BloodPatrol: Revolutionizing Blood Cancer Diagnosis - Advanced Real-Time Detection Leveraging Deep Learning & Cloud Technologies

Jinhang Wei, Longyue Wang, Zhecheng Zhou, Linlin Zhuo, Xiangxiang Zeng, Xiangzheng Fu, Quan Zou, *Senior Member, IEEE*, Keqin Li, *Fellow, IEEE*, and Zhongjun Zhou

***Abstract*—Cloud computing and Internet of Things (IoT) technologies are gradually becoming the technological changemakers in cancer diagnosis. Blood cancer is an aggressive disease affecting the blood, bone marrow, and lymphatic system, and its early detection is crucial for subsequent treatment. Flow cytometry has been widely studied as a commonly used method for detecting blood cancer. However, the high computation and resource consumption severely limit its practical application, especifically in regions with limited medical and computational resources. In this study, with the help of cloud computing and IoT technologies, we develop a novel blood cancer dynamic monitoring diagnostic model named BloodPatrol based on an intelligent feature weight fusion mechanism. The proposed model is capable of capturing the dual-view importance relationship between cell samples and features, greatly improving prediction accuracy and significantly surpassing previous models. Besides, benefiting from the powerful processing ability of cloud computing, BloodPatrol can run on a distributed network to efficiently process large-scale cell data, which provides immediate and scalable blood cancer diagnostic services.**

***Index Terms*—Cloud computing in healthcare, IoT in medical diagnostics, blood cancer diagnosis, intelligent feature weight fusion mechanism, flow cytometry data analysis.**

I. Introduction

THE Internet of Things (IoT) and cloud computing play a significant role in cancer diagnosis. IoT supports seamless interaction and communication between devices and people, promoting healthcare solutions and improving quality of life [1], [2], [3]. IoT-based healthcare systems utilize smart interconnected devices and sensors to provide cancer medical services and business analysis/cloud services [4]. Cloud computing is used to store, process, and handle large volumes of medical data in a highly secure manner [5]. It enables the analysis and processing of soft copies of CT/MRI scans for cancer detection and diagnosis [6], [7], [8]. Furthermore, cloud computing is also used for data transmission, reporting, and enhances cancer treatment through actionable insights and decisions [9]. IoT platforms for cancer treatment focus on reducing costs associated with diagnosis and detection, especially for breast cancer, skin cancer, and lung cancer. IoT technologies, including wearable devices, can monitor and manage the adverse effects, symptoms, and quality of life of cancer patients undergoing treatment [9].

Early diagnosis of blood cancer has a significant impact on treatment outcomes, increasing the chances of a good prognosis and complete remission, and allows for the selection of precision-based personalized immunotherapies [10], [11]. Flow cytometry plays a crucial role in the diagnosis of blood cancers, especially in hematological diseases. It can analyze various diagnostic samples and differentiate between different cell groups based on the expression of specific antigens. Flow cytometry is also used for detecting minimal residual disease [12], which is important for monitoring treatment response. Additionally, flow cytometry is increasingly used for detecting and analyzing circulating tumor cells in the blood of cancer patients. The enhanced capacity for high throughput and the recovery of viable cells for subsequent molecular analysis renders it an indispensable instrument in the fields of cancer research and diagnostics [13]. As a high-throughput technique, flow cytometry can perform

Received 19 December 2023; revised 22 October 2024; accepted 1 November 2024. Date of publication 11 November 2024; date of current version 5 June 2026. This work was supported by the National Natural Science Foundation of China under Grant 62302339. *(Corresponding authors: Linlin Zhuo; Xiangzheng Fu.)*

Jinhang Wei, Zhecheng Zhou, and Linlin Zhuo are with the College of Data Science and Artificial Intelligence, Wenzhou University of Technology, Wenzhou 325027, China (e-mail: zhuoninnin@163.com).

Longyue Wang is with the Tencent AI Lab, Shenzhen 518057, China.

Xiangxiang Zeng and Xiangzheng Fu are with the College of Computer Science and Electronic Engineering, Hunan University, Changsha 410012, China (e-mail: fxz326@hnu.edu.cn).

Quan Zou is with the Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 611730, China.

Keqin Li is with the Department of Computer Science, State University of New York, New Paltz, NY 12561 USA.

Zhongjun Zhou is with the School of Biomedical Sciences, The University of Hong Kong, Hong Kong, China.

We have also created a cloud diagnostic platform to facilitate access to our work, the latest access link and updates are available at: https://github.com/kkkayle/BloodPatrol.

Digital Object Identifier 10.1109/JBHI.2024.3496294

multi-parametric analysis on millions of cells at the single-cell level, providing exhaustive data on each cell's volume, morphology, and specific surface antigens [14]. Given this, the large number of cells and the multi-dimensional information output generate massive datasets, whose processing and interpretation depend on complex data analysis software, requiring advanced computational technologies and expertise [15]. Especially in environments with scarce medical resources, this technological threshold poses a significant barrier to diagnostic efficiency and accuracy for patients, potentially leading to delayed diagnosis and adversely affecting the prognosis of the condition [16].

Deep learning technology has been widely applied in cancer diagnosis, especially in the field of medical imaging. It is often used to identify and classify normal and abnormal patterns in medical images, such as automatically distinguishing between normal and pathological changes like pneumonia or tumors in X-ray or CT images [17]. In addition to medical imaging, deep learning technology is also applied in genomics, drug discovery, survival analysis [18], [19], [20], and other fields. Deep learning models can extract useful features from complex biological data, helping doctors and researchers better understand the occurrence, development, and treatment of cancer [21]. For example, deep learning models can predict the subtype and prognosis of tumors by analyzing genetic mutations in tumors and assist in selecting adjuvant therapies [22]. In the field of blood cancer diagnosis research, deep learning technology has shown its potential. For instance, CSNN [23] is a cellular-level blood cancer discrimination network, and DeepCNN [24] and CellCNN [25] effectively analyze flow cytometry data based on convolutional neural networks [26], which can be applied in blood cancer diagnosis. However, these models still face significant challenges in the precise diagnosis of blood cancer. They mainly focus on the analysis of cell data and require a fixed number of cell data during diagnosis; otherwise, truncation or padding operations must be undertaken, which affect the predictive performance of the models and limit their application in assisting doctors in making diagnostic and therapeutic decisions. Therefore, the application of these models in clinical decision support systems requires further research and development.

Against the current backdrop of blood cancer diagnosis, we propose a dynamic blood cancer monitoring and diagnostic model named "BloodPatrol". The uniqueness of this model lies in its ability to process any number of cells in flow cytometry data without the need for truncation or padding. This feature allows BloodPatrol to perform real-time diagnosis based on the real-time data provided by the flow cytometer. In designing the model, we fully considered the characteristics of single-cell samples and cell feature aspects. The dual-view model design greatly improves the model's predictive accuracy, thereby enhancing its potential application in assisting diagnosis. Additionally, traditional model interpretation methods, such as Shapley value analysis, are difficult to apply effectively to ultra-large-scale flow cytometry data. Based on dynamic flow cytometry data, we propose an interpretability analysis method called Cell Population Impact Analysis (CPIA), which identifies abnormal cell populations from the model's predicted probabilities. Then, this information about abnormal cell populations can be fed back to medical personnel, assisting them in making more accurate judgments in diagnostic and therapeutic decision-making.

Specifically, we use an encoder based on the transformer [27] architecture to efficiently encode features from an arbitrary number of streaming single-cell data, employing an all-pair attention mechanism to effectively capture key features within the dimension of single-cell samples. Subsequently, a dynamic max-pooling layer is used to extract significant features from the cell samples and integrate the variable-length single-cell data into a fixed-size feature matrix. This is followed by the use of a feature weight fusion network for adaptive weight fusion of the pooled cell sample dimensions, a process aimed at learning the interactions within single-cell samples. Finally, we employ a 1D Convolutional Neural Network (CNN) pooling layer to aggregate single-cell feature information. This process not only merges the importance relations within the cell feature dimensions but also further reduces the feature dimensions, simplifies the network structure, and enhances computational efficiency. Most importantly, this method significantly reduces the risk of model overfitting. Considering the presence of some indistinguishable samples in the blood cancer dataset, we introduced the PolyLoss [28] loss function. PolyLoss helps the model to more accurately identify and classify blood cancer by more effectively focusing on these hard-to-classify samples through adjusting the weight of higher-order terms in the loss function. Additionally, we found in our research that smaller experimental datasets could lead to a higher variance in the model's prediction results. On this basis, we proposed a model ensemble layer, which combines multiple models from cross-validation, effectively enhancing the model's prediction stability and further improving predictive performance. Our contributions in this work can be summarized as follows:

1) We proposed a dynamic blood cancer monitoring and diagnostic model named BloodPatrol, based on a transformer encoder and intelligent weight fusion module, significantly surpassing existing blood cancer diagnostic models in multiple aspects of performance.
2) BloodPatrol is capable of rapid diagnosis based on real-time changing streaming cell data and utilizes CPIA technology to effectively identify and label abnormal cell populations, providing more accurate and timely decision support for medical professionals in diagnosis and treatment.
3) This study is the first to utilize cloud computing and IoT technologies for addressing the significant computational challenges in processing millions of cellular data, offering scalable, real-time diagnostics in blood cancer, and introducing a novel paradigm in blood cancer diagnosis.

II. Materials and Methods

A. Problem Definition

Sample representation: Consider a test sample composed of flow cytometry data, denoted as X. This sample contains n single-cell data points, where n is variable. Each cell in the sample is represented by a multidimensional vector X_i ($0 \leq i < n$), reflecting cell characteristics based on the configuration of

the flow cytometer. The vector length of X_i may vary across different datasets, corresponding to different flow cytometer configurations.

Target label: Each sample X is associated with a binary label Y, where $Y = 0$ indicates a healthy (non-diseased) sample, and $Y = 1$ indicates a diseased sample.

Task objective: Given a sample X, our goal is to use a model to predict the probability of the sample being diseased. This is a binary classification problem, where the probability of disease presence is inferred from the attributes of the cells within the sample.

Evaluation metric: When assessing the performance of our model, we set the threshold for Y at 0.5. Predictions exceeding this threshold will classify the sample as diseased, while those below this threshold will classify the sample as healthy.

B. Datasets

The first dataset CLL [23]: Provided by the Diagnostic Laboratory of the Center for Advanced Laboratory Medicine (CALM) at the University of California, San Diego (UCSD), it is used to identify cases of Chronic Lymphocytic Leukemia (CLL). DS1 includes FCM data from 288 subjects, of which 186 were diagnosed with CLL, and 102 were judged as non-CLL. In the clinical FCM testing of each subject, two reagent panels, PB1 and PB2, were used, each containing antibodies for detecting 10 markers (fluorescent parameters): PB1 includes CD3, CD5, CD10, CD19, CD22, CD38, CD43, CD45, CD79b, and CD81; PB2 includes anti-Ig-lambda, anti-Ig-kappa, CD5, CD7, CD19, CD20, CD23, CD38, CD49d, and FMC-7. Each dataset also includes 6 scatter parameters: Forward Scatter (FSC) - Area (A)/Height (H)/Width (W) and Side Scatter (SSC) - A/H/W.

The second dataset B-ALL [23]: Provided by the Diagnostic Laboratory of Stanford University, it is used to identify leukemia cells (blast cells) in B-cell Acute Lymphoblastic Leukemia (B-ALL). The samples in DS2 are bone marrow samples, including diagnostic samples and samples collected from patients undergoing treatment with CD19-targeted Chimeric Antigen Receptor (CAR) T-cell therapy. DS2 includes FCM data from 178 subjects, where 50 were diagnosed with B-ALL, and 128 were judged as non-B-ALL. The FCS files of DS2 come from a reagent panel testing nine markers: CD66b, CD22, CD19, CD24, CD10, CD34, CD38, CD20, CD45, as well as two scatter parameters (FSC/SSC).

C. BloodPatrol

The BloodPatrol model utilizes flow cytometry data, employing a transformer-based encoder to efficiently process large-scale, variable cell data. This encoder adapts to the inherent variability in sequence length characteristic of flow cytometry, capturing critical cellular information through an all-to-all attention mechanism. The subsequent intelligent feature weighting and fusion module refines this process, selectively emphasizing prominent cell features to enhance diagnostic accuracy. This is achieved through adaptive pooling and a feature weighting network, which can identify influential cellular characteristics. Finally, the model employs a decoder with a linear decoder and integration layer to map the processed features to diagnostic outcomes, utilizing ensemble learning to mitigate bias from limited data samples and improve robustness, ultimately achieving robust and accurate diagnostic predictions. An analysis of the time complexity of the BloodPatrol model can be found in Supplementary Information, Section II: "Computational Efficiency Analysis of BloodPatrol". The model diagram is shown in Fig. 1.

D. Encoder

Due to the large scale of cells collected by the flow cytometer and the variable number of cells, how to effectively extract key features from dynamic data is the focus of this study. Therefore, the Encoder based on the transformer architecture becomes a clear choice. The transformer Encoder can effectively adapt to sequences of indefinite length, thus it can efficiently encode real-time data from the flow cytometer. In addition, the all-pair attention mechanism of the transformer encoder can expand the model's perceptual field to the single-cell dimension, extracting more information from abnormal cells. This deep cell-level expression is crucial for subsequent modules when making sample-level predictions. Moreover, the diversity of cell types in flow cytometry data means that the multi-head attention mechanism can capture more layered features of cell types, thereby preventing the heterogeneity among cells from causing the failure of single-level analysis methods. In subsequent ablation experiments, we can also see that the transformer Encoder has a significant impact on the model's performance, suggesting that flow cytometry data processing models based on the transformer architecture have great potential to become the mainstream method in this field.

E. Intelligent Feature Weighting Fusion Module

In the diagnosis of leukemia, it is crucial to accurately identify and analyze abnormal cells, as they are key factors in determining the presence and classification of the disease [29]. Therefore, we designed the intelligent feature weighting fusion module based on this concept. We first utilize a dynamic max pooling layer to extract features processed by the transformer encoder. The dynamic pooling layer can adaptively adjust the size of the pooling window according to the number of single cells in each sample, selecting the most significant cell features and normalizing the size of the feature matrix for subsequent processing. This is formally represented as follows:

$$\text{Window Size} = \left\lceil \frac{n}{m} \right\rceil, \tag{1}$$

$$P_{ij} = \max\left(X_{(i-1)\cdot \text{Window Size}+1:i\cdot \text{Window Size},j}\right), \tag{2}$$

where n represents the number of cells in a sample, m denotes the output dimension, and X is the feature matrix encoded by the encoder.

Next, we employ a feature weighting net to further extract cellular feature information. This network learns the importance weights of each cell feature, enabling the model to distinguish features that are more influential for diagnosis. This weighting strategy enhances key information during feature fusion while suppressing noise or irrelevant data, ensuring that the model's

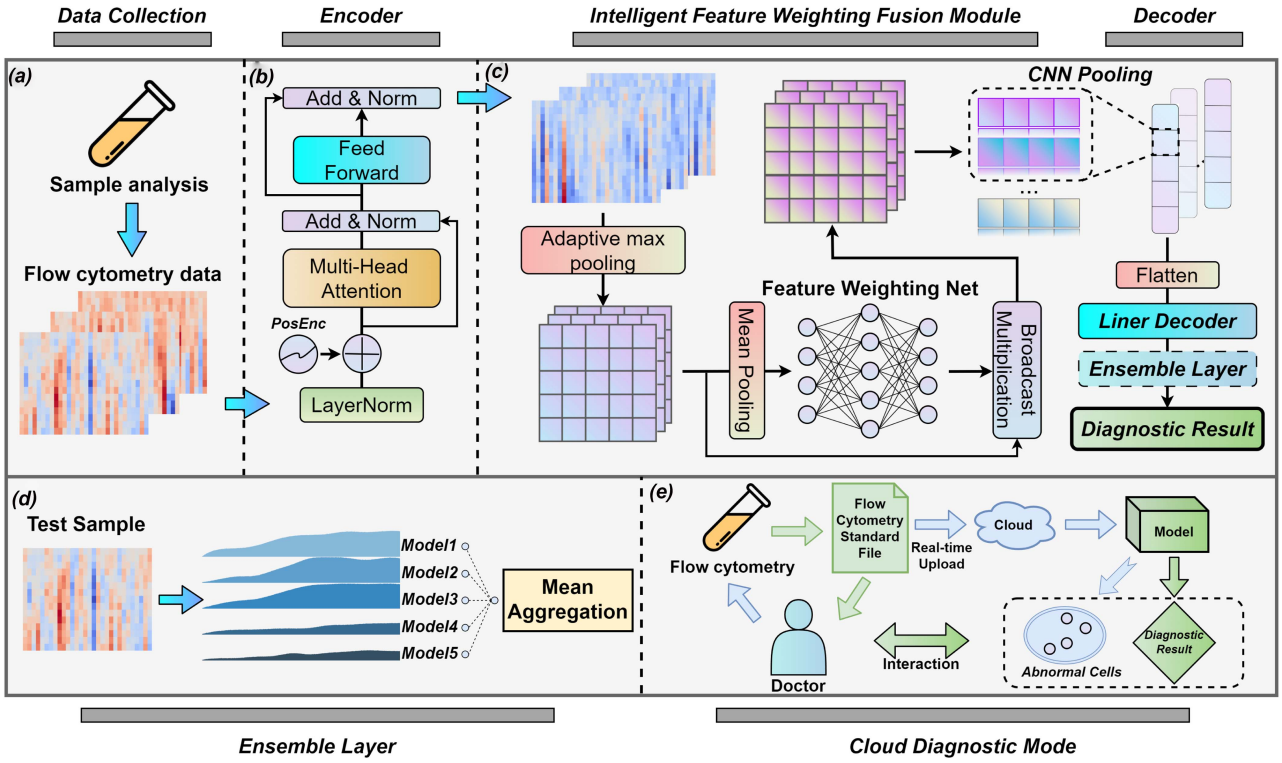

Fig. 1. The model diagram of BloodPatrol. (a) This represents the initial data collection phase where a biological sample is analyzed. (b) This part of the process involves the encoding of the flow cytometry data. (c) Here, the encoded features are further processed through an 'adaptive max pooling', which are techniques to reduce the dimensionality of the data while preserving important features. The output then passes through a 'feature weighting net' that assigns importance to different features before feeding into a 'CNN pooling' process. (d) The data from multiple models (model1 to model5) is aggregated using the 'mean aggregation' method, which is an ensemble learning approach where the predictions from multiple models are combined to improve accuracy. (e) This partoutlines a cloud-based diagnostic workflow where flow cytometry data is uploaded in real-time to the cloud, and a model processes the data to yield a diagnostic result. There is an interaction with a doctor, indicating that the system is designed to work in conjunction with medical professionals. The result includes the identification of abnormal cells, which is crucial for diagnostic purposes.

attention is focused on cell samples that best represent disease characteristics. Specifically, we first use mean pooling to obtain a sample-level representation, which serves as the basis for feature allocation in the feature weighting net. The model then applies broadcasting multiplication to the original features using the output weights, achieving weight allocation on cellular features. The formal definition of the feature weighting net is:

$$\begin{aligned} p &= \text{MeanPooling}(P), \\ W &= \text{Relu}\left(W_2 \cdot \text{Relu}\left(W_1 \cdot p\right)\right), \\ F &= W * P, \end{aligned} \tag{3}$$

where P denotes the feature matrix after dynamic pooling, W denotes the output vector of the feature weighting net, indicating the importance weights of the features, $*$ denotes broadcast multiplication, and F is the final output of this module.

Subsequently, the model applies residual connections [30] to integrate both weighted and unweighted feature information. This process is especially important when dealing with highly heterogeneous leukemia data, ensuring the model's generalization ability on unseen data and helping to avoid performance degradation due to rare data distributions.

At this stage, the feature matrix still has a large size, which, if used directly for decoding, could lead to model overfitting [31]. To prevent loss of information, we use a 1DCNN to learn spatial correlations in the dimension of weighted cell features, further enhancing the model's ability to recognize key information in feature dimensions. This is done while preserving information and outputting a simplified feature representation, avoiding model overfitting due to excessively high feature dimensions. The formal definition of the CNN pooling layer is:

$$\begin{aligned} F_p(l, c_{out}) = &\sum_{k=0}^{K-1} \sum_{c_{in}=0}^{C_{in}-1} W\left(k, c_{out}, c_{in}\right) \\ &\cdot F_{res}\left(l + k, c_{in}\right) + b(c_{out}), \end{aligned} \tag{4}$$

where F_{res} denotes the output of the feature weighting net combined with residual connections, F_p is the result after pooling over the feature dimensions (channel numbers), $W \in \mathbb{R}^{K \times C_{out} \times C_{in}}$ represents the convolutional kernel weights, $b \in \mathbb{R}^{C_{out}}$ is the bias and l is the position of the convolution operation on the input feature map.

The Intelligent feature weighting fusion module, through its meticulously designed feature extraction and weight learning mechanisms, enhances the sensitivity and accuracy of the leukemia diagnosis model, providing robust technical support for clinical diagnosis.

F. Decoder

The Decoder module, consisting of a Linear Decoder and an Ensemble Layer, works in tandem to map the multi-dimensional feature matrix to the final diagnostic results. The model initially flattens the high-dimensional feature matrix into one-dimensional features, which are then fed into the linear layer for decoding, preliminarily yielding the model's estimation of the diagnostic results. Additionally, in this study, the scarcity of blood cancer flow cytometry data poses a challenge, as the division of samples can introduce some bias into the model. To effectively mitigate this issue, we integrated the concept of ensemble learning. By merging the outputs of multiple models and employing an averaging aggregation strategy, we can effectively reduce errors caused by specific biases in models, thereby providing more robust predictions [32]. Furthermore, to ensure the diversity of ensemble learning, we made full use of multiple models from cross-validation in this study, ensuring that each model is trained on data with different distributions. This guarantees that each model learns from data from different perspectives, allowing for complementarity when combined and reducing the risk of overfitting. This ensemble strategy is straightforward yet effective, significantly enhancing the model's discriminative ability and reducing prediction variance, thus improving the model's robustness.

III. Experiment

A. Evaluation Strategies and Metrics

This study employs two datasets, CLL [23] and B-ALL [23], for experimentation. To assess the performance of the model, each dataset is divided into a training set and a test set, in an 8:2 ratio. For the stability and reliability of experimental results, we perform five-fold cross-validation on the training set. That is, the training set is equally divided into five subsets, with four subsets used for training and the remaining one for validation in each round.

The evaluation of the model utilizes multiple metrics to comprehensively reflect its performance. The primary evaluation metrics include: AUC (Area Under the Curve), AUPR (Area Under the Precision-Recall Curve), accuracy (acc), recall, precision (pre), and F1-score. AUC provides an overall assessment of the model's ability to distinguish between different categories, while AUPR focuses more on the performance assessment on positive samples. Accuracy, recall, precision, and F1-score measure the model's predictive accuracy from different perspectives. Detailed explanations of these metrics are as follows:

$$\begin{aligned}
\text{Accuracy} &= \frac{TP+TN}{TP+TN+FP+FN}, \\
\text{Recall (Sensitivity)} &= \frac{TP}{TP+FN}, \\
\text{Precision} &= \frac{TP}{TP+FP}, \\
\text{F1-score} &= 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}},
\end{aligned} \tag{5}$$

where TP and TN denote the number of positive and negative samples correctly predicted by the model, respectively; FP and FN denote the number of positive and negative samples incorrectly predicted by the model, respectively.

B. Experimental Setting

In our study, the Pytorch [33] framework was utilized for designing our model. For the B-ALL dataset, we employed a transformer encoder with 11 heads, while for the CLL dataset, a transformer encoder with 4 heads was used. The target dimension for the dynamic pooling layer was set at 100. The dimensions for the CNN pooling layer were defined as [original dimension, original dimension, 4]. For the linear decoding layer, the dimensions were arranged as [400, 256, 128, 64, 2]. Each layer was integrated with a Dropout rate of 0.2 to mitigate overfitting. The PolyLoss [28] function was employed with a first-order coefficient Λ set at 1. The Adam [34] optimizer was chosen for learning, with a weight decay set to 1e-4. The learning rate was set at 5e-5 for the B-ALL dataset and 1e-4 for the CLL dataset. Training was conducted over a maximum of 50 epochs.

C. Performance Comparison With Benchmarks

In our experiment, we compared the BloodPatrol model and its ensemble learning version, BloodPatrol-ES, with several traditional methods (including CSNN, DeepCNN, and CellCNN), which are used for the application of flow cytometry data in the classification of leukemia. Additionally, we included a Multilayer Perceptron (MLP) in our comparison as a standard deep learning method. Fig. 2 and Table I show the results of five-fold cross-validation. Across all datasets, BloodPatrol-ES achieved the best performance on all evaluation metrics. Most notably, the F1-score indicator showed that in the B-ALL dataset, BloodPatrol-ES achieved an improvement of 14% over other models, and in the CLL dataset, the improvement was over 27%. This indicates that the model proposed in this study has a significantly superior ability to predict positive samples compared to similar models, which is extremely important for the application in leukemia diagnosis. It is worth noting that the standard deviation of each indicator for the BloodPatrol series models is also significantly smaller than that of other models. Our model remains stable in its predictive results when using cross-validation, indicating that it has good generalization error and can learn effective leukemia diagnostic patterns from the training data, thus performing well on unseen data. The key driving factor for these advances is the innovative design of the model. The BloodPatrol model combines the transformer architecture with an intelligent feature weighting fusion module, which is crucial for understanding the complex patterns in flow cytometry data.

Furthermore, the ensemble strategy in BloodPatrol-ES also achieved good results, especially in the B-ALL dataset where it saw a significant performance improvement. We believe this is because in the B-ALL dataset, each sample has nearly 10 times the number of cells as in the CLL dataset, reaching the millions, and contains more types of abnormal cells. The diversity of training data brought by cross-validation in the ensemble

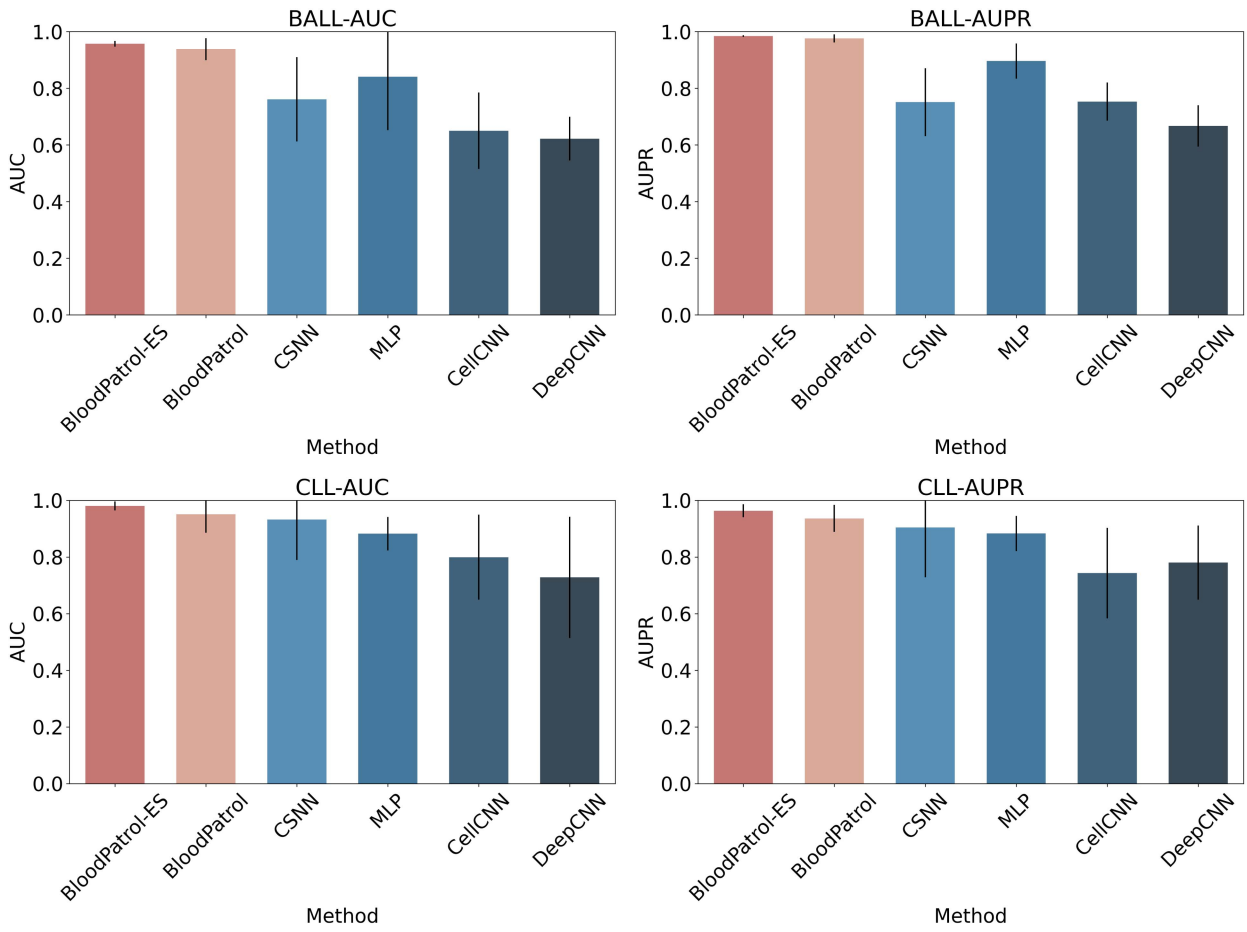

Fig. 2. Comparison of AUC and AUPR values across methods on the B-ALL and CLL datasets. Error bars in the chart represent standard deviations from the mean.

TABLE I
COMBINED PERFORMANCE METRICS FOR BALL AND CLL DATASETS

Dataset	Method	AUC	AUPR	Accuracy	Precision	Recall	F1-score
BALL	DeepCNN	0.799±0.150	0.743±0.160	0.743±0.192	0.693±0.251	0.685±0.215	0.628±0.161
	CellCNN	0.728±0.214	0.780±0.131	0.648±0.300	0.600±0.337	0.822±0.236	0.627±0.246
	MLP	0.882±0.059	0.883±0.062	0.811±0.029	0.884±0.080	0.628±0.138	0.719±0.068
	CSNN	0.932±0.143	0.904±0.176	0.900±0.063	0.900±0.300	0.650±0.220	0.754±0.252
	BloodPatrol	0.951±0.066	0.936±0.048	0.925±0.022	0.909±0.040	0.76±0.101	0.850±0.056
	BloodPatrol-ES	**0.980±0.016**	**0.963±0.023**	**0.942±0.023**	**0.932±0.048**	**0.866±0.094**	**0.891±0.045**
CLL	DeepCNN	0.622±0.077	0.667±0.073	0.588±0.062	0.651±0.090	0.506±0.108	0.559±0.073
	CellCNN	0.650±0.135	0.753±0.067	0.629±0.104	0.602±0.227	0.710±0.282	0.632±0.215
	MLP	0.841±0.189	0.896±0.062	0.838±0.112	0.869±0.128	0.914±0.044	0.882±0.061
	CSNN	0.761±0.149	0.751±0.120	0.580±0.064	0.581±0.343	0.366±0.295	0.384±0.227
	BloodPatrol	0.942±0.039	0.976±0.014	0.866±0.047	0.966±0.040	0.825±0.057	0.888±0.038
	BloodPatrol-ES	**0.957±0.010**	**0.984±0.003**	**0.885±0.008**	**1.000±0.000**	**0.837±0.012**	**0.911±0.007**

Values in bold represent the highest metric for each method within each dataset bloodPatrol-ES is a version of BloodPatrol that applies an ensemble learning layer.

strategy allows the model to have a broader range of abnormal cell detection. It is this strategy that has brought a significant improvement to the Recall indicator of the BloodPatrol model in the B-ALL dataset, fully demonstrating the advantages of the ensemble method in handling large-scale, diverse data.

The experimental results significantly prove the effectiveness of the BloodPatrol series models in providing reliable and consistent diagnostic results. This model not only significantly outperforms similar diagnostic models in various indicators but also provides clinicians with a powerful auxiliary tool, helping

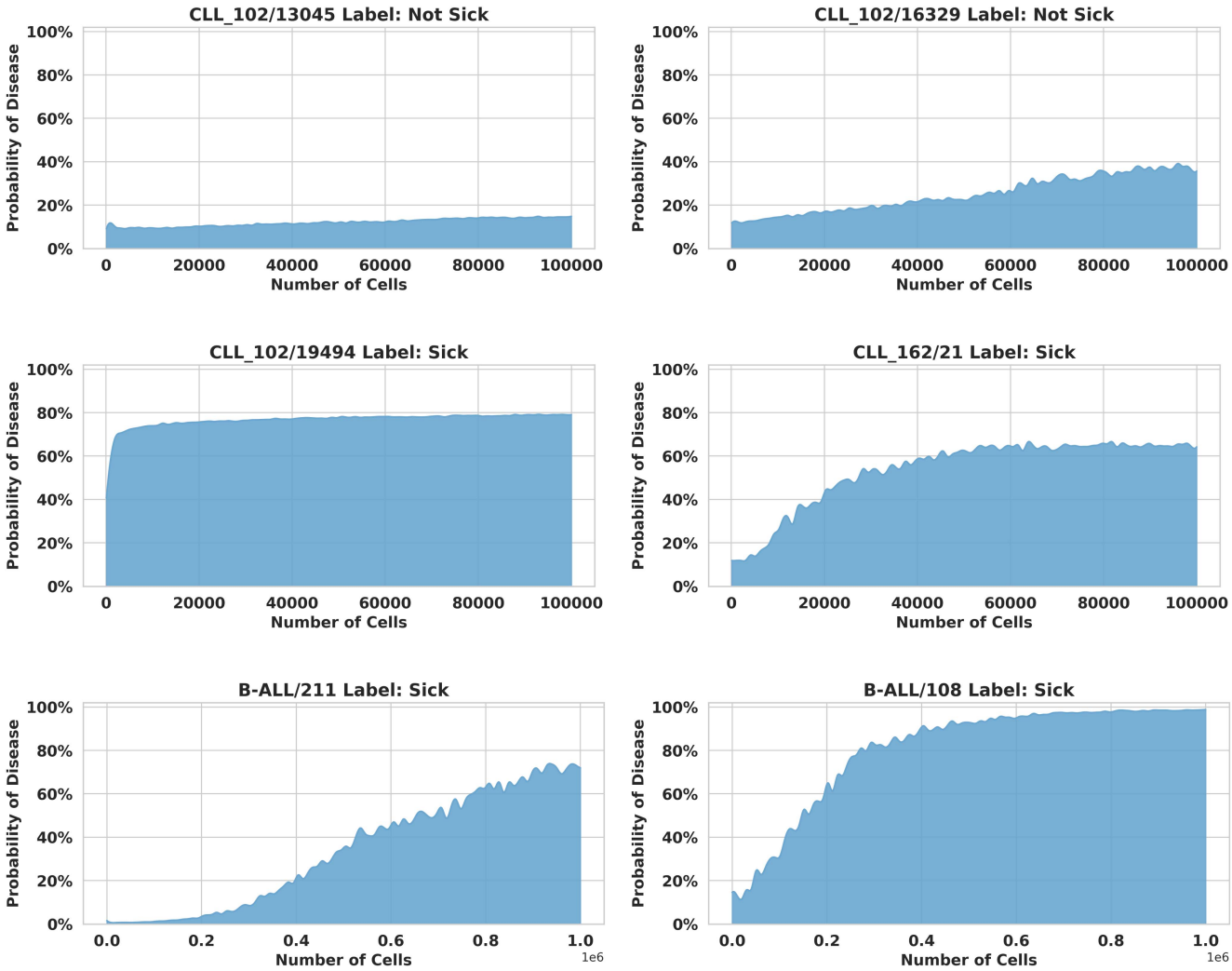

Fig. 3. Simulated dynamic monitoring experimental results of some cases.

them to deeply understand the nature of the disease and develop more accurate treatment plans.

D. Dynamic Monitoring Simulation Experiment

In this study, we presented the BloodPatrol model, which demonstrated the capability of dynamic monitoring and diagnosis of leukemia. This model can process an arbitrary number of cells in flow cytometry data without the need for padding or truncation and achieves accurate prediction of disease conditions. To validate the model's capability in dynamic diagnosis, we designed a series of simulated experiments. In these experiments, each unseen flow cytometry data set was divided into 100 subsets. By gradually increasing the number of subsets input into the model, we observed changes in the model's prediction results to assess its performance in dynamic monitoring.

Fig. 3 shows typical results of some samples in the simulated dynamic monitoring experiment. We found that the prediction probability trends for the vast majority of samples were very stable, and as the number of cells increased, the prediction results tended toward a clear bifurcation. This finding not only confirms the efficiency of BloodPatrol in capturing key features in flow cytometry data but also further validates its practical value in precise diagnosis of leukemia. Compared to traditional flow cytometry, a major advantage of BloodPatrol lies in its ability to process large data sets. Traditional methods may require significant time and resources when handling a large number of samples, whereas BloodPatrol can analyze these data more rapidly, enhancing diagnostic efficiency. Furthermore, with the aid of Internet of Things (IoT) technology, flow cytometers can be networked with cloud computing platforms, enabling real-time diagnosis during cell analysis. This not only significantly reduces the time required for diagnosis but also provides patients with faster disease diagnosis to a great extent, creating possibilities for timely treatment and better prognosis.

In addition, during the experimental process, we observed that many samples only required a small subset of the original data to achieve stable predictive results. Based on this discovery, we selected positive samples from the test sets of Chronic Lymphocytic Leukemia (CLL) and B-cell Acute Lymphoblastic Leukemia (B-ALL) for further study. To simulate the randomness of data collection by flow cytometers, we randomly shuffled

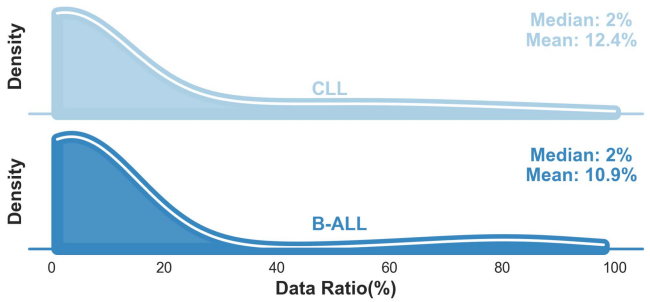

Fig. 4. Density distribution of minimum data requirement for accurate diagnosis by BloodPatrol.

the cell sequences of each sample and repeated this process 20 times to augment the experimental data. Subsequently, we divided each sample into 100 subsets and observed the proportion of original data required for the model to make accurate diagnoses.

The experimental results, as shown in Fig. 4, reveal that BloodPatrol can make correct diagnoses with just 2% of the original data for the majority of samples. These results not only demonstrate BloodPatrol's high sensitivity in capturing features of abnormal cell data but also indicate the potential of our model in terms of robustness to data and rapid diagnosis. Even with reduced data volumes, the model maintains a high level of diagnostic accuracy. This further proves the high efficiency of BloodPatrol in data processing and interpretation, and its modest requirements for data quality and completeness. This characteristic makes BloodPatrol particularly suitable for situations where data acquisition capabilities are limited or where rapid diagnostic responses are needed. This finding is of significant importance to medical practice as it implies that diagnoses can be made more quickly in clinical settings, while also reducing the need for extensive data processing.

E. Key Cell Group Analysis

To assist medical professionals in making more precise decisions during the diagnostic process, a method for locating abnormal cell groups is essential. However, due to the million-level volume of cell data produced by flow cytometry, traditional interpretability analysis methods like Shapely values struggle to balance computational speed and accuracy in this application scenario, posing significant challenges to computational hardware performance. Against this backdrop, we leveraged the model's ability to handle dynamic cell quantities and proposed a rapid interpretability analysis method. We named this method "Cell Population Impact Analysis" (CPIA). We divide all cells into n groups based on a set number of cell groups, n. Using the overall cell sample's predicted probability value as a benchmark, we observe the impact of reducing a particular cell group on the output probability. The results of the analysis are intuitively presented in the form of a heatmap Fig. 5, as shown in Figure. Moreover, CPIA allows for the free setting of start and end indices of the cells to be analyzed, thus adjusting the size of the cell groups. Repeating this process multiple times can pinpoint any single cell. Therefore, CPIA can rapidly identify key abnormal cell groups, aiding in more targeted examination and treatment.

TABLE II
CHANGES IN THE DISEASE PROBABILITY OF REMOVING KEY CELL GROUPS IN DIFFERENT CASES

Case	Standard	Top 3	Top 5
CLL_29860	0.843	0.749	0.736
CLL_330	0.814	0.723	0.634
CLL_149	0.787	0.686	0.644
B-ALL_134	0.865	0.820	0.791
B-ALL_189	0.903	0.892	0.881
B-ALL_62	0.972	0.969	0.964

To further verify the effectiveness of CPIA, we selected several positive samples from the test set for analysis. During the experiment, we first used CPIA to analyze key cell groups in these samples. We then ranked all cell groups based on their contribution to the pathogenic probability and manually removed the top 3 and top 5 contributing groups before predicting again. The experimental results, as shown in Table II, indicate that the prediction probabilities of all CLL samples significantly decreased after removing key cell groups, especially for the CLL_330 sample, which showed an 18% reduction in prediction probability after removing 5 key groups. For B-ALL samples, due to their cell count being approximately ten times that of CLL samples and correspondingly increased number of abnormal cells, a single CPIA's effectiveness is not as pronounced as in CLL samples. However, this is not a limitation of the CPIA method but rather due to overly large partition granularity failing to precisely identify all abnormal cell groups. In fact, when dealing with such large-scale samples, simply calling CPIA multiple times can refine the division granularity of cell groups, thereby improving the precision of the analysis.

The experimental results, as shown in Table II, demonstrate that CPIA, supported by a robust predictive model, can effectively analyze key cell groups and reveal interactions among them. Overall, CPIA, with its dynamic data handling capability, has successfully analyzed the importance of various cell groups, providing valuable auxiliary diagnostic information for medical professionals and making a significant contribution to the screening of abnormal cell groups. Our online diagnostic webpage also offers an interface for using this method.

F. Ablation Experiment

To evaluate the effectiveness of each component in our proposed BloodPatrol model, we conducted a series of ablation experiments. These experiments aimed to analyze the roles of the transformer encoder and feature weighting net in the model. We used Accuracy, Precision, Recall, F1-score, and AUC and AUPR as evaluation metrics, and compared the performance of the complete model with that of models with specific components removed. And "w/o TE" corresponds to the model variant with the transformer encoder module removed, while "w/o FWN" corresponds to the variant with the feature weighting net module removed. In addition, we validated the necessity of using the

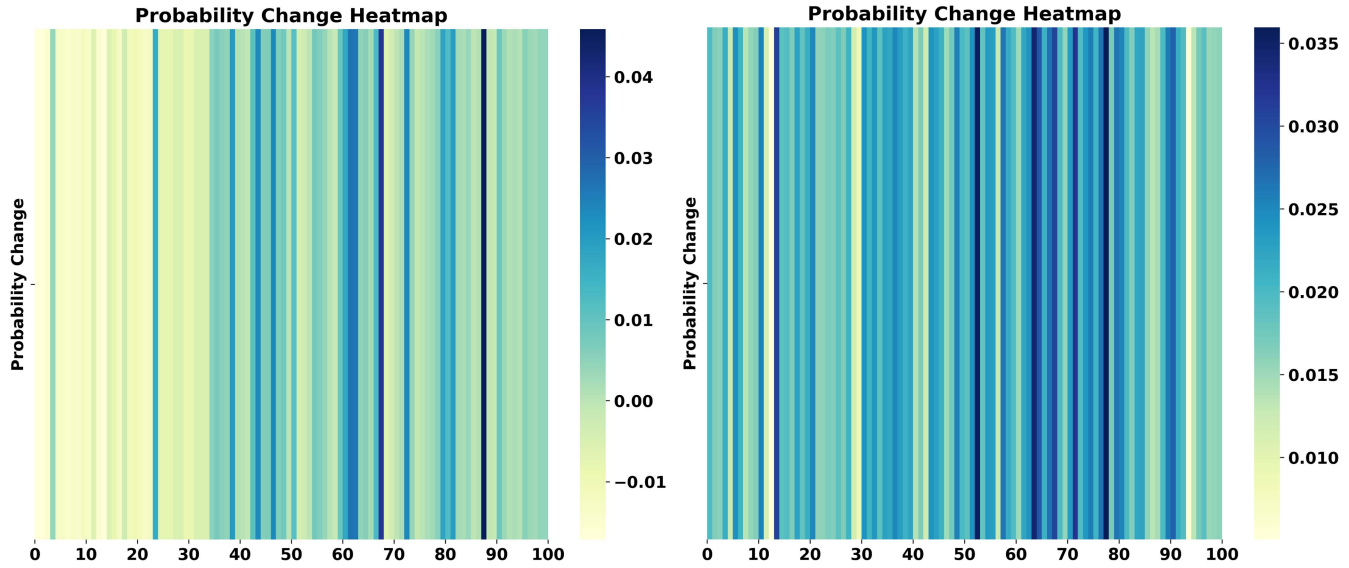

Fig. 5. Analysis of key cell group using CPIA.

TABLE III
RESULTS OF ABLATION EXPERIMENT

Dataset	Method	AUC	AUPR	Accuracy	Precision	Recall	F1-score
BALL	w/o FWN	0.948±0.057	0.931±0.041	0.908±0.027	0.883±0.095	0.760±0.120	0.816±0.049
	w/o TE	0.643±0.165	0.532±0.119	0.342±0.088	0.306±0.032	**1±0.000**	0.467±0.037
	Replaced TE with CNN	0.893±0.046	0.903±0.030	0.855±0.049	0.685±0.186	0.942±0.078	0.774±0.111
	BloodPatrol	**0.951±0.066**	**0.936±0.048**	**0.925±0.022**	**0.909±0.040**	0.760±0.101	**0.850±0.056**
CLL	w/o FWN	0.931±0.028	0.975±0.011	0.852±0.042	0.961±0.047	0.825±0.047	0.886±0.033
	w/o TE	0.749±0.013	0.897±0.009	0.677±0.017	0.786±0.034	0.745±0.033	0.764±0.011
	Replaced TE with CNN	0.894±0.034	0.942±0.020	0.820±0.037	0.847±0.067	**0.853±0.064**	0.847±0.030
	BloodPatrol	**0.942±0.039**	**0.976±0.014**	**0.866±0.047**	**0.966±0.040**	0.825±0.057	**0.888±0.038**

Transformer Encoder in contrast to a CNN-based encoder. The results are shown in Table III.

Data from the table shows that in both BALL and CLL types of blood cancer, the complete model demonstrated higher performance on most metrics. Removing the transformer encoder and feature weighting net both affected the model's performance in processing blood cancer (BALL and CLL types) data. The feature weighting net had a significant impact in the B-ALL dataset, especially for metrics related to positive samples. This indicates that even after pooling, the relative importance of cell samples remains crucial to the model. This module enhances the overall performance of the model by strengthening key features and suppressing noise. When the feature weighting net was removed, the model lost its ability to effectively weight important features, leading to a performance decline.

More notably, removing the transformer encoder resulted in a significant decline in performance on both datasets, especially on the key metrics of AUC and AUPR. This result emphasizes the importance of the transformer encoder in processing flow cytometry data. The advantage of the transformer Encoder lies in its ability to effectively process large-scale, variable data sequences, making it particularly suitable for data collected by flow cytometers. Its all-pair attention mechanism greatly expands the model's perceptual range, crucially extracting key features from dynamic data. Removing this module means the model loses its ability to learn interactions between cell features, leading to decreased performance in predicting blood cancer.

Furthermore, we conducted comparative experiments by replacing the Transformer encoder with a CNN encoder. The experimental results demonstrated a significant decline in model performance on both datasets. We attribute this primarily to the limited capacity of CNN to capture global features and long-range dependencies in cell samples. Although CNN are proficient at extracting local features through convolutional operations, they struggle to model the overall distribution and complex interactions among cells. In the context of leukemia diagnosis, both the overall level of abnormal cells and the presence of a few critical abnormal cells are of great importance. The Transformer's self-attention mechanism excels at weighing the importance of each cell within the entire sample context, thereby effectively capturing these crucial aspects. Detailed experimental results and hyperparameter settings for the CNN module can be found in Supplementary Information, Section I: CNN Encoder Hyperparameters Experiments.

IV. Conclusion and Discussion

This study proposes a novel dynamic monitoring diagnostic model for blood cancer, named BloodPatrol, which significantly outperforms existing models. The uniqueness of the BloodPatrol model lies in its ability to process any number of cells in flow cytometry data without the need for data padding or truncation. The model adopts an Encoder based on the transformer architecture and integrates an intelligent feature weighting and fusion module, effectively enhancing its capability to capture cell features. This has been validated through comprehensive experiments, proving the model's high application value in blood cancer diagnosis. The model can apply Internet of Things (IoT) technology, enabling real-time uploading of cell data through flow cytometers and rapid diagnosis through the computational power of cloud platforms. This significantly alleviates the demand for experimental analysis personnel in areas with scarce medical resources, thus offering greater possibilities for early diagnosis and favorable prognosis for patients. Moreover, our model has strong dynamic monitoring capabilities. Experiments have shown that the model can make robust diagnoses with just 2% of the standard number of cells in most samples. This is particularly suitable for scenarios with limited data acquisition capabilities and those requiring rapid diagnostic responses. It also demonstrates that the model is not stringent in data requirements and possesses strong robustness. We have also proposed a CPIA method for key cell group analysis, which, utilizing the model's dynamic data processing capabilities, quickly identifies key abnormal cell groups, providing valuable auxiliary diagnostic information for medical personnel and aiding in more precise examination and treatment strategy formulation.

Through innovative model design and algorithm implementation, this research has taken an important step forward in improving the accuracy and efficiency of blood cancer diagnosis. Future work could focus on further optimizing the model structure to handle more complex or heterogeneous data types. Additionally, further clinical trials and multicenter studies will provide deeper insights into the practicality and reliability of the model. This will not only furnish us with richer performance data but also test the model's universality in different medical settings and populations. Moreover, exploring the potential applications of the model in the diagnosis of other types of cancer will be an important step in expanding its application scope. Simultaneously, developing detailed and practical schemes to widely apply deep learning models in real diagnostic scenarios through IoT and cloud computing technologies is also one of the key directions of our future work.

V. Data Availability Statement

The source code of the model developed in this study, along with the key cell group analysis algorithm, has been made publicly available for use by the research community and professionals in related fields. These resources can be accessed through the following GitHub link: https://github.com/kkkayle/BloodPatrol.

Additionally, an online diagnostic platform has been established to provide auxiliary diagnostic services for patients with blood cancer and medical professionals. This platform offers not only basic diagnostic services but also performs key cell group analysis. The platform can currently be accessed at http://14.29.210.22:8001. If this address is inaccessible, please visit our GitHub repository at https://github.com/kkkayle/BloodPatrol to obtain the latest online diagnostic URL.

Acknowledgment

The authors acknowledge the assistance of AI-based language tools to enhance the manuscript's clarity and readability.

References

[1] M. Salimibeni, Z. Hajiakhondi-Meybodi, A. Mohammadi, and Y. Wang, "TB-ICT: A trustworthy blockchain-enabled system for indoor contact tracing in epidemic control," *IEEE Internet Things J.*, vol. 10, no. 7, pp. 5992–6017, Apr. 2023.

[2] L. Zhao, F. Zhang, H. Zhang, Y. Liang, A. Zhou, and H. Ma, "Robust respiratory rate monitoring using smartwatch photoplethysmography," *IEEE Internet Things J.*, vol. 10, no. 6, pp. 4830–4844, Mar. 2023.

[3] L. Gu, M. Mukherjee, M. Guo, J. Lloret, and R. Matam, "Low-cost assistive body temperature screening system to combat communicable infectious diseases leveraging edge computing and long-range and low-power wireless networks," *IEEE Internet Things J.*, vol. 10, no. 5, pp. 4174–4183, Mar. 2023.

[4] A. Onasanya and M. Elshakankiri, "Smart integrated IoT healthcare system for cancer care," *Wireless Netw.*, vol. 27, pp. 4297–4312, 2021.

[5] Y.-C. Wang, J. Xue, C. Wei, and C.-C. J. Kuo, "An overview on generative AI at scale with edge-cloud computing," *IEEE Open J. Commun. Soc.*, vol. 4, pp. 2952–2971, 2023.

[6] D. Villegas, A. Martínez, C. Quesada-López, and M. Jenkins, "IoT for cancer treatment: A mapping study," in *Proc. 15th IEEE Iberian Conf. Inf. Syst. Technol.*, 2020, pp. 1–6.

[7] M. Liu et al., "A deep learning method for breast cancer classification in the pathology images," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 10, pp. 5025–5032, Oct. 2022.

[8] A. U. Haq et al., "IIMFCBM: Intelligent integrated model for feature extraction and classification of brain tumors using MRI clinical imaging data in IoT-healthcare," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 10, pp. 5004–5012, Oct. 2022.

[9] D. A. d. Queiroz, C. A. d. Costa, E. A. I. F. d. Queiroz, E. F. d. Silveira, and R. d. R. Righi, "Internet of things in active cancer treatment: A systematic review," *J. Biomed. Inform.*, vol. 118, 2021, Art. no. 103814.

[10] E. E. Yörüker, S. Holdenrieder, and U. Gezer, "Blood-based biomarkers for diagnosis, prognosis and treatment of colorectal cancer," *Clinica Chimica Acta*, vol. 455, pp. 26–32, 2016.

[11] J. D. Schiffman, P. G. Fisher, and P. Gibbs, "Early detection of cancer: Past, present, and future," *Amer. Soc. Clin. Oncol. Educ. Book*, vol. 35, no. 1, pp. 57–65, 2015.

[12] M. I. D. Principe et al., "Applications and efficiency of flow cytometry for leukemia diagnostics," *Expert Rev. Mol. Diagnostics*, vol. 19, no. 12, pp. 1089–1097, 2019.

[13] P. F. Virgo and G. J. Gibbs, "Flow cytometry in clinical pathology," *Ann. Clin. Biochem.*, vol. 49, no. 1, pp. 17–28, 2012.

[14] S. Montante and R. R. Brinkman, "Flow cytometry data analysis: Recent tools and algorithms," *Int. J. Lab. Hematol.*, vol. 41, pp. 56–62, 2019.

[15] A. Rahim et al., "High throughput automated analysis of big flow cytometry data," *Methods*, vol. 134, pp. 164–176, 2018.

[16] K.-C. Chai, Y.-B. Zhang, and K.-C. Chang, "Regional disparity of medical resources and its effect on mortality rates in China," *Front. Public Health*, vol. 8, p. 8, 2020.

[17] T. Liu, E. Siegel, and D. Shen, "Deep learning and medical image analysis for COVID-19 diagnosis and prediction," *Annu. Rev. Biomed. Eng.*, vol. 24, pp. 179–201, 2022.

[18] J. Wei, Y. Zhu, L. Zhuo, Y. Liu, X. Fu, and F. Li, "Efficient deep model ensemble framework for drug-target interaction prediction," *J. Phys. Chem. Lett.*, vol. 15, no. 30, pp. 7681–7693, 2024.

[19] M. Kaneko et al., "The novel green learning artificial intelligence for prostate cancer imaging: A balanced alternative to deep learning and radiomics," *Urologic Clin.*, vol. 51, no. 1, pp. 1–13, 2024.

[20] M. Kaneko et al., "MP55-20 a novel machine learning framework for automated detection of prostate cancer lesions confirmed on MRI-informed target biopsy," *J. Urol.*, vol. 209, 2023, Art. no. e771.
[21] A. Kumar, A. Sharma, V. Bharti, A. K. Singh, S. K. Singh, and S. Saxena, "MobiHisNet: A lightweight CNN in mobile edge computing for histopathological image classification," *IEEE Internet Things J.*, vol. 8, no. 24, pp. 17778–17789, Dec. 2021.
[22] K. A. Tran, O. Kondrashova, A. Bradley, E. D. Williams, J. V. Pearson, and N. Waddell, "Deep learning in cancer diagnosis, prognosis and treatment selection," *Genome Med.*, vol. 13, no. 1, pp. 1–17, 2021.
[23] E. E. Robles et al., "A cell-level discriminative neural network model for diagnosis of blood cancers," *Bioinf.*, vol. 39, 2023, Art. no. btad585.
[24] Z. Hu, A. Tang, J. Singh, S. Bhattacharya, and A. J. Butte, "A robust and interpretable end-to-end deep learning model for cytometry data," *Proc. Nat. Acad. Sci.*, 2020, vol. 117, no. 35, pp. 21373–21380.
[25] E. Arvaniti and M. Claassen, "Sensitive detection of rare disease-associated cell subsets via representation learning," *Nat. Commun.*, vol. 8, no. 1, 2017, Art. no. 14825.
[26] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, Nov. 1998, vol. 86, no. 11, pp. 2278–2324.
[27] A. Vaswani et al., "Attention is all you need," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2017, vol. 30.
[28] Z. Leng et al., "Polyloss: A polynomial expansion perspective of classification loss functions," 2022, *arXiv:2204.12511*.
[29] A. Bouzid-Daho, M. Boughazi, C. Tanougast, and O. B. Medjahed, "Textural analysis of bio-images for aid in the detection of abnormal blood cells," *Int. J. Biomed. Eng. Technol.*, vol. 25, no. 1, pp. 1–13, 2017.
[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[31] R. Liu and D. F. Gillies, "Overfitting in linear feature extraction for classification of high-dimensional image data," *Pattern Recognit.*, vol. 53, pp. 73–86, 2016.
[32] A. Mohammed and R. Kora, "A comprehensive review on ensemble deep learning: Opportunities and challenges," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 35, pp. 757–774, 2023.
[33] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2019, vol. 32 .
[34] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.