



Developing a Hybrid Machine Learning Algorithm for Anemia Diagnosis

Safa S. Abdul-Jabbar^{1,*}

Alaa K. Farhan²

Abdul Hafeez Kandhro³

¹ Computer Science Department, College of Science for Women, University of Baghdad, Baghdad, Iraq

² Computer Science Department, University of Technology, Baghdad, Iraq

³ Institute of Medical Technology, Jinnah Sindh Medical University, Karachi, Pakistan

Emails: Safa.s@cs.w.uobaghdad.edu.iq • alaa.k.farhan@uotechnology.edu.iq • abdulhafeez.kandhro@jsmu.edu.pk

Received: November 02, 2024 Revised: December 26, 2024 Accepted: January 20, 2025 ★ Corresponding author

ABSTRACT

The utilization of artificial intelligence (AI) algorithms has significantly transformed the field of blood disease diagnosis, enabling enhanced capabilities in prediction, categorization, and optimization. However, there is still a lack of research exploring hybrid machine learning models that combine qualitative and quantitative datasets to address issues associated with blood diseases. We evaluate algorithmic combinations using datasets that include key characteristics from complete blood count (CBC) examinations. Prominent deep learning models, including CNN, RNN, and RCNN, are assessed. XGBoost is identified as the strongest machine learning algorithm and RCNN as the strongest deep learning model. We therefore propose RCNNX, a hybrid model integrating Robust Scaler, SelectKBest feature selection, RCNN, and XGBoost. RCNNX achieves testing accuracies of 100% and 95.12% on the Anemia Diagnosis Dataset and a second dataset, respectively, with recall rates of 100% and 94.64%. These findings indicate that feature selection reduces the number of input variables and the risk of overfitting, while XGBoost improves the predictive accuracy of RCNN.

Keywords: Complete Blood Cell Count ▪ CBC Test Parameters ▪ Machine Learning Algorithms ▪ Anemia Classifications ▪ Blood Diseases Prediction

1. INTRODUCTION

Anemia is a widespread condition characterized by a lower-than-normal hemoglobin concentration in the blood. It can affect overall health and well-being, causing fatigue, weakness, dizziness, shortness of breath, and an increased risk of infections [1]. Anemia is a major public health concern affecting a large portion of the global population. It can result from nutrient deficiencies, genetic disorders, chronic diseases, and certain medications. Early and accurate diagnosis and treatment are therefore essential to preventing complications and improving the quality of life of those affected [2].

In medicine, artificial intelligence (AI) focuses on designing

accurate algorithms and diagnostic methods. Diagnosis involves identifying the diseases or conditions responsible for a patient's signs and symptoms [3]. Despite advances, unmet needs remain in early disease detection, particularly when signals are complex and patients exhibit diverse underlying symptoms, challenging healthcare professionals developing diagnostic tools [4].

Creating an efficient and accurate AI model requires data pre-processing to enhance inputs and improve model precision. Feature scaling and selection are important for prediction efficiency [5]. Their combined use can enhance diagnosis [6, 7]. Integrating RCNN aims to combine the strengths of RNN and CNN with XGBoost; the latter can handle diverse datasets,

perform well across different challenges, and produce accurate predictions at low computational cost [8, 9].

2. LITERATURE REVIEWS

Anemia is a widespread condition characterized by a lower-than-normal hemoglobin concentration in the blood. It can cause fatigue, weakness, dizziness, shortness of breath, and an increased risk of infections. As a major public health concern, anemia can result from nutrient deficiencies, genetic disorders, chronic diseases, or medications. Early diagnosis and treatment are essential to preventing complications and improving quality of life [2].

AI research in medicine aims to develop accurate algorithms and diagnostic methods. Despite advances, early disease detection remains challenging when signals are complex and patients have diverse symptoms [10]. Preprocessing, feature scaling, and feature selection work together to improve model inputs and diagnostic performance. Combining RCNN with XGBoost is intended to leverage the strengths of recurrent and convolutional networks while maintaining efficient classification.

Table 1 summarizes selected studies in hematological diagnosis.

3. THEORETICAL BACKGROUND

This section introduces concepts that support the design of an AI model for disease detection and provide context for the research process.

3.1 Data Preprocessing

Feature scaling and selection are applied to preprocess data before it is fed into the proposed model.

3.1.1 Feature Scaling

Feature scaling adjusts data values to a defined range so that a model can perform more effectively. Depending on the data and the problem, common techniques include [15, 16]:

- **Min–Max scaling:** Scales features to a defined range, typically 0 to 1.
- **Robust scaling:** Uses the interquartile range, improving resistance to outliers.
- **Normalization:** Scales features to equal unit norms. It is useful when feature magnitude is irrelevant and only direction matters, as in some clustering algorithms.

3.1.2 Feature Selection

Feature selection identifies the features that contribute most to prediction and reduces the number of input variables. Common techniques include [13, 17]:

- **Filter methods:** Select features using statistical tests such as chi-squared, correlation, mutual information, SelectKBest, or variance threshold.
- **Wrapper methods:** Evaluate feature subsets by model performance using recursive feature elimination, forward selection, or backward elimination.

- **Embedded methods:** Integrate selection into model training, as in Lasso regression, decision trees, and random forests.
- **Dimensionality reduction:** Map high-dimensional data to a lower-dimensional space while retaining information, using methods such as PCA and t-SNE.

3.2 Machine Learning and Deep Learning Models

AI has the potential to improve disease diagnosis through more accurate and efficient methods [18]. Several pivotal models can be used to address this challenge.

3.2.1 Machine Learning Models

Machine learning (ML) analyzes samples and draws conclusions using mathematical and statistical methods. Random Forest (RF) and Multilayer Perceptron (MLP) can analyze large datasets, identify patterns, and generate predictions. Their use in anemia diagnosis has been studied [19, 20]. Ensemble methods such as XGBoost have also been applied to disease prediction [14].

3.2.2 Deep Learning Models

Deep learning (DL), a subfield of ML, uses neural-network architectures and is applied in healthcare, visual recognition, and text analytics [21, 12, 22]. Real-world data for DL modeling may be sequential, two-dimensional (such as images), or tabular. Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) are widely used; a Recurrent Convolutional Neural Network (RCNN) combines convolutional and recurrent layers, using convolutional features to model temporal relationships [11, 23].

Long Short-Term Memory (LSTM) units capture temporal dependencies using forget, input, and output gates. Optimization algorithms further support efficient training, including Stochastic Gradient Descent (SGD), Adam, and AdaDelta [24, 25, 26].

4. PROPOSED MODEL DESIGN

Machine learning algorithms that learn to assign instances to classes are central to classification in diagnostic and prediction systems. We first evaluated 11 algorithms: LogitBoost, Random Forest (RF), Decision Tree (DT), XGBoost (XGB), Gradient Boosting (GB), Multilayer Perceptron (MLP), AdaBoost, Logistic Regression (LR), Support Vector Machine (SVM), CatBoost, and K-Nearest Neighbor (KNN). These algorithms process datasets, detect patterns and correlations, and produce predictions.

4.1 Algorithm Selection and Refinement

Experiments with parameter values, feature scales, and selection criteria were used to refine the algorithms. LogitBoost, RF, XGBoost, LR, and MLP were identified as promising candidates. Robust Scaler was used for feature scaling and SelectKBest with the F-score for feature selection. Hyperparameters were tuned to improve performance.

4.2 Deep Learning Model Comparison

CNN, RNN, and RCNN were evaluated on the datasets described in Section 5.1. RCNN was selected as the candidate for the hybrid model based on its testing results.

Table 1. Summary of Literature on Hematological Diagnostics

Reference	Year	Methodology	Problem	Main Contribution	Accuracy	Limitations
Kandhro et al. [11]	2017	Random Forest and Decision Tree	Distinguishing thalassemia traits (TTs) from iron-deficiency anemia (IDA)	Improved a cutoff value and developed formulas to differentiate TTs from IDA.	100%	551 TTs and 441 IDA records.
Jaiswal et al. [12]	2019	Naive Bayes, Random Forest, and Decision Tree	Selecting an appropriate algorithm for CBC data	Proposed automated anemia diagnosis and reported 96.09% accuracy for Naive Bayes.	96.09%	200 records.
Alsheref [13]	2019	LogitBoost, Random Forest, Decision Tree, regression, KNN, Bayesian Network, MLP, Naive Bayes, and SVM	Selecting an algorithm for CBC data	Identified algorithms for blood-data prediction to support diagnosis.	98.16%	668 records.
Haider et al. [14]	2022	Artificial Neural Network	Diagnosing hematological emergencies	Reported training and testing results and used hidden trends in disease signatures.	83.1% training; 89.47% testing	Clinical usefulness limited.
Vohra et al. [8]	2022	Decision Tree, Logistic Regression, MLP, Naive Bayes, Random Forest, and SVM	Detecting anemia at different stages	Compared ML algorithms and filter, wrapper, and embedded feature-selection methods.	94.44%	400 records.

Table 2. Comparison of CNN, RNN, and RCNN Models

Property	CNN	RNN	RCNN
Type of data	Image data	Sequential data	Image and sequential data
Parameter sharing	Yes	Yes	Yes
Fixed-length input	Yes	No	No
Recurrent connections	No	Yes	Yes
Spatial relationship	Yes	No	Yes
Temporal relationship	No	Yes	Yes
Main advantages	Strong image-recognition performance	Predicts time series	Combines RNN and CNN strengths
Disadvantages	Requires substantial training data	Gradient vanishing/exploding	—
Applications	NLP, facial recognition	Text-to-speech	Image/video object detection

4.3 Proposed Hybrid Model: RCNNX

The proposed model combines RCNN, feature scaling, feature selection, and XGBoost. RCNN generates predicted class probabilities, which are used to train XGBoost to refine the predictions. This approach uses RCNN for feature extraction and XGBoost for classification.

4.4 RCNNX Algorithm

Input: Dataset in an Excel file.

Output: Training and testing loss/accuracy plots and saved models.

1. Load the data and split it into feature matrix **X** and target vector **y**.
2. Encode labels and handle missing values.
3. Split the data into training, testing, and validation sets.
4. Scale the data using Robust Scaler.
5. Select the top *K* features using SelectKBest with the F-score function, and remove unselected features.
6. Reshape the inputs for the RCNN model.
7. Build the RCNN with a bidirectional LSTM layer, convolutional layer, pooling layer, flatten layer, and Softmax output layer.
8. Compile the RCNN using Adam, categorical cross-entropy loss, and accuracy metrics.
9. Define callbacks for early stopping, model checkpointing, and learning-rate reduction.
10. Train and validate the RCNN model.
11. Extract and reshape the RCNN predicted probabilities into a two-dimensional matrix.
12. Define a custom XGBoost objective function using the gradient and Hessian.
13. Train and evaluate XGBoost using the RCNN probabilities, then save both trained models.
14. Plot the accuracy and loss results.

4.5 Proposed Model Architecture

The RCNNX architecture incorporates a bidirectional LSTM layer, convolutional and pooling layers, flattening, and output layers. Figure 1 shows how the design combines CNN and RNN technologies. XGBoost is used to handle high-dimensional data efficiently and produce accurate predictions at low computational cost.

5. EXPERIMENTAL RESULTS AND DISCUSSION

This section reports results from combining ML and DL methods and compares the performance of the evaluated models.

5.1 Dataset Description

Two CBC datasets were used to train and test the proposed model. Dataset2 is a private dataset containing encoded and unlinked clinical laboratory data from 2012–2015, obtained

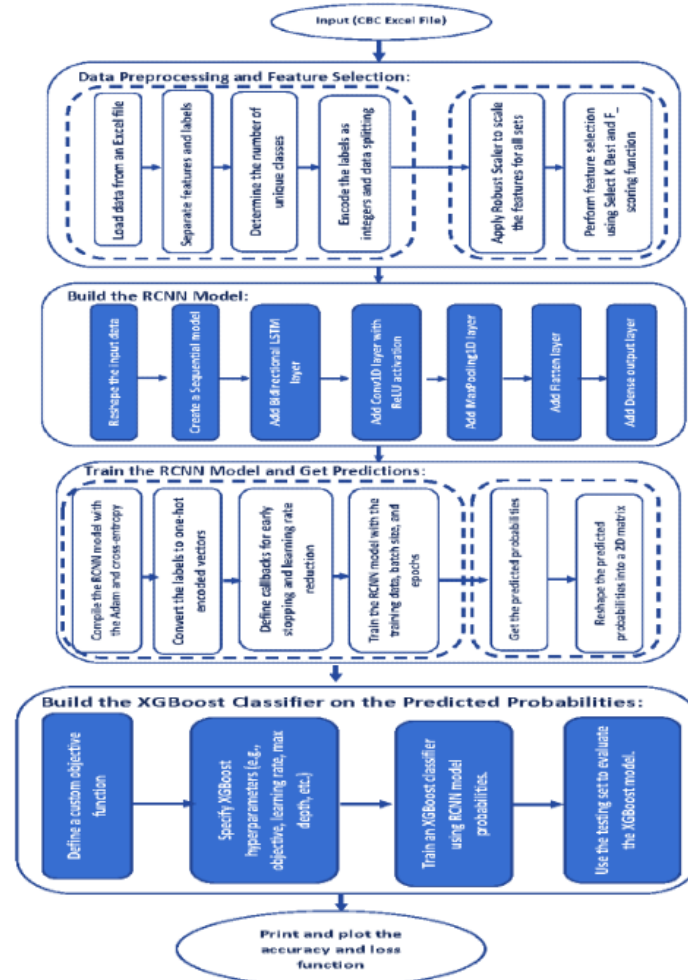


Figure 1. Diagram of the proposed RCNNX learning model.

from the Healthcare Molecular and Diagnostic Laboratory in Hyderabad, Pakistan. Table 3 summarizes the datasets.

Table 3. Description of the Tested Datasets

Dataset and author	Year, disease, country	Size and access
Anemia Dataset	Diagnosis 2021	1,442 samples
Shahane [3]	Anemia India	5 features
Dataset2	2017	410 samples
Kandhro et al. [11]	Anemia Pakistan	21 features
		Closed access

5.2 Machine Learning Comparative Results

The Anemia Diagnosis Dataset contains two classes (anemia and non-anemia) and 1,442 records. Figure 2 shows the training-phase accuracy metrics using 10-fold cross-validation; Table 4 reports test results for 25% of the dataset. Dataset2 contains 410 records and 21 features. Figure 3 presents training-phase accuracy metrics, and Table 5 shows the test results.

5.3 Deep Learning Comparative Results

The Sequential model was trained for 100 epochs. Preprocessing used Standard Scaler and SelectKBest; Adam was used with a learning rate of 0.001. The RNN and RCNN models used LSTM layers with 64 units. Table 6 summarizes the test accuracy.

Table 4. Accuracy Metrics for 25% of the Anemia Diagnosis Dataset (Testing)

Algorithm	Accuracy	Recall	Score	TNP	TRP	NPV	PPV
LogitBoost	1	1	66.66	1	1	1	1
RF	1	1	66.66	1	1	1	1
DT	1	1	66.66	1	1	1	1
NB	93.82	96.95	65.97	89.93	96.95	95.97	92.27
XGB	1	1	66.66	1	1	1	1
GB	1	1	66.66	1	1	1	1
CatBoost	99.43	99.04	66.45	1	99.04	98.65	1
SVM	99.71	1	66.66	99.33	1	1	99.51
KNN	97.19	99.00	66.44	94.83	99.00	98.65	96.13
LR	99.43	1	66.66	98.67	1	1	99.03
MLP	1	1	66.66	1	1	1	1
AdaBoost	1	1	66.66	1	1	1	1

Table 6. Test Results for Common Deep Learning Models

Model	Anemia Dataset (%)	Dataset2 (%)
CNN	99.32	53.48
RNN	100.00	55.81
RCNN	100.00	58.13

RCNN achieved the highest accuracy for both datasets.

5.4 Proposed Model Results

RCNNX was designed to achieve high accuracy with low execution time. The model explored Min–Max scaling, Robust scaling, and normalization, alongside filter, wrapper,

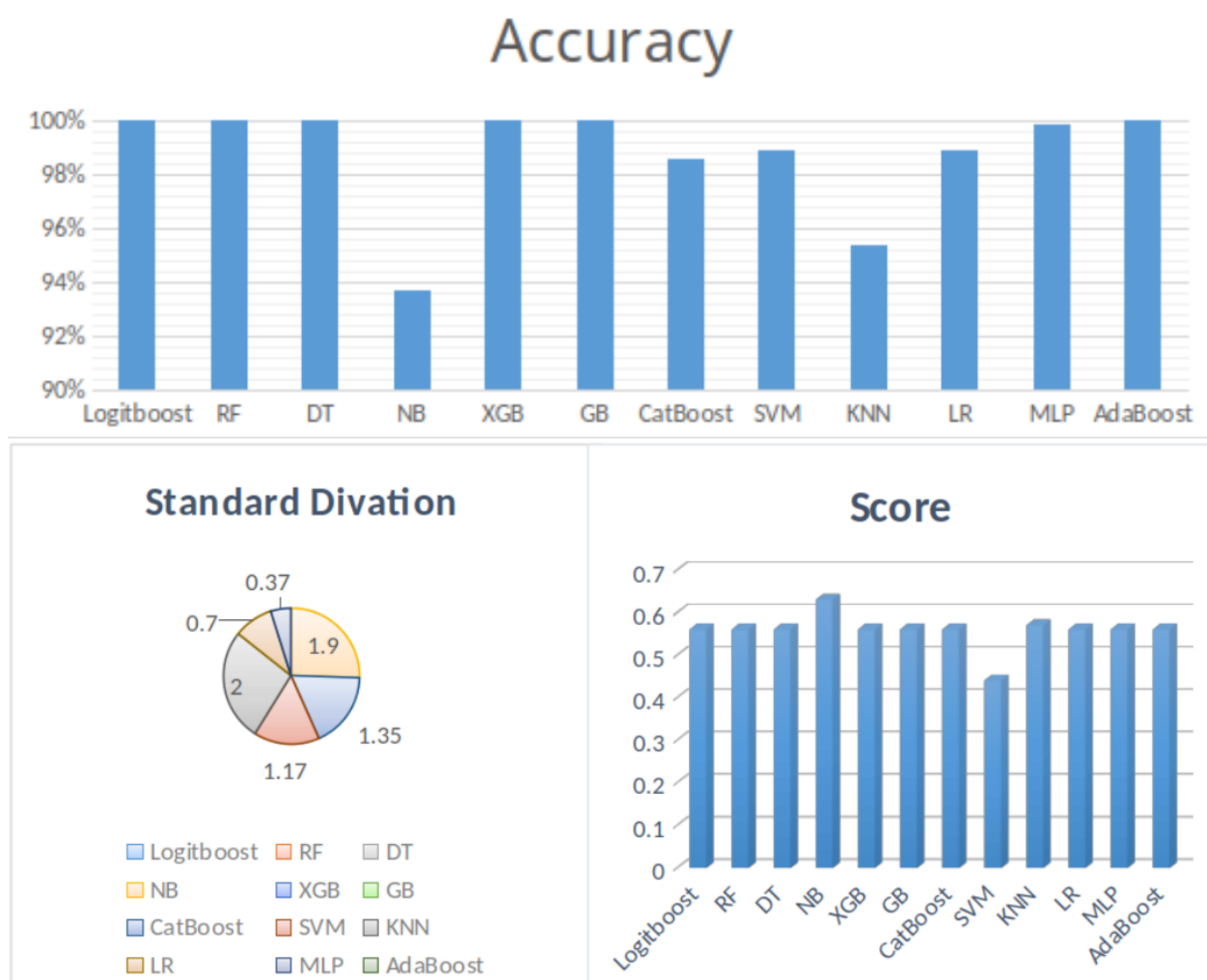


Figure 2. Training-phase accuracy metrics for the Anemia Diagnosis Dataset (10-fold cross-validation).

Table 5. Accuracy Metrics for 25% of Dataset2 (Testing)

Algorithm	Accuracy	Recall	Score	TNP	TRP	NPV	PPV
LogitBoost	52.04	52.71	51.32	51.03	52.71	41.67	61.91
RF	47.16	48.89	49.44	42.43	48.89	23.34	69.85
DT	49.51	52.45	51.20	45.23	52.45	39.58	58.18
XGB	54.36	58.57	53.94	45.45	58.57	34.09	69.49
NB	53.39	53.46	51.67	50.00	53.46	20.83	69.18
GB	51.45	53.84	51.85	47.36	53.84	37.50	63.63
CatBoost	45.63	49.33	49.66	35.71	49.33	20.83	67.27
SVM	48.54	51.21	50.60	38.09	51.21	16.66	76.36
KNN	46.60	50.00	50.00	38.70	50.00	25.00	65.45
LR	52.42	53.57	51.72	47.36	53.57	18.75	81.81
MLP	52.33	54.16	52.00	48.57	54.16	34.00	68.42
AdaBoost	52.42	54.28	52.05	48.48	54.28	33.33	69.09

dimensionality-reduction, and embedded feature-selection methods. Final results used 70% of data for training, 10% for validation, and 20% for testing. Table 7 summarizes the accuracy and recall.

Table 7. Results of the Proposed RCNNX Model

Dataset	Accuracy (%)	Recall (%)
Anemia Diagnosis Dataset	100.00	100.00
Dataset2	95.12	94.64

Table 8. Comparison of the Proposed Model with Prior Studies

Study, records, and main techniques	Accuracy (%)
Jaiswal et al. [12] (2019) 200 records; Naive Bayes, RF, DT	96.09
Alsheref [13] (2022) 1,577 records; ANN	89.47
Vohra et al. [8] (2022) 400 records; cross-validation with ML models	94.44
Proposed model (2023) 1,442 and 410 records; Robust Scaler, SelectKBest, RCNN, XGBoost	100.00; 95.12

5.5 Comparison with Prior Studies

Table 8 compares RCNNX with prior studies. The reported results show higher accuracy for the proposed model on the two evaluated datasets.

5.6 Visualization of Model Performance

Training and validation accuracy and loss were monitored for both datasets. Figure 4 shows the recorded curves.

6. CONCLUSIONS

Healthcare systems have advanced substantially, and analysis of pathological reports using deep learning (DL) and machine learning (ML) is a growing research area. This paper exam-

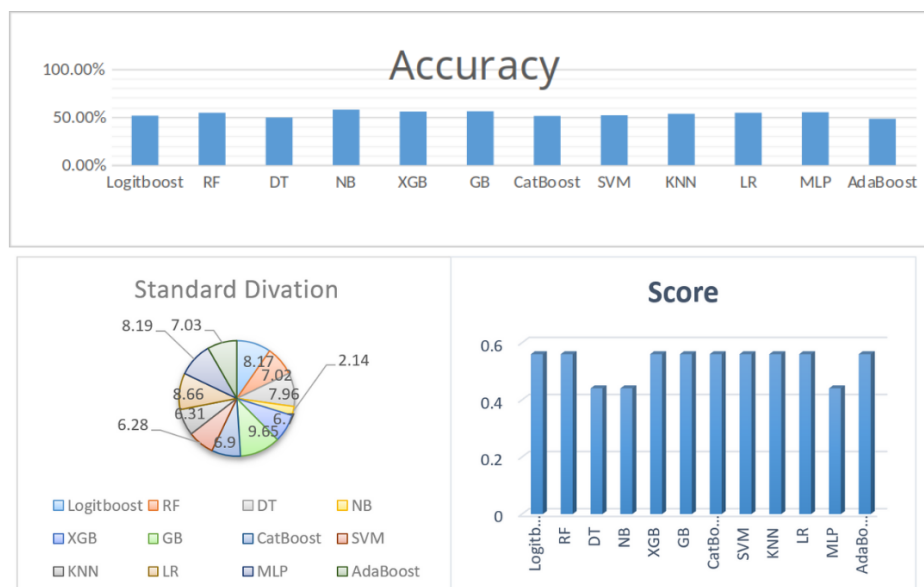


Figure 3. Training-phase accuracy metrics for Dataset2 (10-fold cross-validation).

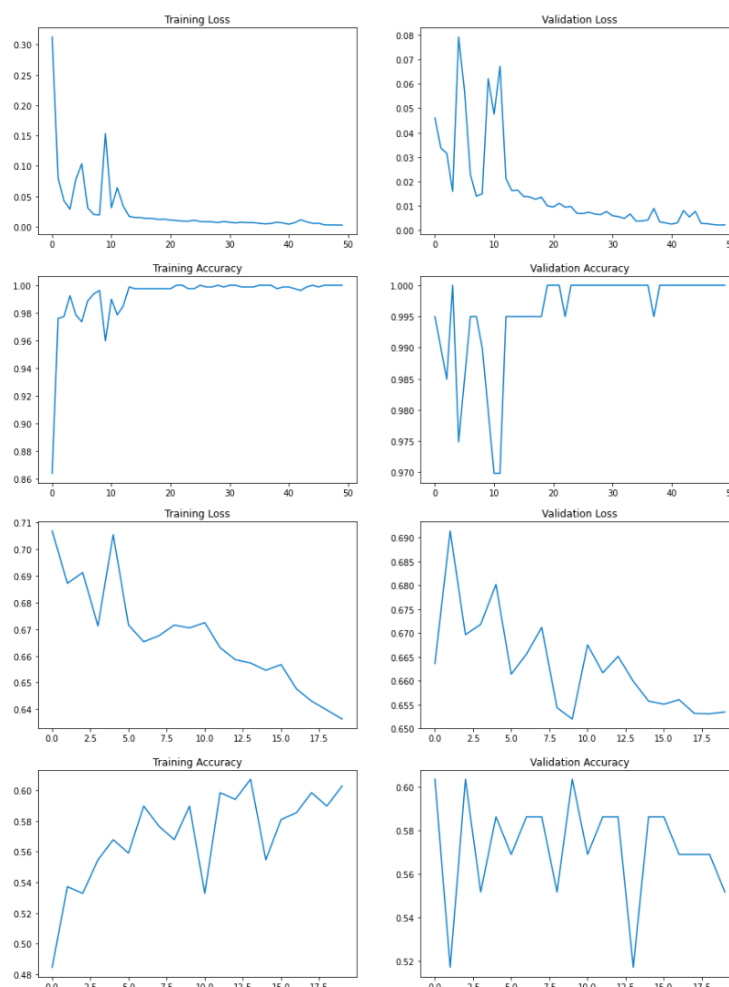


Figure 4. Training and validation loss and accuracy for both datasets.

ined DL and ML approaches for analyzing complete blood count (CBC) test data to support anemia diagnosis.

The developed hybrid model combines RCNN with XGBoost to balance accuracy and processing time. Random Forest was used to determine feature importance and inform feature-selection thresholds, while XGBoost was applied as an ensemble classifier. This approach was selected for its flexibility

and scalability and for reducing preprocessing requirements. RCNNX integrates Robust Scaler and SelectKBest for scaling and feature selection, RCNN for extracting patterns, and XGBoost for classification of high-dimensional data.

RCNNX achieved accuracies of 100% and 95.12% on the Anemia Diagnosis Dataset and Dataset2, respectively, as reported in Table 7. These results indicate that integrating

XGBoost with RCNN can help manage high-dimensional datasets and produce accurate predictions.

6.1 Future Work

Although RCNNX achieved strong results, several directions merit further study:

- **Dataset diversity:** Add larger and more diverse datasets to improve generalization across populations.
- **Feature engineering:** Explore advanced feature-engineering methods to improve classification accuracy.
- **Real-time applications:** Investigate deployment for real-time anemia diagnosis in clinical settings.
- **Explainability:** Improve interpretability to provide healthcare professionals with clearer insights into predictions.

This research highlights the potential of hybrid AI models for disease diagnosis by combining ML and DL techniques. RCNNX provides a framework for CBC data analysis and improved diagnostic support.

REFERENCES

- [1] I. H. Sarker, "Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions," *SN Computer Science*, vol. 2, no. 6, pp. 1–20, 2021.
- [2] D. C. E. Saputra, K. Sunat, and T. Ratnaningsih, "A new artificial intelligence approach using extreme learning machine as the potentially effective model to predict and analyze the diagnosis of anemia," *Healthcare*, vol. 11, no. 5, p. 697, 2023.
- [3] S. Shahane, "Anemia diagnosis dataset," Kaggle, 2020.
- [4] L. R. Al-Khazraji, A. R. Abbas, A. S. Jamil, and A. J. Hussain, "A hybrid artistic model using deep-dream model and multiple convolutional neural networks architectures," *IEEE Access*, 2023.
- [5] E. Stevens, L. Antiga, and T. Viehmann, "Deep learning with pytorch," 2020.
- [6] R.-Y. Sun, "Optimization for deep learning: An overview," *Journal of the Operations Research Society of China*, vol. 8, no. 2, pp. 249–294, 2020.
- [7] S. Suner, J. Rayner, I. U. Ozturan, G. Hogan, C. P. Meehan, A. B. Chambers, J. Baird, and G. D. Jay, "Prediction of anemia and estimation of hemoglobin concentration using a smartphone camera," *PLOS ONE*, vol. 16, p. e0253495, 2021.
- [8] R. Vohra, A. Hussain, A. Dudyala, J. Pahareeya, and W. Khan, "Multi-class classification algorithms for the diagnosis of anemia in an outpatient clinical setting," *PLOS ONE*, vol. 17, no. 7, pp. 1–18, 2022.
- [9] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "CNN-RNN: A unified framework for multi-label image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2285–2294, 2016.
- [10] L. R. Ali and A. S. Jamil, "Effect of changing targeted layers of the deep dream technique using VGG-16 model," *International Journal of Online and Biomedical Engineering*, vol. 19, no. 3, 2023.
- [11] A. H. Kandhro, W. S. V. Nuchnoi, and P. Pornlada, "New bioinformatics-based discrimination formulas for differentiation of thalassemia traits from iron deficiency anemia," *Laboratory Medicine*, vol. 48, no. 3, pp. 230–237, 2017.
- [12] M. Jaiswal, A. Srivastava, and T. J. Siddiqui, "Machine learning algorithms for anemia disease prediction: Select proceedings of IC3E 2018," in *IC3E 2018 Proceedings*. Springer, 2019.
- [13] F. K. Alsheref, "Blood diseases detection using classical machine learning algorithms," *Journal of Biomedical Science and Engineering*, vol. 10, no. 7, pp. 77–81, 2019.
- [14] R. Z. Haider, I. U. Ujjan, N. A. Khan, and E. Urrechaga, "Beyond the in-practice CBC: The research CBC parameters-driven machine learning predictive modeling for early differentiation among leukemias," *Diagnostics*, vol. 12, no. 1, p. 138, 2022.
- [15] M. Ahsan, M. A. P. Mahmud, P. K. Saha, K. D. Gupta, and Z. Siddique, "Effect of data scaling methods on machine learning algorithms and model performance," *Technologies*, vol. 9, pp. 5–9, 2021.
- [16] M. M. Ahsan, S. A. Luna, and Z. Siddique, "Machine-learning-based disease diagnosis: A comprehensive review," *Healthcare*, vol. 10, p. 541, 2022.
- [17] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of XGBoost," arXiv preprint, 2019.
- [18] N. Colakoglu and B. Akkaya, *Recent Advances in Data Science and Business Analytics*. Mimar Sinan Fine Arts University Publications, 2019.
- [19] R. Cui, H. Liu, and C. Zhang, "Recurrent convolutional neural networks for continuous sign language recognition by staged optimization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7361–7369, 2017.
- [20] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, no. 7, pp. 2121–2159, 2011.
- [21] A. A. Ibrahim, R. L. Ridwan, M. M. Muhammed, R. O. Abdulaziz, and G. A. Saheed, "Comparison of the cat-boost classifier with other machine learning methods," *Journal of Intelligent and Fuzzy Systems*, vol. 11, pp. 1–7, 2020.

-
- [22] A. Jović, K. Brkić, and N. Bogunović, “A review of feature selection methods with applications,” in *38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 1200–1205, 2015.
 - [23] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint, 2015.
 - [24] J. Li, H. Zhang, J. Zhao, X. Guo, W. Rihan, and G. Deng, “Embedded feature selection and machine learning methods for flash flood susceptibility mapping in the mainstream,” *Remote Sensing*, vol. 14, no. 21, p. 5523, 2022.
 - [25] K. Maharana, S. Mondal, and B. Nemade, “A review: Data pre-processing and data augmentation techniques,” *Global Transitions Proceedings*, pp. 91–99, 2022.
 - [26] P. O. Pinheiro and R. Com, “Recurrent convolutional neural networks for scene labeling,” in *International Conference on Machine Learning (PMLR)*, pp. 82–90, 2014.