Leukemia Diagnosis using Machine Learning Classifiers based on MRMR Feature Selection

Sipan M. Hameed
Computer Information Systems Department, Duhok Polytechnique University, Iraq
sipan.post@gmail.com

Walat A. Ahmed
Computer Information Systems Department, Duhok Polytechnique University, Iraq
walat.ali@dpu.edu.krd (corresponding author)

Masood A. Othman
Computer Information Systems Department, Duhok Polytechnique University, Iraq
masood.abdulrahman@dpu.edu.krd

Received: 2 May 2024 | Revised: 28 May 2024 | Accepted: 6 June 2024

 DOI: https://doi.org/10.48084/etasr.7720

ABSTRACT

Early and accurate diagnosis of leukemia is crucial for effective treatment. Machine Learning (ML) offers promising tools for leukemia diagnosis classification, but the required high-dimensional datasets pose challenges. This study explores the effectiveness of ML algorithms for leukemia disease classification and investigates the impact of feature selection with the Minimum Redundancy Maximum Relevance (MRMR) technique. MRMR was implemented to select informative features and evaluate four ML algorithms (Naïve Bayes (NB), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Artificial Neural Networks (ANNs)) using feature subsets with varying levels of relevance based on MRMR scores. Our results demonstrate that MRMR effectively reduced dimensionality while maintaining and even improving classification accuracy. KNN and SVM achieved the highest accuracy (100% for 67, 30, and 24 feature subsets), suggesting the benefit of focusing on highly relevant features. NB exhibited consistent accuracy across all feature sets.

Keywords-leukemia; machine learning; feature selection; KNN; SVM; ANN; Naïve Bayes

I. INTRUDUCTION

The incorporation of Machine Learning (ML) serves as a pivotal catalyst, altering the manner in which investigators tackle intricate issues and derive knowledge from extensive data collections. As the production of data continues to surge across a wide range of fields, including genomics, environmental science, astronomy, and finance, the demand for sturdy analytical instruments capable of distilling valuable insights has become increasingly critical [1]. ML, with its proficiency in identifying patterns, forecasting outcomes, and automating tasks, has surfaced as a crucial facilitator of scientific progression. By employing advanced algorithms and computational methodologies, researchers are now able to reveal concealed correlations, anticipate phenomena, and expedite the process of discovery in ways that were once beyond conception [2].

Feature selection is a crucial step in the ML workflow, focusing on selecting the most relevant and informative features from the dataset to improve model performance. It essentially involves choosing a subset of features that best contribute to the prediction task. Choosing the right feature selection technique depends on the specific dataset, the utilized ML algorithm, and the desired outcome. It is often beneficial to experiment with different techniques to find the one that yields the best results. Minimum Redundancy Maximum Relevance (MRMR) aims to strike a balance between these two aspects, measuring relevance and redundancy, and feature ranking and selection [4].

Leukemia research has greatly benefited from microarray technology, which allows scientists to analyze the expression levels of thousands of genes simultaneously. However, this vast amount of data presents a challenge, i.e. identifying the most informative genes that contribute to leukemia classification. Here is where MRMR comes into play. MRMR is a feature selection technique particularly suitable for high-dimensional datasets like gene expression profiles in leukemia [5, 6].

This paper use four widely used ML algorithms: Naïve Bayes (NB), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Artificial Neural Networks (ANNs) [3]. The present study utilizes a publicly available dataset originating from a foundational work published in 1999 [7].

This seminal investigation demonstrated the feasibility of classifying novel cancer cases through the analysis of gene expression profiles obtained via DNA microarray technology. This approach offered a generalizable framework for both the identification of new cancer subtypes and the classification of tumors within established categories. Notably, the referenced dataset was specifically employed for the classification of patients diagnosed with either Acute Myeloid Leukemia (AML) or Acute Lymphoblastic Leukemia (ALL) [8].

II. RELATED WORK

Authors in [9] investigated the utility of ML for immature leukocyte detection and classification in AML diagnosis. Their approach involved feature extraction, where 16 features were identified, including two novel features based on nuclear color. Subsequently, a Random Forest (RF) algorithm was trained to achieve leukocyte detection with an accuracy of 92.99% and classification with an accuracy of 93.45%. Notably, all classes achieved precision values exceeding 65%, demonstrating improvement over the existing methods. The nucleus-to-cytoplasm area ratio emerged as a significant feature for both detection and classification tasks. Additionally, the two novel nuclear color features contributed significantly to classification accuracy. These findings suggest the potential of the proposed model to aid in AML diagnosis. Furthermore, the identified salient features provide a valuable foundation for further research endeavors. Authors in [10] presented a novel approach for leukemic blood cell identification using morphological analysis of microscopic images. This method leverages image analysis, eliminating the need for physical blood samples, making it potentially suitable for low-cost and remote diagnostic systems. The system operates in a sequential manner: first, it differentiates leukocytes from other blood cells within the image. Then, it focuses on lymphocyte cells, commonly associated with leukemia. Following this, the system evaluates morphological features of these lymphocytes for leukemia classification. Notably, the segmentation process generates two distinct enhanced images for each cell – one for the cytoplasm and another for the nucleus. These images serve as the foundation for extracting characteristic features specific to different leukemia types, facilitating their identification. The KNN classifier was employed to classify ALL on a dataset of 72 samples derived from 66 correctly identified samples. The proposed system achieved an accuracy of 91.66%. Authors in [11]investigated the application of ML algorithms for leukemia classification using gene expression data. Their study explored three algorithms: XGBoost, RF, and ANN. Prior to analysis, the data underwent dimensionality reduction using Principal Component Analysis (PCA) to address potential issues of high dimensionality. The gene expression data, encompassed 72 individuals, each profiled for 7129 genes. The authors specifically focused on the Precision metric, which reflects the model's ability to correctly identify individuals with ALL. The RF model achieved a Precision value of 73.7%. It is important to note that, due to the nature of the metric, accuracy would be identical in this context. Authors in [12] presented an ML approach for classifying gene expression profiles in the context of leukemia diagnosis. Their study utilized a dataset comprising 72 bone marrow expression profiles. The objective of the classification task was to differentiate between ALL and AML. The authors employed a typical training-validation split, allocating 80% of the data for model training and 20% for validation. The proposed method achieved promising results, demonstrating high classification accuracy. Notably, the model achieved a classification accuracy of 98.2%. While this accuracy represents a positive outcome, it is important to acknowledge that advancements in this field have likely led to even higher accuracies in more recent studies. This research highlights the potential of computer-aided analysis using gene expression data for leukemia diagnosis. The insights gleaned from such approaches may prove valuable in future research endeavors undertaken by geneticists and virologists.

Authors in [13] developed a more accurate and efficient method to diagnose different leukemia subtypes. They employed gene expression data from the CuMiDa dataset for leukemia classification. They used linear programming, a computational modeling technique, to classify subtypes like AML, Bone Marrow, and PBSC CD34. The target was to reduce the high dimensionality (22283 features) by selecting the 25 most informative genes through feature selection. They achieved a high classification accuracy of 98%. Authors in [14] explored the use of a Deep Learning (DL) model, specifically a Convolutional Neural Network (CNN), for leukemia diagnosis. The model utilized an AML dataset ALL-ADB1 for training, containing a total of 100 photos. The CNN achieved high accuracy (over 98%), sensitivity (94.73%), and specificity (98.87%).

Authors in [15] developed a new classification model using Optimized Convolutional Neural Networks (OCNNs) for analyzing blood microscopic images. The model differentiates between leukemia-free and leukemia-affected images. Classification is utilizing OCNN to categorize images as "normal" or "abnormal" (leukemia). Fuzzy optimization was employed to fine-tune the OCNN hyperparameters for optimal performance. The model achieved a very high accuracy of 99.99%. Authors in [16] proposed an automated method for detecting ALL by classifying individual lymphocyte images extracted from peripheral blood smears. Their approach focuses on two key objectives: Cell Isolation and Leukemia Classification. The authors investigated the effectiveness of combining shape and histogram features for classification. They employed the KNN algorithm with various values of K (number of neighbors considered) ranging from 1 to 15. The model achieved the highest accuracy of 90% when using a combination of area, perimeter, mean, and standard deviation features with k = 7. Authors in [17] presented an algorithm for developing automated systems for detecting acute leukemia. The implemented method makes use of basic enhancement, morphology, filtering, and segmentation techniques to extract regions of interest using the k – means clustering algorithm. The proposed algorithm achieved a 92.8% accuracy when compared to the KNN and NB classifiers on a 60-sample dataset.

III. LEUKEMIA DISEASE

Leukemia is a cancer originating from bone marrow, producing various blood cells with distinct roles. It is categorized into chronic and acute types. It is a prevalent cancer type in children and among the top 15 in adults.

Characterized by abnormal cell growth and spread, known as metastasis, it is a significant global health concern [18, 28]. Predictions indicate 62,130 new leukemia cases and 245,000 severe or fatal cases in the U.S. alone. Its main types are AML, ALL, Chronic Myeloid Leukemia (CML), and Chronic Lymphoblastic Leukemia (CLL).

A. AML

A blood cancer with excess immature white blood cells. Symptoms include fever, bone pain, and fatigue. Treatment includes chemotherapy and stem cell transplants.

B. ALL

Affects lymphocytes in the bone marrow. Symptoms include bruising and bone ache. Treatment often involves chemotherapy and bone marrow transplantation.

C. CML

A white blood cell cancer due to a genetic defect. Symptoms include bone pain and fever. Treatment involves chemotherapy and bone marrow transplantation.

D. CLL

Starts in bone marrow's lymphocytes. Symptoms include enlarged lymph nodes and abdominal pain. Treatment is complex and involves a team of healthcare specialists.

IV. LEUKEMIA DATASET

The Leukemia dataset, is a collection of microarray datasets. This dataset encompasses 7130 attributes, of which 7129 are gene features and one is a class attribute, distributed across 72 instances (47 instances for the ALL class and 25 for the AML class). All attributes are numerical, with the exception of the final column, which denotes the class and includes two categories: ALL and AML [7].

V. METHODOLOGY

This section outlines the methodology employed for developing and evaluating ML models for leukemia disease classification using the leukemia disease dataset [7]. Figure 1 shows the flowchart diagram of the proposed model

A. Data Acquisition and Preprocessing

1) Feature Selection with MRMR

To address the high dimensionality and identify the most informative features, we employed the MRMR feature selection technique. MRMR is particularly well-suited for high-dimensional datasets as it selects a subset of features that are:

- Highly relevant to the target class (leukemia type).
- Minimally redundant with each other

The MRMR algorithm leverages Mutual Information (MI) to quantify both relevance and redundancy:

- MI between a feature (x) and the target class (y): This measures the dependence between the feature and the class label. A higher MI indicates stronger relevance.
- MI between two features (x and z): This captures the redundancy between features. A lower MI signifies less redundancy.

We implemented MRMR with all the 7129 features in the dataset. The algorithm ranked all features based on their MRMR scores. These scores represent a balance between a feature's relevance to the class and its redundancy with previously selected features. A larger score indicates a more important predictor, as it signifies a higher relevance to the target class while exhibiting minimal redundancy with other informative features. A score values is between 0 and 1. We utilized MATLAB, a widely used platform for scientific computing and ML. MATLAB provides functionalities suitable for working with high-dimensional data and allows for efficient implementation of ML algorithms.

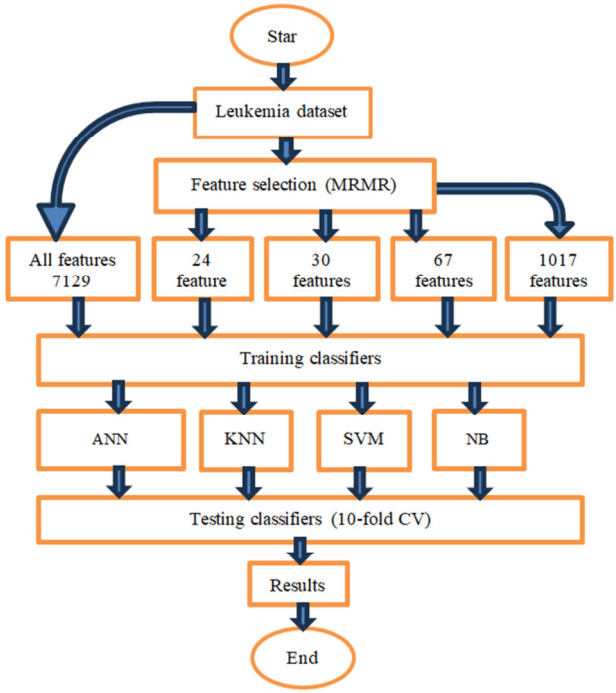

Fig. 1. Flowchart of the proposed leukemia diagnosis model.

2) Analyzing MRMR Scores and Feature Subsets

The score assigned to a feature represents its importance for classification. We analyzed the resulting MRMR scores and created multiple feature subsets based on different score thresholds:

- Features with all scores: This selection incorporates all 7129 features.
- Features with scores greater than 0.0: This selection incorporates features with some level of relevance to the target class based on the MRMR score (1017 features).
- Features with scores greater than 0.01 (67 features): This subset further restricts features to those with a higher degree of relevance.

- Features with scores greater than 0.05 (30 features): This selection represents features with even stronger relevance based on the MRMR scores.
- Features with scores greater than 0.08 (24 features): This subset focuses on the most relevant features identified by MRMR.

By creating these subsets, we can investigate the impact of feature selection stringency on model performance.

B. Model Selection and Training

The performance of four widely used ML algorithms for classification tasks was investigated:

- NB: This probabilistic classifier assumes independence between features, making it efficient for large datasets [20].
- KNN: This non-parametric method classifies data points based on the majority vote of their K nearest neighbors in the training data. The tuned hyperparameters were: K= 17, distance metric was Euclidean, and distance weight was squared inverse [21].
- SVM: SVMs aim to find a hyperplane that best separates the data points belonging to different classes. The tuned hyperparameters were: linear kernel function, automatic kernel scale, and one-vs-one multiclass coding [22].
- ANN: ANNs are powerful tools capable of learning complex relationships between features and the target variable. There were three fully connected layers of neurons, and the hyperparameter were tuned as follows: sigmoid activation function and hidden layer neurons size was 10 [23, 24].

Each model was trained and evaluated using the five feature subsets, the original one (7129) and the four generated from the MRMR scores (1017, 67, 30, and 24 features). This allows us to compare the impact of feature selection on the performance of each algorithm.

C. Model Evaluation

The performance of each model-feature set combination was evaluated using a standard classification metric, Accuracy which is the proportion of correctly classified samples [25].

1) Testing Method: 10-fold Cross-Validation

A robust evaluation approach, 10-fold cross-validation, was employed. This method randomly splits the data into 10 folds. The model is trained on 9 folds and tested on the remaining fold. This process is repeated 10 times, ensuring each fold is used for testing once. The final accuracy reported is the average of the results obtained across all 10 folds [26].

2) Experimental Results and Discussion

The application of MRMR feature selection and four ML algorithms for leukemia disease classification yielded interesting findings. We observed the impact of feature reduction on model performance and the effectiveness of different algorithms. A breakdown of the key observations follows.

- High accuracy across all models: All four models achieved high classification accuracy (above 90%) using all features (7129). This suggests the dataset inherently possesses strong discriminatory power between leukemia types. Table I shows all the accuracy results.
- Limited impact of feature reduction on the accuracy for NB: The accuracy remained relatively stable across all feature set sizes (97.2% - 98.6% for NB). This indicates that this algorithm might be less sensitive to feature redundancy and may perform well even with a broader set of features.
- Conversely, KNN and SVM exhibited a clear improvement in accuracy with increasing feature selection stringency. They achieved peak performance (100% accuracy) using the most selective feature sets (around 30-67-24 features). This suggests these algorithms benefit from focusing on a smaller set of highly relevant features, potentially leading to more robust classification models.

Overall, these findings highlight the potential benefits of feature selection with MRMR for specific ML algorithms in leukemia classification. It can improve model performance for KNN and SVM while potentially enhancing interpretability by focusing on the most informative features.

TABLE I. ACCURACY PERCENTAGE OF ALL CLASSIFIER

Feature set size	NB	KNN	SVM	ANN
All 7129 features	97.20%	98.60%	90.30%	95.80%
1017 features	97.20%	98.60%	98.60%	95.80%
67 features	98.60%	100%	100%	100%
30 features	98.60%	100%	100%	98.6%
24 features	98.60%	100%	100%	97.2%

A comparison with relevant studies is presented in Table II. Notably, researchers have utilized a wide range of techniques for both feature selection and classification. Additionally, the studies leveraged various datasets, each containing a differing number of leukemia samples.

TABLE II. COMPARISON WITH RELATED WORKS

Ref	Year	Dataset	Feature selection	Classifier	Result
[27]	2020	Clinically collected dataset	-	RF	92.99%
[13]	2023	CuMiDa	PCA	Linear programming	98%
[14]	2023	ALL-ADB1	-	CNN	98%
[15]	2024	C-NMC_Leukemia	-	OCNN	99.99%
[11]	2020	Microarray ALL	-	Classifier for separating	98.2%
[10]	2019	Microarray ALL	PCA	Xgboost and RF	92.3%, 80.8%
[12]	2019	TCIA	Lightgbm model	SVM, GBDT	79.4%, 85.6%
[16]	2018	Clinically collected dataset	KNN	KNN	92.8%
[9]	2017	Clinically collected dataset	-	KNN	91.66%
This study		Microarray ALL	MRMR (67 features)	NB, KNN, SVM, ANN	98.60%, 100%, 100%, 100%

VI. CONCLUSION

This study investigated the effectiveness of ML algorithms for leukemia disease classification using the dataset from [7]. Feature selection with the Minimum Redundancy Maximum Relevance (MRMR) technique was employed to address the dataset's high dimensionality and identify the most informative features.

Our findings demonstrate that:

- MRMR effectively reduced dimensionality. By selecting feature subsets based on MRMR scores, we achieved comparable or even improved classification accuracy compared to using all features. This highlights the value of MRMR in identifying the most relevant features for the classification task.
- Feature selection can improve model performance. For KNN, SVM, and ANNs, using feature subsets with higher MRMR score thresholds (indicating more relevant features) resulted in the highest accuracy (100% for KNN and SVM with 67, 30, and 24 features). This suggests that focusing on highly relevant features can enhance model performance.
- Naïve Bayes maintained consistent accuracy: The Naïve Bayes classifier exhibited consistent accuracy across all feature sets, meaning that it is less sensitive to feature selection compared to other algorithms.

REFERENCES

[1] M. Javaid, A. Haleem, R. Pratap Singh, R. Suman, and S. Rab, "Significance of machine learning in healthcare: Features, pillars and applications," *International Journal of Intelligent Networks*, vol. 3, pp. 58–73, Jan. 2022, https://doi.org/10.1016/j.ijin.2022.05.002.

[2] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, Jul. 2015, https://doi.org/10.1126/science.aaa8415.

[3] K. Theofilatos, S. Likothanassis, and A. Karathanasopoulos, "Modeling and Trading the EUR/USD Exchange Rate Using Machine Learning Techniques," *Engineering, Technology & Applied Science Research*, vol. 2, no. 5, pp. 269–272, Oct. 2012, https://doi.org/10.48084/etasr.200.

[4] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of Bioinformatics and Computational Biology*, vol. 3, no. 2, pp. 185–205, Apr. 2005, https://doi.org/10.1142/S0219720005001004.

[5] T. Haferlach *et al.*, "Clinical Utility of Microarray-Based Gene Expression Profiling in the Diagnosis and Subclassification of Leukemia: Report From the International Microarray Innovations in Leukemia Study Group," *Journal of Clinical Oncology*, vol. 28, no. 15, pp. 2529–2537, May 2010, https://doi.org/10.1200/JCO.2009.23.4732.

[6] V. A. Rajendran and S. Shanmugam, "Automated Skin Cancer Detection and Classification using Cat Swarm Optimization with a Deep Learning Model," *Engineering, Technology & Applied Science Research*, vol. 14, no. 1, pp. 12734–12739, Feb. 2024, https://doi.org/10.48084/etasr.6681.

[7] T. R. Golub *et al.*, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, Oct. 1999, https://doi.org/10.1126/science.286.5439.531.

[8] Y. L. Ng, X. Jiang, Y. Zhang, S. B. Shin, and R. Ning, "Automated Activity Recognition with Gait Positions Using Machine Learning Algorithms," *Engineering, Technology & Applied Science Research*, vol. 9, no. 4, pp. 4554–4560, Aug. 2019, https://doi.org/10.48084/etasr.2952.

[9] S. Dasariraju, M. Huo, and S. McCalla, "Detection and Classification of Immature Leukocytes for Diagnosis of Acute Myeloid Leukemia Using Random Forest Algorithm," *Bioengineering*, vol. 7, no. 4, Dec. 2020, Art. no. 120, https://doi.org/10.3390/bioengineering7040120.

[10] P. M. Gumble and S. V. Rode, "Analysis & Classification of Acute Lymphoblastic Leukemia using KNN Algorithm," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 5, no. 2, pp. 94–98, 2017.

[11] U. K. Dey and Md. S. Islam, "Genetic Expression Analysis To Detect Type Of Leukemia Using Machine Learning," in *1st International Conference on Advances in Science, Engineering and Robotics Technology*, Dhaka, Bangladesh, Dec. 2019, pp. 1–6, https://doi.org/10.1109/ICASERT.2019.8934628.

[12] P. K. Mallick, S. K. Mohapatra, G.-S. Chae, and M. N. Mohanty, "Convergent learning–based model for leukemia classification from gene expression," *Personal and Ubiquitous Computing*, vol. 27, no. 3, pp. 1103–1110, Jun. 2023, https://doi.org/10.1007/s00779-020-01467-3.

[13] M. Ilyas, K. M. Aamir, S. Manzoor, and M. Deriche, "Linear programming based computational technique for leukemia classification using gene expression profile," *PLOS ONE*, vol. 18, no. 10, Sep. 2023, Art. no. e0292172, https://doi.org/10.1371/journal.pone.0292172.

[14] K. A. Kadhim, F. H. Najjar, A. A. Waad, I. H. Al-Kharsan, Z. N. Khudhair, and A. A. Salim, "Leukemia Classification using a Convolutional Neural Network of AML Images," *Malaysian Journal of Fundamental and Applied Sciences*, vol. 19, no. 3, pp. 306–312, May 2023, https://doi.org/10.11113/mjfas.v19n3.2901.

[15] F. M. Talaat and S. A. Gamel, "Machine learning in detection and classification of leukemia using C-NMC_Leukemia," *Multimedia Tools and Applications*, vol. 83, no. 3, pp. 8063–8076, Jan. 2024, https://doi.org/10.1007/s11042-023-15923-8.

[16] E. Purwanti and E. Calista, "Detection of acute lymphocyte leukemia using k-nearest neighbor algorithm based on shape and histogram features," *Journal of Physics: Conference Series*, vol. 853, no. 1, Feb. 2017, Art. no. 012011, https://doi.org/10.1088/1742-6596/853/1/012011.

[17] S. Kumar, S. Mishra, P. Asthana, and Pragya, "Automated Detection of Acute Leukemia Using K-mean Clustering Algorithm," in *Advances in Computer and Computational Sciences*, S. K. Bhatia, K. K. Mishra, S. Tiwari, and V. K. Singh, Eds. New York, NY, USA: Springer, 2018, pp. 655–670.

[18] V. R. Minciacchi, R. Kumar, and D. S. Krause, "Chronic Myeloid Leukemia: A Model Disease of the Past, Present and Future," *Cells*, vol. 10, no. 1, Jan. 2021, Art. no. 117, https://doi.org/10.3390/cells10010117.

[19] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005, https://doi.org/10.1109/TPAMI.2005.159.

[20] I. Rish, "An Empirical Study of the Naive Bayes Classifier," IBM, IBM Research Report RC 22230 (W0111-014), Nov. 2001.

[21] P. Cunningham and S. J. Delany, "k-Nearest Neighbour Classifiers - A Tutorial," *ACM Computing Surveys*, vol. 54, no. 6, Apr. 2021, Art. no. 128, https://doi.org/10.1145/3459665.

[22] S. Huang, N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang, and W. Xu, "Applications of Support Vector Machine (SVM) Learning in Cancer Genomics," *Cancer Genomics & Proteomics*, vol. 15, no. 1, pp. 41–51, Jan. 2018, https://doi.org/10.21873/cgp.20063.

[23] S. Shanmuganathan, "Artificial Neural Network Modelling: An Introduction," in *Artificial Neural Network Modelling*, S. Shanmuganathan and S. Samarasinghe, Eds. New York, NY, USA: Springer, 2016, pp. 1–14.

[24] W. Samek and K.-R. Muller, "Towards Explainable Artificial Intelligence," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Muller, Eds. New York, NY, USA: Springer, 2019, pp. 5–22.

[25] L. Zhang *et al.*, "A review of machine learning in building load prediction," *Applied Energy*, vol. 285, Mar. 2021, Art. no. 116452, https://doi.org/10.1016/j.apenergy.2021.116452.

[26] L. A. Yates, Z. Aandahl, S. A. Richards, and B. W. Brook, "Cross validation for model selection: A review with examples from ecology,"

Ecological Monographs, vol. 93, no. 1, 2023, Art. no. e1557, https://doi.org/10.1002/ecm.1557.

[27] S. Aljawarneh, M. B. Yassein, and M. Aljundi, "An enhanced J48 classification algorithm for the anomaly intrusion detection systems," *Cluster Computing*, vol. 22, no. 5, pp. 10549–10565, Sep. 2019, https://doi.org/10.1007/s10586-017-1109-8.

[28] N. Bibi, M. Sikandar, I. Ud Din, A. Almogren, and S. Ali, "IoMT-Based Automated Detection and Classification of Leukemia Using Deep Learning," *Journal of Healthcare Engineering*, vol. 2020, no. 1, 2020, Art. no. 6648574, https://doi.org/10.1155/2020/6648574.