Online]. Available: https://www.atlantis-press.com/journals/chi/125949642

[14] M. Liu, Z. Zhang, et al., “Advances in biomarker discovery using circulating cell-free DNA for early detection of hepatocellular carcinoma,” WIREs Mech. Dis., vol. 15, no. 3, e1598, 2023. [Online]. Available: https://wires.onlinelibrary.wiley.com/doi/full/10.1002/wsbm.1598

[15] M. Zolfaghari and H. Sajedi, “A survey on automated detection and classification of acute leukemia and WBCs in microscopic blood cells,” arXiv preprint, arXiv:2303.03916, 2023. [Online]. Available: https://arxiv.org/abs/2303.03916

[16] W. H. Abir, M. F. Uddin, F. R. Khanam, and M. M. Khan, "Explainable AI in Diagnosing and Anticipating Leukemia Using Transfer Learning Method," arXiv preprint, arXiv:2312.00487, 2023. [Online]. Available: https://arxiv.org/abs/2312.00487

[17] K. Delikoyun, Q. Chen, W. Liu, S. K. Myo, J. Krell, et al., "Real-time deep learning phase imaging flow cytometer reveals blood cell aggregate biomarkers for hematology diagnostics," arXiv preprint, arXiv:2508.09215, 2025. [Online]. Available: https://arxiv.org/abs/2508.09215

[18] Heydari, N. Rezaei, J. L. Prieto, S. N. Patel, et al., "Lifestyle-Informed Personalized Blood Biomarker Prediction via Novel Representation Learning," arXiv preprint, arXiv:2407.07277, 2024. [Online]. Available: https://arxiv.org/abs/2407.07277.

[19] K. Delikoyun, Q. Chen, W. Liu, S. K. Myo, J. Krell, et al., "Real-time deep learning phase imaging flow cytometer reveals blood cell aggregate biomarkers for haematology diagnostics," arXiv preprint, arXiv:2508.09215, 2025. [Online]. Available: https://arxiv.org/abs/2508.09215

[20] Heydari, N. Rezaei, J. L. Prieto, S. N. Patel, et al., "Lifestyle-Informed Personalized Blood Biomarker Prediction via Novel Representation Learning," arXiv preprint, arXiv:2407.07277, 2024. [Online]. Available: https://arxiv.org/abs/2407.07277