IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 21, 2022, accepted April 18, 2022, date of publication May 18, 2022, date of current version June 3, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3176274

Symptom Based Explainable Artificial Intelligence Model for Leukemia Detection

MOHAMMAD AKTER HOSSAIN[1], A. K. M. MUZAHIDUL ISLAM[1], (Senior Member, IEEE), SALEKUL ISLAM[1], (Senior Member, IEEE), SWAKKHAR SHATABDA[1], AND ASHIR AHMED[2], (Member, IEEE)

[1]Department of Computer Science and Engineering, United International University, Dhaka 1212, Bangladesh
[2]Department of Advanced Information Technology, Kyushu University, Fukuoka 819-0395, Japan

Corresponding author: A. K. M. Muzahidul Islam (muzahid@cse.uiu.ac.bd)

This work was supported by the Institute of Advanced Research (IAR), United International University (UIU), Dhaka, Bangladesh, under Research Project UIU/IAR/02/2019-20/SE/07.

ABSTRACT Leukemia is not only fatal in nature, it is also extremely expensive to treat. However, leukemia detection at early stage can save lives and money of the affected people, specially children among whom leukemia as a cancer type is very common. In this paper, we propose an explainable supervised machine learning model that accurately predicts the likelihood of early-stage leukemia based on symptoms only. The proposed model is developed based on primary data collected from two major hospitals in Bangladesh. Sixteen features of the datasets are collected through a survey on leukemia and non-leukemia patients in consultation with a specialist physician. Our explainable supervised model is based on a decision tree classifier which provides significantly better results compared to other algorithms and generates explainable rules that are ready to use. We have employed Apriori algorithm for generating explainable rules for leukemia prediction. In addition, feature analysis and feature selection are performed on the dataset to show the strength of individual features and enhance the performance of the classification models. Several classifiers are experimented on the dataset to show how the proposed model that is simple yet explainable, performs significantly better compared to most other models that we have used. The decision tree model proposed in our experiments has achieved 97.45% of accuracy, 0.63 of Mathew's Correlation Coefficient (MCC) and 0.783 of area under Receiver Operating Characteristic (ROC) curve on the test set. We have also made the dataset and the source code of the methods used in this work available for future use by the researchers.

INDEX TERMS Explainable AI, leukemia, machine learning, symptom-based detection.

I. INTRODUCTION

The National Cancer Institute, USA estimated that only in USA, approximately 60,300 new patients have admitted to the hospital due to leukemia in 2019, where 24,370 of them have died [1]. This has been a major concern in recent years. In spite of major research endeavours taken to tackle leukemia and its different variants worldwide, the death rate from leukemia is alarming having severe consequences specially in children [2].

Leukemia cells are blood cells that are under-developed and shows abnormal behavior of growing and dividing in an indomitable manner. It is the most common type of cancer that is prevalent in children. Based on the cell types leukemia is broadly categorized into two types: lymphocytic (or lymphoblastic) and non-lymphocytic (or myeloid). They can occur in chronic or acute form. A body with leukemia cells develops signs and symptoms. However, often it is diagnosed at a later stage, thus makes the treatment more difficult. An early screening of leukemia can make a great difference by reducing the cost and related fatality rate and also improving the quality of life among the patients.

In general, leukemia detection and screening is being executed in the hospitals using various sophisticated methods. They employ blood samples [3], complete blood counts [4]–[7], bone marrow based tests [8], [9], etc. Bone marrow is the source or starting point of leukemia where lymphocytic or myeloid cells start to develop [10]. Imaging of blood cells too help to detect leukemia as shown in different researches [11]–[13]. A very popular dataset of leukemia detection is Acute Lymphoblastic Leukemia Image Database (ALL-IDB) for image processing, which is extensively

The associate editor coordinating the review of this manuscript and approving it for publication was Antonio J. R. Neves.

discussed in literature. These image based methods are often susceptible to sophisticated devices and imaging techniques deployed [14]–[16]. In recent times, genomics methods [17], [18] along with clinical data are in use [19]–[22]. Though the combinations of various methods and multi-modal data are effective, they might be available at a late stage [23]–[25].

To detect and predict leukemia, various machine learning based methods and algorithms have already been used in literature. With the increase in available data, it is now possible to formulate the problem as a supervised machine learning problem where knowledge based algorithms are applicable. Some of the successfully deployed algorithms are Support Vector Machines (SVM) [4], [7], [12], Random Forest [19], [20], Decision Trees [4], [5], Neural Networks [9], [11], k-Nearest Neighbor (k-NN) [4], [12], Fuzzy Systems [6], ensemble methods [23], [24], etc. Often these machine learning methods do not provide explainable outcomes and act as black-box prediction models only. Explainable artificial intelligence often helps to assist the business logic by extracting knowledge and rules from the domain. On the other hand the black-box models are only suitable when a specific task such as prediction/classification is required to be performed.

In the context of Bangladesh, which has lately elevated to a developing country, not much work has been done in this regard [26]–[29]. This is due to many fold reasons. Firstly, there is a lack of dataset. Since most of the hospitals have paper based data recording systems, often initial screening records are not stored and maintained properly. Secondly, the immense workload on the physicians and the diagnostic system delay the overall digitization and decision support systems to be used. Thirdly, the financial conditions of a patient often does not support the monitoring within a health framework which is available in the developed countries.

However, in recent years things have started changing due to the digital transformation in the healthcare sector of Bangladesh. On the other hand, Bangladesh is among the highest growing nations in the world in terms of smart mobile phone based Internet users. This have led us to envision smart phone based screening applications to detect leukemia based on symptoms only. A system overview of symptom based leukemia detection is shown in Figure 1. Such a system, if implemented will be able to detect leukemia at an early stage and the screening system may help to reduce the overall load of the physicians. It is important to note that there have been a few works in symptom based disease detection [30] and particularly for cancer detection [31]. However, to the best of our knowledge, there is no such work for leukemia detection using symptoms only.

In this paper, we present a symptom based leukemia screening method based on explainable AI models. This work is an extended version of our initial work [32]. In this work, we have collected primary data from the pediatrics leukemia wards of two top government hospitals located in Dhaka, the capital city of Bangladesh. The dataset is collected following a guidance and policy administered by the consenting physicians and subjects. It is observed that decision tree based supervised learning method is able to predict leukemia based on the early symptoms collected from the patients data and provide explanatory analysis. Moreover, the explainable model performs significantly better compared to other complex and sophisticated black-box type models. The noteworthy contributions made in this paper are as follows:

- A primary dataset on leukemia screening based only on the symptoms is collected from pediatric leukemia ward of two top hospitals in Bangladesh, which can be found in the following link: `https://github.com/AkterHossain312/LeukemiaDataset`.
- Experimental analysis is carried out to show the performances of different machine learning models including a detailed hyper-parameter study.
- A feature analysis and selection study is completed to identify the suitable features that can further enhance the performances of the classification models.
- Experimental results demonstrate that Explainable AI deploying decision tree and Apriori algorithms show satisfactory results compared to other methods. Moreover, the related confidence and support of the rules are generated.

The rest of the paper is organized as follows: a brief literature review is presented in section II; the details of the materials and methods are given in section III; the experimental results and the discussion are presented in section IV and the paper concludes with a summary and outline of the future work in section V.

II. RELATED WORK

There have been several studies to predict leukemia where researchers have applied various machine learning techniques. In this section, first we review the existing works in the global context followed by the present works carried out in the context of Bangladesh.

A. GLOBAL ML BASED RESEARCH ON LEUKEMIA DETECTION

In this subsection, we briefly discuss about the cancer detection work done in the literature in the global context. Most of the works differ from the source of the samples from where the data is collected and the type of the data and the algorithms that have been applied. We have organized the section in terms of the type of the data that is used. However, a summary of methods is given in the upper part of Table 1.

1) BLOOD SAMPLE BASED SCREENING METHODS

Blood samples are often used to screen leukemia. Zelig *et al.* [3] investigated the effectiveness of Fourier Transform Infrared Microscopy (FTIR-MSP) for prescreening and follow-up of leukemia patients undergoing chemo-therapy. They collected blood samples from leukemia patients before and during the treatment, and from healthy subjects who served as control groups. Often the Complete Blood Count (CBC) test taken on blood samples are

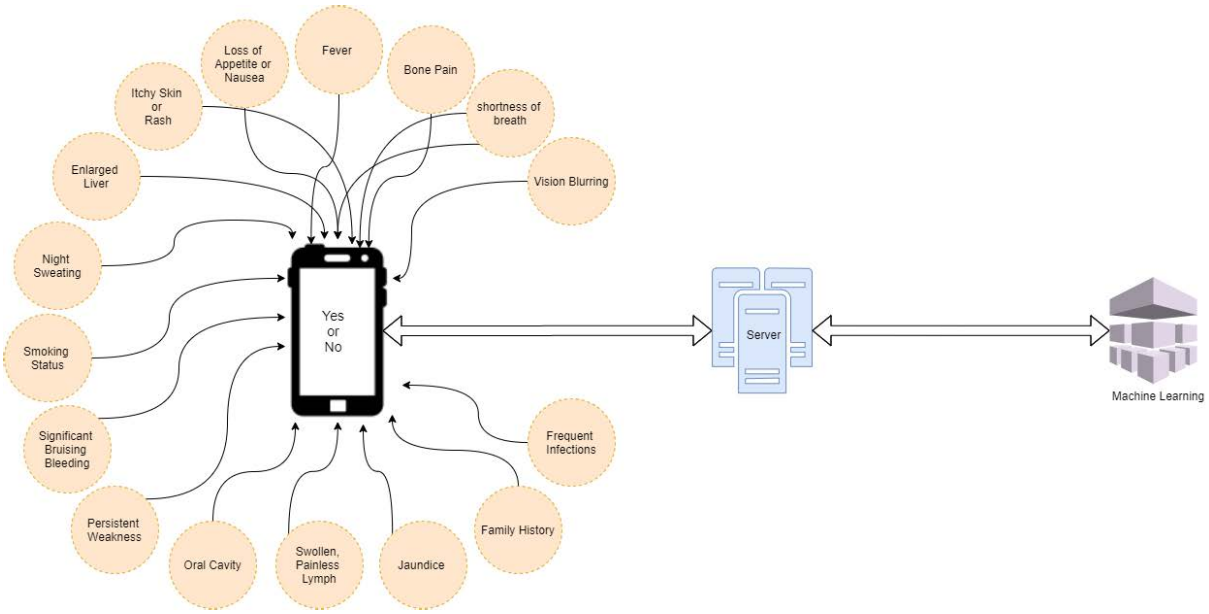


FIGURE 1. A proposed framework for symptom based leukemia detection.

used to screen leukemia [4]–[6]. Daqqa *et al.* [4] achieved 77.30% accuracy using decision tree on patients' gender, age and health status data, along with blood characteristic from CBC test. Mahmood *et al.* [5] obtained a very high accuracy using Classification and Regression Trees (CART) with CBC, Renal Function Test (RFT) and Liver Function Test (LFT) data. Fathi *et al.* [6] investigated the differences between the cute lymphoblastic leukemia and the acute myeloid leukemia using CBC test based data from children. Markiewicz *et al.* [7] used SVM classifiers to recognize the blood cells of myelogenous leukemia.

2) BONE MARROW BASED SCREENING METHODS

Bone marrow data based screening methods had been explored in [8]–[10]. Hsieh *et al.* [8] used SVM on bone marrow and blood peripheral data. Ritter *et al.* [9] developed a supervised machine learning method using a combination of multiple Gaussian mixing models (GMMs). Leinoe *et al.* [10] worked on predicting bleeding in the early stages of acute myeloid leukemia by flow cytometry analysis of platelet function flow.

3) IMAGE BASED SCREENING METHODS

French–American–British (FAB) classification was used by Shafique *et al.* [11] to find sub-types of acute lymphoblastic leukemia based on ALL-IDB dataset. Das *et al.* [12] proposed to use optimized Support Vector Neural Network (SVNN) on the same dataset. Fatma *et al.* [13] also used the same dataset. They applied a color model considering linear contrast. Using neural networks they gained up to 91% accuracy. Rawat *et al.* [14] proposed to analyze color, morphology and textual features from blood images. A genetic algorithm was applied for feature optimization using the SVM classifier. Jha *et al.* [15] also developed a FAB classification-based identification from the Blood Smear images (BSI). The size, texture properties and color of the segmented image extracted by the neural networks were fed to the SVM and Naive Bayes Classifiers.

4) GENOMICS DATA BASED METHODS

In the recent years, genomics and transcriptomics data analysis are playing a very crucial role in cancer related research. Warnat-Herresthal *et al.* [17] used combinations of transcriptomics data for the acute myeloid leukemia prediction. Lee *et al.* [18] used gene expressions for the targeted treatment of acute myeloid leukemia.

5) CLINICAL DATA BASED METHODS

Pan *et al.* [19] applied forward feature selection algorithm to rank the clinical variables and suggested to use Random Forest as a classifier. Chen *et al.* [20] explored different methods for sensing chronic lymphocytic leukemia using ensemble methods. Lin *et al.* [21] used auto-encoders to extract high-level features and used them to predict the acute myeloid leukemia. Fuse *et al.* [22] demonstrated the effectiveness of decision tree algorithms for relapse of acute Leukemia.

6) OTHER METHODS

Kashef *et al.* [23] used paper-based files and analyzed 31 attributes using stacked ensemble classifier with the high area under receiver operating characteristic (auROC) value. Agius *et al.* [24] addressed chronic lymphocytic leukemia (CLL) and used 28 different machine learning algorithms on data from 4,149 patients. Karimi *et al.* [25] conducted a study on the spread of leukemia and lymphoma signs and symptoms in childhood in the southern Iranian province of Fars. They analyzed different symptoms that are highly correlated with different types of leukemia. However,

they did not apply machine learning for decision making based on the symptoms.

B. RESEARCH WORKS IN BANGLADESH

Hossain *et al.* [26] showed a pre-analysis of more than 5,000 confirmed hematological cancer cases from 10 specialized hospitals between January 2006 and December 2012. They mainly showed the prevalence of different types of leukemia among various age groups. Hossain *et al.* [27] counted the sub-types of blood from microscopic images then based on the count of the object they attempted to detect leukemia. The authors used the Faster RCNN models. For this study, they collected approximately 256 images from Dhaka Shishu Hospital and National Institute of Cancer Research and Hospital (NICRH). Abedy *et al.* [28] proposed a scalable leukemia prognosis method based on the universally available ALL_IDB dataset. In another work, Zahra *et al.* [29] investigated the relationship of gene polymorphism in patients with acute lymphoblastic leukemia (ALL) from Bangladesh.

C. SUMMARY

A summary of the literature review is shown in Table 1. It is observed that various types of features have been used for leukemia prediction and screening. It is to be noted that early leukemia detection is possible only from symptoms. Though a good number of methods are used to analyze the symptoms and their correlations with the types of cancers, they are mostly used in combination with other sophisticated features for the prediction model. Often, these models are black-box machine learning and can not provide insights about the decision making process. Moreover, in the context of Bangladesh, not much work have been performed in this regard. Thus, we find a clear research gap and propose a symptom based early screening method for leukemia using explainable AI models.

III. METHODOLOGY

In this section, we present the methodology used in the proposed framework of symptom based leukemia detection as shown in Figure 1. The figure shows that the screening starts from a simple smart phone based questionnaire system that is filled up by a patient. The data sent by the phone is then processed by a server to feed into a machine learning model to find the desired detection.

The complete machine learning workflow is presented in Figure 2. First, we have identified the parameters in consultation with specialist physicians. Then we have collected data from the patients using a survey form. The missing data values and unnecessary columns are removed as a part of pre-processing and data cleaning. The dataset is then divided into train and test datasets. Machine learning models are trained using the train dataset and the resulting model is put into experiments using the test dataset to validate the results. The details of these steps are presented in the rest of the section.

A. DATA COLLECTION

The data collection step is guided by the specialist physicians from one of the largest medical universities of Bangladesh, namely Bangabandhu Sheikh Mujib Medical University (BSMMU). After consulting with the physicians, 16 features or symptoms of leukemia are identified. The data collection is performed from two leading hospitals of Bangladesh: Dhaka Shishu (Children) Hospital and the pediatric ward of the National Institute of Cancer Research and Hospital (NICRH), Dhaka. The data collection is performed with necessary permission and ethical clearance from the authority of the hospitals and only from the consenting subjects.

In total 840 subjects have given consent and participated in data collection. Among them, 131 patients are from NICRH with 103 leukemia patients and 28 non-leukemia patients. Whereas, 709 patients' data is collected from Dhaka Shishu Hospital; 510 of them are leukemia patients and 199 are non-leukemia patients. A summary of the collected data is shown in Table 2.

Thereafter, we separate the datasets into train and test datasets. We have kept the dataset from the National Institute of Cancer Research and Hospital (NICRH) as a test set and used the dataset from the Dhaka Shishu Hospital as a train set.

B. DESCRIPTION OF THE FEATURES

Table 3 presents the 16 features of the dataset along with the class label collected in our research. Note that all of the features have binary meaning i.e., the presence and absence of a symptom. The binary levels are shown as zero (0) and one (1). The distribution of the features is also given in the last two columns of the table. Binary levels make the identification of the symptoms simpler from the users' point of view.

The features used in this research are telltale symptoms. *Shortness of breath* indicates whether someone has long term or short term shortness of breath. Long-term shortness of breath increases the risk of leukemia. People with *bone pain* have an 80% chance of developing leukemia, especially with pain around the spine. Bone pain at night and *fever* are also observed in people who have frequent infections. Family history denotes if there is any genetic linkage with the disease. People who have *frequent infections* are more likely to have leukemia. A leukemia patient with *rash and itching* on her skin may find small red or purple spots on her skin caused by ruptured blood vessels and capillaries under the skin. Among other important features are *loss of appetite or nausea* leading to weight loss, *persistent weakness* and fatigue. *Swollen lymph nodes* in armpits, neck or groin might be one of the early symptoms of leukemia. If the blood vessels under the skin are damaged patients experience *bruising*. Leukemia cells can grow in the liver and spleen and make them bigger. It can be noticed as fullness or bloating, or full feeling after eating only a small amount. The other symptoms related to this are *enlarged liver*, *oral cavity*, *vision blurring*, *jaundice* and *night sweats*. We have also included *smoking* as a feature if the patient is exposed to smoke.

TABLE 1. Summary of the literature review.

Authors	Data Size	Data Type	Data Source	Algorithms
Global Context				
Zelig et al. [3]	15	Peripheral Blood	Soroka University Medical Center	Statistical Analysis
Daqqa et al. [4]	4000	CBC	European Gaza Hospital	SVM, KNN, DT
Mahmood et al. [5]	50	CBC, LFTs, RFTs	Children Hospital and Institute of Child Health	CART, RF, Gradient Boosting, C5.0 DT
Fathi et al. [6]	346	CBC Test	Urmia University of Medical Sciences	Adaptive Neuro-fuzzy Inference System
Markiewicz et al. [7]	17	Blood Cell	_	SVM
Hsieh et al. [8]	7,129	Bone marrow and peripheral blood	BROAD Institute	IG-SVM
Reiter et al. [9]	337	Bone Marrow	CCRI, Charité Berlin, Garrahan Hospital	SVM, Deep Neural Network
Leinoe et al. [10]	45	Bone Marrow	Bayer A/S Diagnostics	Statistical analysis
Shafique et al. [11]	368	Image	ALL-IDB	Deep Neural Network
Das et al. [12]	368	Image	ALL-IDB	GFNB, ELM, KNN, SVM, Naive Bayes, SSDE-based SVNN
Fatma et al. [13]	368	Image	ALL-IDB	Neural Network
Rawat et al. [14]	420	Image	American society of hematology	FAB Classification
Jha et al. [15]	_	Image	Dataset-master and Cellavison	FAB, SVM, Naive Bayesian
Mohapatra et al. [16]	108	Image	Ispat General Hospital, Rourkela	K-means Clustering
Heresta et al. [17]	12,029	Transcriptomic data	Omnibus Database	FAB & Neural Networks
Lee et al. [18]	30	Gene Expression	Local Hospital	MERGE Algorithm
Pan et al. [19]	661	Clinical Data	Guangzhou Women & Children's Medical	RF, DT,LR, SVM
Chen et al. [20]	737	Clinical Data	Mayo Clinic	LR, RF, GBM,Cox, RSF
Lin et al. [21]	94	Clinical Data	TCGA Database	Deep Learning
Fuse et al. [22]	217	Clinical Data	Niigata University and Nagaoka Red Cross Hospital	AD Tree
Kashef et al. [23]	241	Cranial Radiotherapy	Mahak Charity Hospital	Stacked Ensemble Classifier
Agius et al. [24]	4,149	Heterogeneous	Danish National CLL registry	CLL-TIM Ensemble Algorithm
Karimi et al. [25]	368	Categorical Data	Shiraz University of Medical Sciences	Statistical Analysis
Bangladesh Context				
Hossain et al. [26]	5013	Hematological Malignancy Data	Bangladeshi Local Hospitals	FAB Classification
Hossain et al. [27]	256	Image	Local Hospitals	Faster RCNN
Abedi et al. [28]	300	Images	ALL-IDB	Logistic Regression
Zahra et al. [29]	160	Genotyped data	Local Hospitals	Statistical Analysis

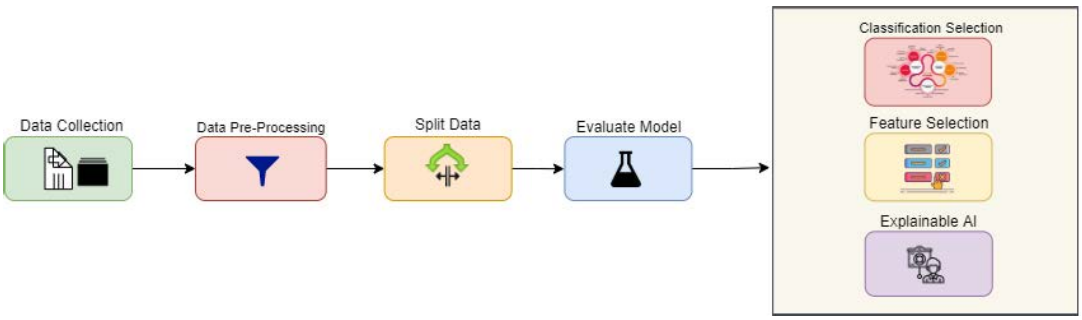


FIGURE 2. Machine learning workflow for early detection of Leukemia.

TABLE 2. Summary of the dataset collected from two hospitals.

Hospital	Total Samples	Positive Samples	Negative Samples
NICRH	131	103	28
Dhaka Sishu Hospital	709	510	199
Total	840	613	227

C. FEATURE SELECTION ALGORITHM

We have used the Least Absolute Shrinkage and Selection Operator (LASSO) model [33] for feature selection. The purpose of the LASSO model is to find the important or dominating features, and thus to regulate the data models. It uses the regression coefficient to select the features. They have been previously used in the recent literature for symptom based machine learning methods in disease diagnosis [30] and particularly in cancer detection [31]. In our experiments, we have observed the effectiveness of the feature selection. The importance of the selected features are determined using different types of classification algorithms and evaluation metrics.

D. CLASSIFICATION METHODS

In our experiments on the dataset, we have used seven different machine learning models: Decision Tree (DT) classifier [34], Random Forest [35], k-Nearest Neighbor [36], Adaboost Classifier [37], Logistic Regression Classifier [38], Naive Bayesian Classifier [38] and Artificial Neural Network [38]. In this section, we briefly discuss about these classifiers. The details parameter studied on each of these classifiers are given in section IV.

TABLE 3. Distribution of the binary features and class label in the dataset.

SL.	Feature Name	Distribution	
		zero	one
1	shortness_of_breath	323	517
2	bone_pain	203	637
3	fever	278	562
4	family_history	229	611
5	frequent_infections	242	598
6	Itchy_skin_or_rash	261	579
7	loss_of_appetite_or_nausea	235	605
8	Persistent_weakness	231	609
9	swollen,painless_lymph	235	605
10	significant_bruising_bleeding	161	679
11	enlarged_liver	162	678
12	oral_cavity	242	598
13	vision_blurring	242	598
14	jaundice	208	632
15	night_sweats	234	606
16	Smoke	259	581
17	**Class label: Leukemia**	600	240

1) DECISION TREE (DT) CLASSIFIER

Decision tree classifiers create a structured decision flow by making decision based on the selected features at each decision node. The features are selected based on the information content. DT classifiers are used to generate rules that are suitable for explainable AI. Often these rules are interpreted by the domain experts. We have used gini index as feature selection metric for decision tree. Gini index [34] is defined as follows.

$$Gini(q) = \sum_{i=1}^{k} p_i(1 - p_i) \tag{1}$$

Here, p_i denotes the probability of an instance being classified in class i among all possible branches created by the attribute q.

2) RANDOM FOREST (RF) CLASSIFIER

Random Forest classifier uses a bootstrapped method to sample the feature space and creating decision tree ensemble based on the selected features. In our experiments, we have set *gini* as the attribute selection metric for the decision trees. The decision of the ensemble is the weighted average of all the predictions made by the decision trees. Random forest is often used successfully for classification of large datasets. However, they are not interpretable compared to DTs. They are used in feature importance analysis. A random forest classifier does the classification based on the predictions made by the constituent tree classifiers defined as in the following equation [35].

$$\hat{y} = sign(\sum_{i=1}^{m} w_i y_i) \tag{2}$$

Here, y_i and w_i denotes the class label predicted and weights assigned to tree i.

3) k-NN CLASSIFIER

The k-Nearest Neighbor classifier is a lazy instance based classifier that uses a weighted voting mechanism to predict the class label of an instance based on its neighboring instances. The neighborhood and the weights are defined by the specific distance metrics selected. k-NN classifiers do not explicitly train, rather selects suitable hyper-parameters and the classification is done on real time. However, they might not be well interpreted for categorical data and depends on the specific label encoding method. The prediction made by a k-NN classifier is a weighted avaerage of the class of the neighbors defined by the following equation [36].

$$\hat{y} = \sum_{i=1}^{k} w_i y_i \tag{3}$$

Here, w_i is the weight of the instance i in the neighborhood assigned based on the distance metric and y_i is the class label of the instance.

4) ADABOOST ALGORITHM

Adaboost is an ensemble algorithm that adaptively improves the performances of the classifiers by changing the weights of the wrongly classified instances dynamically over the iteration. The final classifier provides a weighted prediction of all the single weak classifier predictions that are learned in the iterations. The classification rule of the Adaboost ensemble is given as follows [37].

$$H(\vec{x}) = sign(\alpha_1 h_1(\vec{x}) + \alpha_2 h_2(\vec{x}) + \cdots + \alpha_T h_T(\vec{x})) \tag{4}$$

5) LOGISTIC REGRESSION CLASSIFIER

Logistic regression classifier is a linear classifier that finds a linear boundary to divide the instances. It often uses regularization parameters and a sigmoid function along with the learned weights or parameters. The weights of the logistic regression parameters are learned using a gradient based optimization algorithm. The predicted label of logistic regression is given below [38].

$$\hat{y} = \sigma(w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n) \tag{5}$$

6) NAIVE BAYESIAN CLASSIFIER

The Naive Bayesian classifier uses a simple form of Bayesian net by formulating the class label. This is because the dependent variable or parent variable and all the features that are directly connected to it show direct causal relationship and confirm the conditional independence among the feature variables given in the class label. The classification rule is shown in the following equation [38].

$$\arg\max_k P(c_k|\vec{x}) = \arg\max_k \frac{P(c_k)P(\vec{x}|c_k)}{P(\vec{x})} \tag{6}$$

7) ARTIFICIAL NEURAL NETWORK

Artificial neural networks are models created to imitate the brain functions. However, they can be thought of as layers to nodes inter-connected to each other processing input features using hidden layers and eventually generating prediction at the output layer. Each node in the layers are logistic units. The output of the artificial neural network is defined by a sigmoid

function defined as following [38].

$$\hat{y} = \frac{1}{1 + e^{-z}} \tag{7}$$

Here, z is the input to the activation in the last layer.

E. PERFORMANCE EVALUATION

We have used separate train and test sets to evaluate the performances of the algorithms and methods employed in this paper. Cross-validation is applied to the train set. Here, the dataset is first divided into k non-overlapping sets and in each iteration, $k-1$ sets are used to train and the rest set is used to validate. This is done in k turns or iterations. We have used different values of k to show the robustness of the training of the classifiers.

We have used a set of metrics that are suitable for binary classification performance measurement. These are accuracy, precision, recall, Mathew's Correlation Coefficient (MCC) and area under Receiver Operating Characteristic (auROC) curve. The first four metrics are dependent on the confusion matrix. In the confusion matrix, true positives (TP) are the positive instances that are correctly classified. True negative (TN) denotes the negative instances that are correctly classified. On the other hand, false positive (FP) and false negative (FN) are the instances that are wrongly classified by the classifier as positives and negatives, respectively. Whereas their real class is the opposite. Based on these the metrics are defined in the following equations:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{11}$$

Note that accuracy, precision and recall have their values in the range of [0,1], where 1 represents the perfect classifier and 0 represents the worst classifier. MCC has values in the range of $[-1, +1]$. A positive MCC valued classifier is preferred over a classifier with a negative MCC value. The other metrics, i.e., auROC, has been used without any threshold values selected by the binary classifiers. Note that auROC is often more effective in imbalanced datasets. ROC curve is the curve that plots the true positive rates against the false positive rates. This metric has values in the range of [0, 1], where 0.5 is a random classifier and 1 represents the best classifier.

IV. RESULTS AND DISCUSSION

All the modules we have used in our experiments are based on Python 3.6 and sci-kit learn library [39]. We have used Kaggle notebooks to run the experiments. We have executed all the experiments 10 times and reported the average values only.

A. PERFORMANCE OF THE TRAINING SET

We have applied all the classifiers–DT, RF, k-NN, Adaboost, Naive Bayes, Neural Network and Logistic regression–on the training set where they have been validated using k-fold cross validation ($k = 3, 5, 10$). Table 4 shows the metric values corresponding to each value of k. Please note that these experiments are conducted only on the training set to show the robustness of the methods and to tune or select the hyper-parameters. After this set of experiments, the models with the set of selected parameters are used and applied on the test set.

From Table 4, we can notice that DT algorithm gives the highest accuracy values of 97.14%, 97.54% and 97.74% for 3-, 5- and 10-fold, respectively. Naive Bayes model gives less accuracy values than other models, such as 85.65%, 85.55% and 85.69% for 3-, 5- and 10-fold, respectively. From Random Forest, we get the highest accuracy of 97.74% for 10-fold. With k-NN model the highest value is 87.07% for 10-fold. Moving on to Adaboost, the accuracy value is 92.22% for 10-fold. Finally, Logistic Regression gives 89.37% for 10-fold.

From all these results, it is clear that the DT model gives the highest accuracy value which is 97.74% for 10-fold. It is also to be noted that the other metrics such as precision, recall and MCC are also high for the DT classifier. Moreover, auROC also gives very satisfactory values for the classifiers. Finally, MCC and recall are low for the classifiers such as logistic regression, Naive Bayes and k-NN in all experiments.

B. TEST SET PERFORMANCE

After performing cross-validation and learning models on the training data, we have applied the model to classify the instances from the test set. Table 5 shows comparison between all classification algorithms in terms of their performance in the test set. We have also drawn spider plots based on each of the metrics used in our experiments as shown in Figure 3.

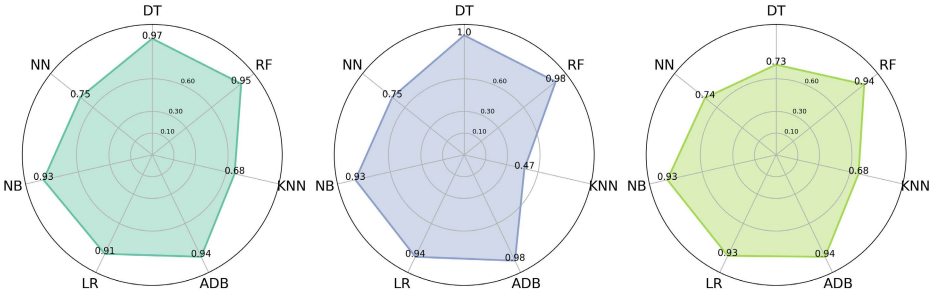
From the table and the figure, we note that using KNN, the results are not much satisfactory with 68.70% accuracy. Logistic Regression and Adaboost have the accuracy of 91.60% and 94.66% respectively. Using Naive Bayes we get an accuracy of 93.13%. Finally, with Decision Tree model we obtain the highest accuracy of 97.45%. Furthermore, using the Random Forest Tree we receive accuracy of 95.41%. It is also to be observed that the performance in the test set in terms of recall, MCC and auROC are not that satisfactory for Decision Tree classifier which is an indication of the possible overfitting in the train set and encourage us to go for further experiments to reduce the overfitting by selecting important features and hyper-parameter tuning of the learning algorithms. We have performed Wilcoxon Sign-Ranked test to ensure the statistical significance with a p-value of 0.046 for the decision tree which we have selected as the best classifier.

TABLE 4. Cross validation results on the train set using different classification algorithms.

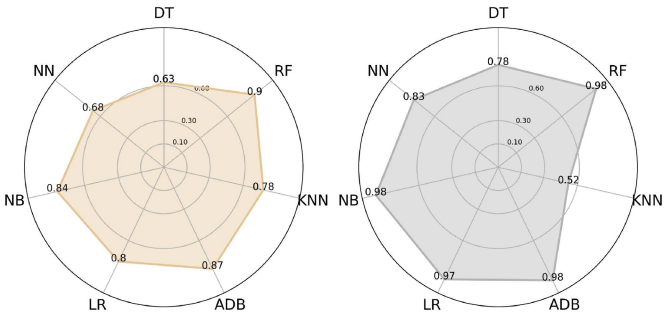
	Algorithms	Accuracy	Precision	Recall	MCC	auROC
3-fold	Decision Tree	97.14%	99.22%	96.30%	0.952	0.975
	Random Forest	96.54%	96.84%	95.44%	0.945	0.995
	KNN	82.99%	100%	71.60%	0.755	0.972
	Adaboost	91.82%	95.10%	91.29%	0.877	0.980
	Logistic Regression	87.38%	96.84%	85.16%	0.851	0.974
	Naive Bayes	85.65%	98.87%	71.19%	0.729	0.958
	Neural Network	89.27%	90.96%	94.44%	0.840	0.979
5-fold	Decision Tree	97.54%	100%	96.51%	0.961	0.979
	Random Forest	98.41%	97.24%	95.48%	0.953	0.998
	KNN	86.12%	100%	71.21%	0.755	0.974
	Adaboost	91.89%	93.85%	89.92%	0.871	0.979
	Logistic Regression	88.37%	97.24%	89.37%	0.841	0.972
	Naïve Bayes	85.55%	100%	72.23%	0.742	0.958
	Neural Network	93.11%	91.99%	95.55%	0.796	0.977
10-fold	Decision Tree	97.74%	100%	96.31%	0.966	0.982
	Random Forest	97.67%	97.19%	96.10%	0.955	0.997
	KNN	87.07%	100%	71%	0.757	0.974
	Adaboost	92.22%	94.11%	90.35%	0.877	0.979
	Logistic Regression	85.82%	97.19%	85.61%	0.854	0.972
	Naïve Bayes	85.69%	99.45%	71.40%	0.734	0.956
	Neural Network	93.85%	94.25%	93.33%	0.727	0.970

TABLE 5. Comparison of each of the classification models based on the test set.

Algorithms	Accuracy	Precision	Recall	MCC	auROC
Decision Tree	97.45%	100%	73.33%	0.630	0.783
Random Forest	95.41%	98.83%	94.44%	0.900	0.987
KNN	68.70%	47.20%	68.70%	0.783	0.521
Adaboost	94.66%	98.83%	94.44%	0.879	0.987
Logistic Regression	91.60 %	94.38%	93.33%	0.806	0.978
Naïve Bayes	93.13%	93.36%	93.13%	0.844	0.984
Neural Network	75.00%	75.58%	74.38%	0.683	0.834



(a) Accuracy, Precision, Recall (from left)



(b) MCC, auROC (from left)

FIGURE 3. Plot of Accuracy, Precision, Recall, MCC and auROC for different classification algorithms.

C. FEATURE SELECTION

We have used the LASSO model to find the feature importance. Figure 4 shows the feature importance of our dataset. We can see through the bar chart that some features have their regression coefficient values very close to 0. Those features are Itchy_skin_or_rash, night_sweats, shortness of breath, oral_cavity, smokes and vision_blurring. We exclude those features and select the features with higher correlation

TABLE 6. Performance of different algorithms on the test set before and after feature selection.

	Algorithms	Accuracy	Precision	Recall	MCC	auROC
Before feature selection	Decision Tree	97.45%	100%	73.33%	0.630	0.783
	Random Forest	95.41%	98.83%	94.44%	0.900	0.987
	Adaboost	94.66%	98.83%	94.44%	0.879	0.987
	KNN	68.70%	47.20%	68.70%	0.783	0.521
	Logistic Regression	91.60 %	94.38%	93.33%	0.806	0.978
	Naïve Bayes	93.13%	93.36%	93.13%	0.844	0.984
	Neural Network	75.00%	75.58%	74.38%	0.683	0.834
After feature selection	Decision Tree	89.31%	95.24%	88.89%	0.765	0.946
	Random Forest	95.41%	97.72%	95.56%	0.895	0.993
	Adaboost	93.64%	97.72%	95.56%	0.893	0.993
	KNN	91.60%	92.47%	90.08%	0.805	0.966
	Logistic Regression	91%	92.39%	94.44%	0.784	0.976
	Naïve Bayes	84.73%	89.74%	84.73%	0.723	0.982
	Neural Network	91.60%	95.40%	92.22%	0.809	0.975

with the class label. We have also applied extra tree classifier based feature ranking and the ranking is shown in Figure 5. We could see the similarities between the feature rankings found by two of the methods.

We have reported the results before and after the feature selection together in Table 6 for different classifiers. Please note that performances have improved after the feature selection for different algorithms. However, accuracies might not reflect that. For example, the accuracy of Decision Tree is 89.31%, which was 97.45% earlier. In case of k-NN and Neural Network the accuracy has been greatly improved. Accuracy is unchanged in Random Forest, slightly degraded in Adaboost and Logistic Regression, and significantly degraded in Naive Bayes. Note that the dataset is a imbalanced one. The improved performances are reflected in MCC and auROC scores for each of these classifiers. We have performed Wilcoxon Sign-Ranked test to ensure the statistical significance with a p-value of 0.0277.

However, the changes are more pronounced in terms of recall, MCC and ROC. Note that for Decision Tree we have noticed recall, MCC and ROC have significantly improved from 73.33%, 0.630 and 0.783 to 88.89%, 0.765 and 0.946, respectively. We also observe very similar results for the Random Forest and Adaboost algorithms. For other algorithms, the values of recall, MCC and ROC are either improved or degraded insignificantly. Over all we can conclude that feature selection has improved different performance parameters.

D. COMPARATIVE ANALYSIS

In this section, we show the comparative analysis of the different classification algorithms used in this paper based on the hyper-parameter space and overfitting.

1) DECISION TREE ALGORITHM

The max_depth hyper-parameter is used to restrict the size of the decision tree and thus reduce overfitting. The graph of Max-Depth vs Accuracy is shown in Figure 6 for the train and test sets where the change of the max-depth is from 3 to 23. We can see from the graphs that for the train set the highest accuracy is obtained for the max-depth at 8 for while the accuracy does not vary much in case of the test set. On the other hand, we can also notice the corresponding performances on the test set for the max_depth.

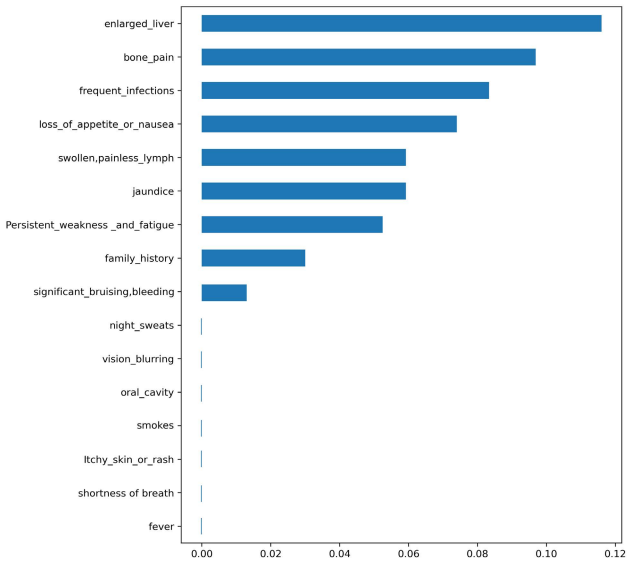


FIGURE 4. Feature importance using LASSO Model.

2) RANDOM FOREST ALGORITHM

In Figure 7, we present the plot of accuracy vs n_estimators for the Random Forest classifier on the train and test sets. We have changed the value of n_estimator from 20 to 1000 estimators and to see the change of accuracy. We have plotted the corresponding accuracy for both the train and test sets. We can observe that the maximum accuracy is obtained when the value of n_estimators is 150 for the train set.

3) ADABOOST ALGORITHM

Similar experiments are performed using the Adaboost algorithm and the results are plotted in Figure 8. Figure 8(a) shows that the accuracy is higher when the value of n_estimators is 100 for the train set.

4) k-NN ALGORITHM

In Figure 9, we have plotted two graphs based on n_neighbours (denoted by k) and accuracy. We have taken the value of n_neighbours between 3 to 23 and reported

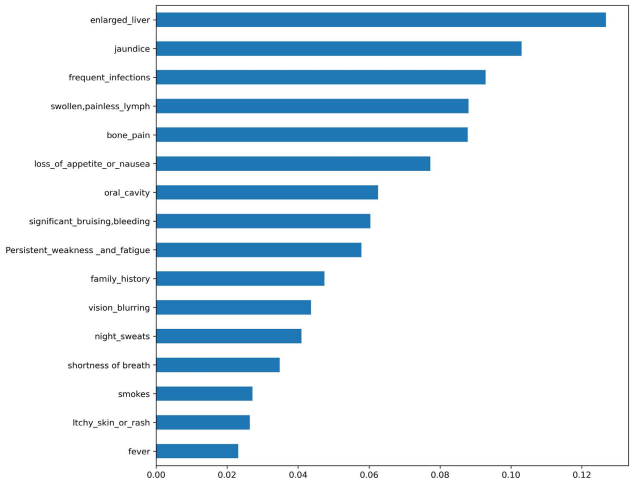


FIGURE 5. Feature importance using Extra-tree Classifier Model.

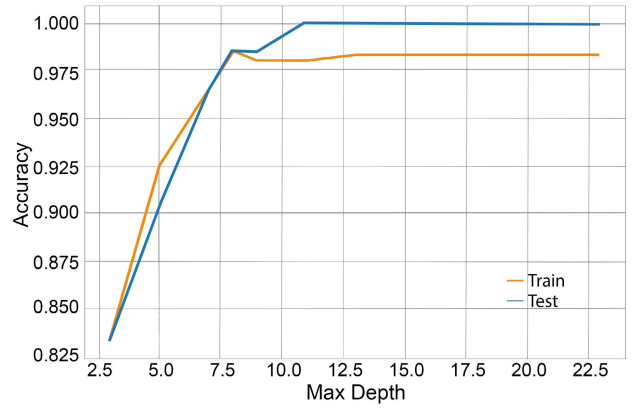


FIGURE 6. Plot of accuracy vs max-depth for the Decision Tree algorithm.

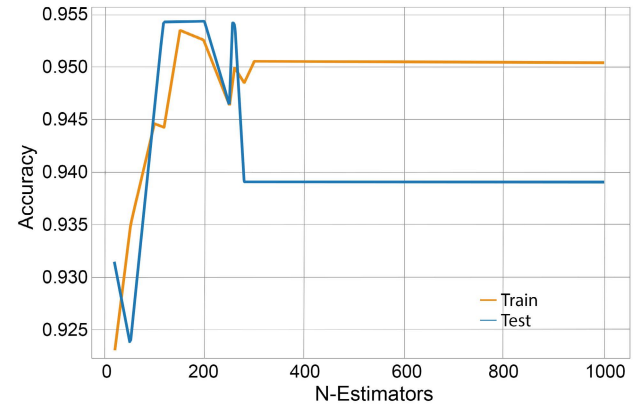


FIGURE 7. Plot of accuracy vs n_estimators for the Random Forest algorithm.

the corresponding accuracy. From Figure 9(a) we note that the highest accuracy for the train set is achieved at n_neighbors=3. We have also applied distance weighted k-NN on the test set to see the performances thereof. However, varying the hyper-parameter k similar to the values for the majority voting k-NN, we could not get higher performances in terms of accuracy.

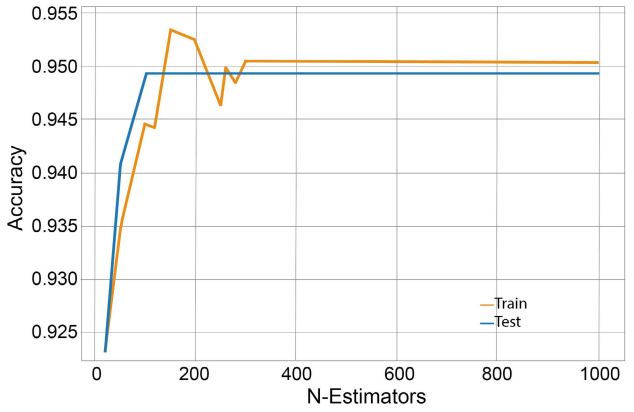


FIGURE 8. Plot of accuracy vs n_estimators for the Adaboost algorithm.

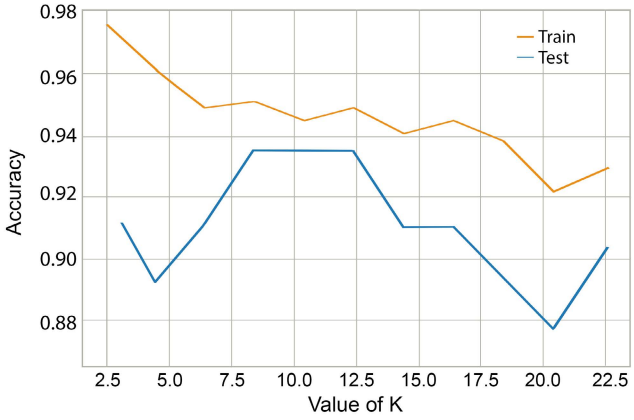
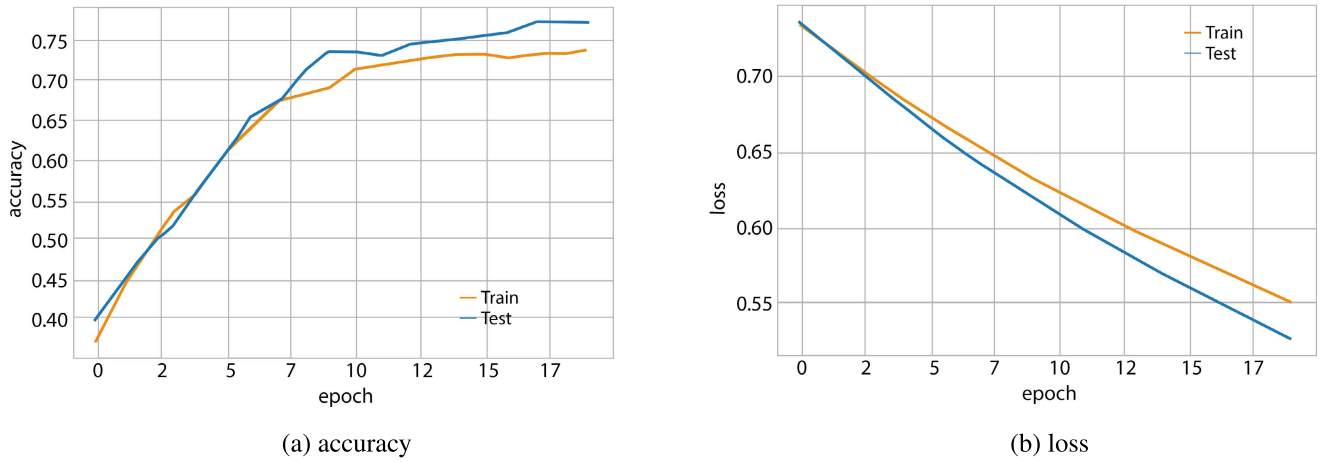


FIGURE 9. Plot of accuracy vs k for the k-NN algorithm.

5) ARTIFICIAL NEURAL NETWORK ALGORITHM

Figure 10 shows the training accuracy vs validation accuracy and the training loss vs validation loss for the Artificial Neural Networks algorithm. From Figure 10(a) we can see that the model can probably be trained more as the trend of accuracy in both datasets is still increasing for the last few epochs. We can further see that the model has not yet learned much more from the training dataset by showing comparable skills with both datasets. On the other hand, From Figure 10(b), we can see that the model has better performance. If these parallel plots begin to exit consistently, it could be a sign of stopping training at an early epochs. Please note that the graphs in Figures 6–10 are showing the model performances of different algorithms with different hyper-parameters. It is interesting to see the fluctuations of a few models for the settings used. However, for each of the classifiers we also note a stable performance for a region in the landscape of the parameters. We have selected the parameters for optimization according to the model behavior. The most fluctuations are shown by Random Forest, which is explained by the randomly selected features by the number of estimators. However, note that after we increase the number of classifiers there are no changes in performances which is due to the relatively small number of features in our dataset.



(a) accuracy (b) loss

FIGURE 10. Plot of accuracy and loss for the Artificial Neural Network algorithm.

The parameters that we have used to get the highest accuracy in case of different algorithms are listed in Table 7. Note that we have not used any parameter tuning for the Logistic Regression algorithm as the regularization parameters are the only parameters to be tuned. We have used the suggested settings from the literature. Also note that the Naive Bayes classifier is a parameter free algorithm. In the table, we show the default settings for these two algorithms. We have reported ROC curves for two of our best performing algorithms on the test set: decision tree and Random Forest as shown in Figure 11.

E. EXPLAINABILITY OF MODELS

Often in machine learning and AI, the results of the algorithms and the system are not explainable due to the black-box nature of the mathematical models related with. In our experiments, we have shown the effectiveness of the Decision Tree model, which gives significantly better results than other algorithms in most of the evaluation metrics. We have already seen the important features that are revealed in the feature selection process and ranking done by the LASSO algorithm. In this section we further extend the explainability analysis by visualizing the models and automatically generating the rules. The rules generated by the models often confirm the existing knowledge and reports new information. There are two types of rules: classification rules and association rules. In this section, we have generated both types of rules and make a comparative analysis on them.

1) CLASSIFICATION RULES

First, we visualize two decision trees in Figure 12 and Figure 13 with max-depth values set to 4 and 8, respectively. Due to the large size of the figure, the decision tree with max-depth has been shown partially in Figure 13.

These figures give us a clear picture of the importance of the features selected as different levels by the Decision Tree. Note that, the selection depends on the attribute selection parameter, information gain, gini, etc. The nodes of the decision trees are making a binary decision by comparing the values associated with it to a threshold value 0.5. Thus the branches actually denote the presence or absence of the particular attribute. From the trees, it is a very simple procedure to generate the classification rules. A path from the root to a leaf node denotes a path for classification. In the following we are showing three classification rules generated from the decision tree (shown in Figure 12) with max-depth equals to 4:

- **rule: 1** frequent_infections $\wedge$ swollen, painless_lymph $\wedge$ Persistent_weakness $\wedge$ enlarged_liver $\Longrightarrow$ Leukemia
- **rule: 2** $\neg$frequent_infections $\wedge$ $\neg$swollen, painless_lymph $\wedge$ $\neg$jaundice $\Longrightarrow$ $\neg$Leukemia
- **rule: 3** frequent_infections $\wedge$ oral_cavity $\Longrightarrow$ $\neg$Leukemia

In the following we show four classification rules derived from the decision tree (shown in Figure 13) with max-depth equals to 8.

- **rule: 1** frequent_infections $\wedge$ swollen, painless_lymph $\wedge$ jaundice $\wedge$ enlarged_liver $\wedge$ shortness_of_breath $\wedge$ smokes $\wedge$ significant_bruising_bleeding $\wedge$ Itchy_skin_or _rash $\Longrightarrow$ Leukemia
- **rule: 2** $\neg$frequent_infections $\wedge$ $\neg$swollen, painless_lymph $\wedge$ $\neg$jaundice $\Longrightarrow$ $\neg$Leukemia
- **rule: 3** frequent_infections $\wedge$ oral_cavity $\wedge$ shortness_of_breath $\wedge$ Persistent_weakness $\wedge$ loss_of _appetite_or _nausea $\wedge$ Itchy_skin_or _rash $\wedge$ fever $\neg$ Leukemia
- **rule: 4** frequent_infections $\wedge$ oral_cavity $\wedge$ shortness_of_breath $\neg$ Leukemia

Note the similarities between the classification rules generated by the two different decision trees vary in depth. Lower-depth decision trees provide better generalization capacity compared to higher-depth trees. Although, higher-depth decision trees achieve higher accuracy, they are sometimes prone to overfitting. These rules indicate consistency of the decision trees and they are very much interpretable and ready to be used in any application for leukemia screening.

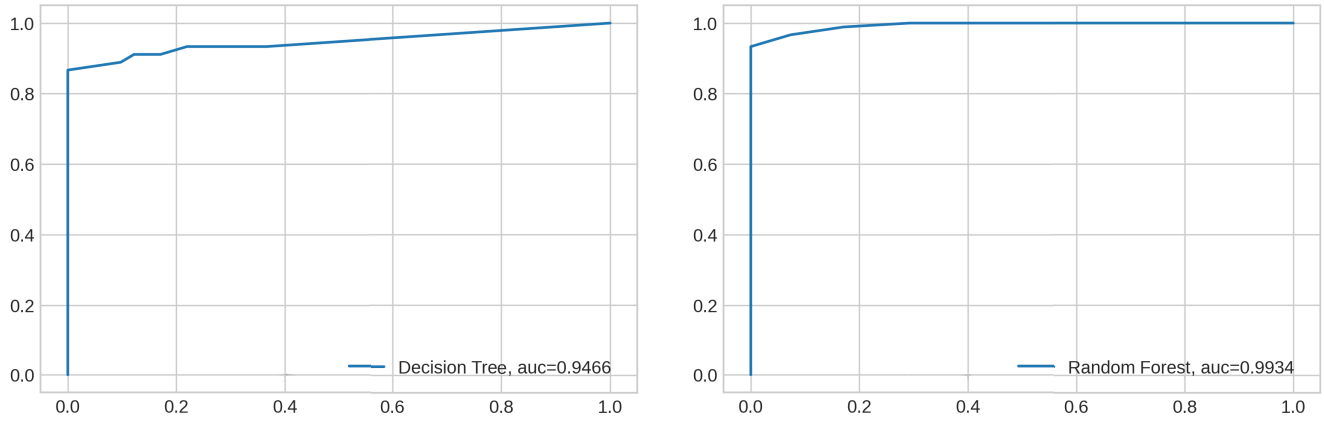


FIGURE 11. Receiver Operating Characteristic Curves for (a) Decision Tree and (b) Random Forest Classifier.

TABLE 7. Best hyper-parameters that are used in the experiments.

Algorithms	Parameters
Decision Tree	Max-depth=8, criterion=gini, random size=0, splitter=best, min-samples-split=2
Random Forest	n_estimators=150, criterion=gini, max-features=auto, random-state=0, verbose=0
k-NN	n_neighbors=3,5,10, weights=uniform, n-jobs=none
Adaboost	n_estimators=100, random-state=90, learning-rate=1
Logistic Regression	penalty=l2, c=0.1, random-state=none, dual=false
Naïve Bayes	priors, var-smoothing=default
Neural Network	validation-split=0.33, batch-size=200, epochs=20, verbose=0

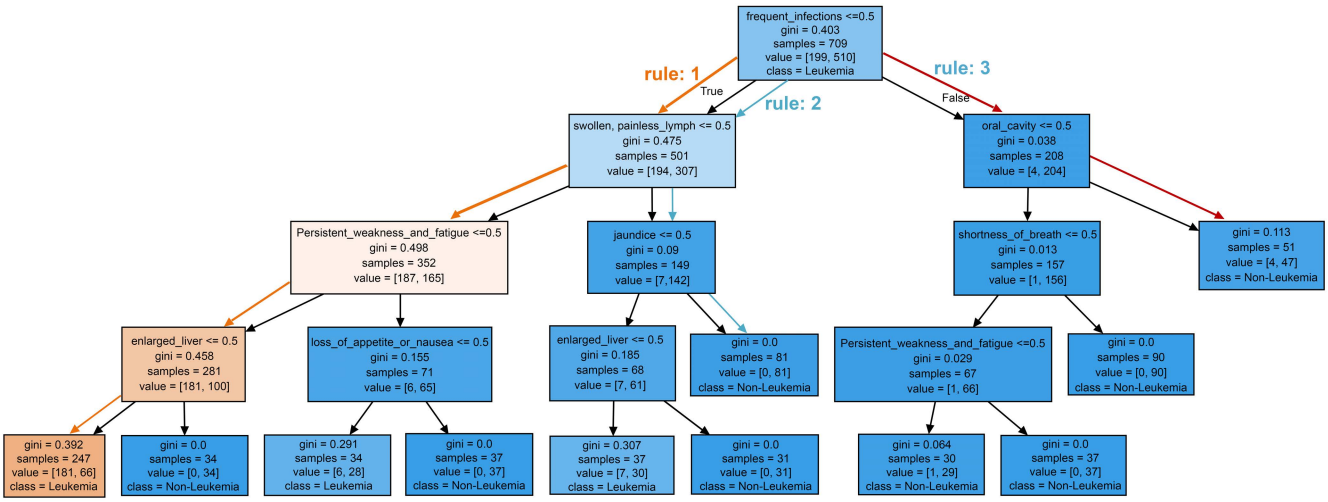


FIGURE 12. Decision Tree graph with max-depth equals to 4.

F. ASSOCIATION RULES

In explainable AI, often association rules are generated from the frequent itemsets that are present in the datasets. Often the quality of the rules are measured in terms of confidence, lift and support as defined below.

$$Support = \frac{|X \subseteq |T}{|T|}$$

$$Confidence(X \implies Y) = \frac{Support(X \cup Y)}{Support(X)}$$

$$Lift(X \implies Y) = \frac{Support(X \cup Y)}{Support(X) \times Support(Y)}$$

Here X and Y are two non-empty itemsets that are subsets of a frequent itemset based on a threshold and works as premise and conclusion of a rule. T denotes the set of all instances in the dataset. Here, support is denoted as how popular an itemset is, which is measured by the proportion of the transactions where an itemset appears. Confidence denotes as the presence of one itemset, which indicates the likelihood of the presence of another itemset. Here, items are indicating the symptoms in the dataset we have considered.

Table 8 shows the rules of the Apriori algorithm without including the class labels as an item. These are the rules that show the co-occurrences among the items or symptoms that

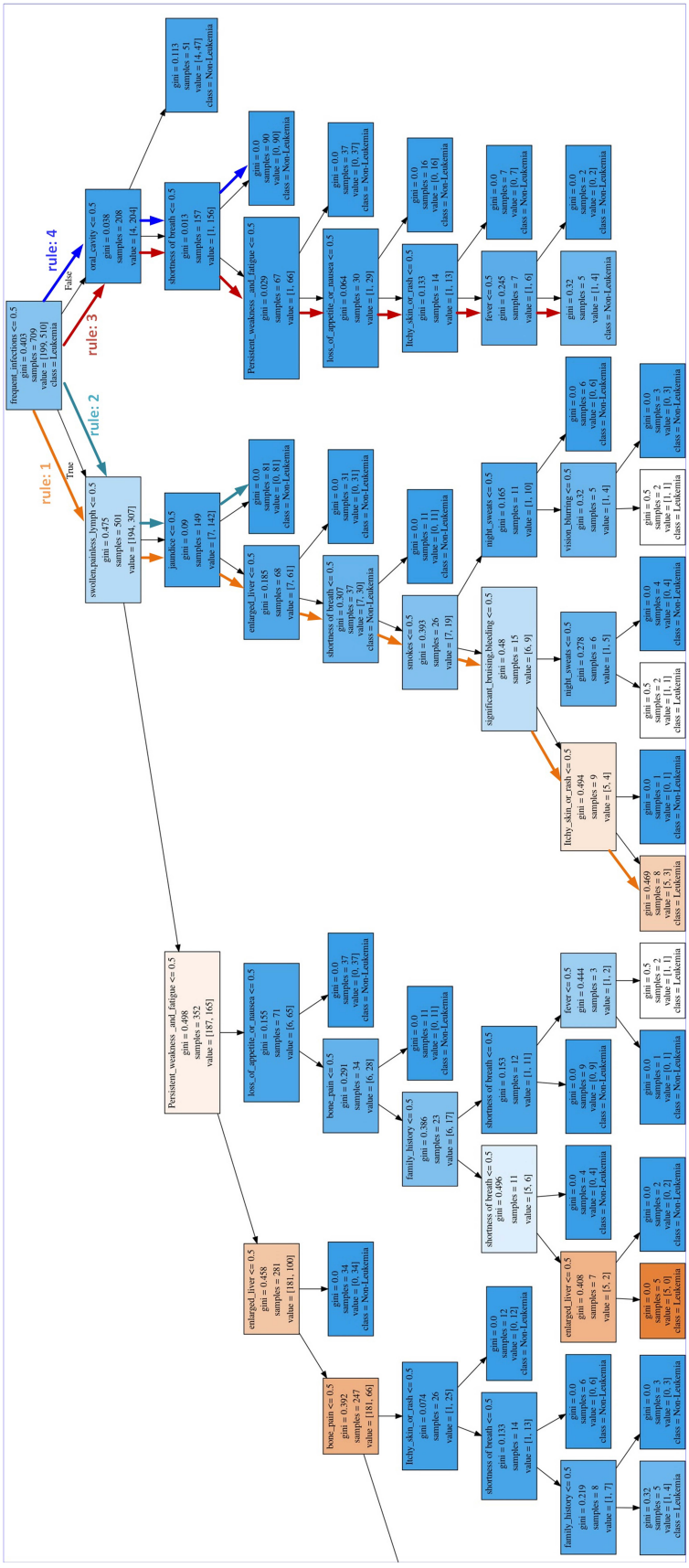


FIGURE 13. Decision Tree graph with max-depth equals to 8.

TABLE 8. Rules from the Apriori algorithm without using the class labels.

Rules	Support	Confidence	lift
rules{enlarged_liver ⟹ oral_cavity}	0.0148	0.2444	6.4603
rules{ bone_pain ⟹ enlarged_liver}	0.0148	0.2444	8.2222
rules{family_history ⟹ frequent_infections}	0.0445	0.2750	3.7000
rules{jaundice ⟹ oral_cavity}	0.0148	0.1447	3.8251
rules{frequent_infections ⟹ jaundice}	0.0148	0.1447	4.8684
rules{significant_bruising,bleeding ⟹ swollen, painless_lymph}	0.0162	0.1304	4.0217
rules{frequent_infections ⟹ Persistent_weakness_and _fatigue & family_history}	0.0148	0.2000	5.2857
rules{Persistent_weakness_and_fatigue ⟹ frequent _infections & loss_of_appetite_or_nausea}	0.0297	0.0936	3.1489
rules{family_history ⟹ bone_pain & frequent_infections}	0.0175	0.1083	6.1667
rules{shortness of breath ⟹ bone_pain & fever}	0.0378	0.3500	4.9807
rules{ shortness of breath ⟹ bone_pain & jaundice}	0.0135	0.1250	7.7083

TABLE 9. Rules from the Apriori algorithm using the class labels.

Rules	Support	Confidence	lift
rules{enlarged_liver ⟹ oral_cavity}	0.0148	0.2444	6.4603
rules{ bone_pain ⟹ enlarged_liver}	0.0148	0.2444	8.2222
rules{family_history ⟹ frequent_infections}	0.0445	0.2750	3.7000
rules{jaundice ⟹ oral_cavity}	0.0148	0.1447	3.8251
rules{frequent_infections ⟹ jaundice}	0.0148	0.1447	4.8684
rules{significant_bruising,bleeding ⟹ swollen, painless_lymph}	0.0162	0.1304	4.0217
rules{frequent_infections ⟹ Persistent_weakness_and _fatigue & family_history}	0.0148	0.2000	5.2857
rules{Persistent_weakness_and_fatigue ⟹ frequent _infections & loss_of_appetite_or_nausea}	0.0297	0.0936	3.1489
rules{family_history ⟹ bone_pain & frequent_infections}	0.0175	0.1083	6.1667
rules{shortness of breath ⟹ bone_pain & fever}	0.0378	0.3500	4.9807
rules{ shortness of breath ⟹ bone_pain & jaundice}	0.0135	0.1250	7.7083
rules{ leukemia ⟹ Persistent_weakness_and_fatigue}	0.0535	0.1724	3.1484
rules{ smokes ⟹ leukemia}	0.0535	0.1948	3.6364
rules{ bone_pain ⟹ leukemia}	0.0357	0.1149	3.1145
rules{ loss_of_appetite_or_nausea ⟹ leukemia}	0.0547	0.1762	3.1499
rules{ shortness_of_breath ⟹ leukemia}	0.0654	0.2107	3.2183
rules{ significant_bruising_bleeding ⟹ leukemia}	0.0357	0.1442	4.0384
rules{swollen_painless_lymph ⟹ leukemia}	0.0654	0.2644	3.9663
rules{leukemia ⟹ frequent_infection & enlarged_liver}	0.0535	0.2163	4.1379
rules{leukemia ⟹ Persistent_weakness_and_fatigue & jaundice}	0.0214	0.0647	3.0215
rules{leukemia ⟹ loss_of_appetite_or_nausea & bone_pain}	0.0214	0.0647	3.0215
rules{leukemia ⟹ loss_of_appetite_or_nausea & bone_pain & family_history}	0.0369	0.1527	4.1379

are important for leukemia screening. It is interesting to see the effect of the association rules generated when the class labels are included in the dataset. The association rules are shown in Table 9. Please note the similarities between the two sets of the rules. However, this time, a few of the rules where leukemia is present in the conclusion or premise encourage us to use them as similar to the classification rules. We can also see the symptoms that were selected by the decision tree such as 'frequent_infection', 'loss_of_appetite_or_nausea' or 'bone_pain' are selected as important features by all three types of analysis: feature ranking, decision tree based rule generation and association analysis by Apriori algorithm.

G. AVAILABILITY OF METHOD

To ensure that our method is reproducible and usable for other researchers in the community we have made all the necessary source code and dataset freely available. It can be accessible from here: `https://github.com/AkterHossain312/LeukemiaDataset`.

V. CONCLUSION

In this paper, we have presented an explainable machine learning model for leukemia detection. We have used a primary dataset collected from two government hospitals in Bangladesh. In a developing country like Bangladesh, data collection is a great challenge. However, after the data collection we have applied various ML models and shown a comparative analysis. We have also performed explainable analysis to generate classification and association rules that are interpretable and usable in leukemia screening.

In our experiments we have seen that simple and explainable model like the decision tree classifier performs best results when compared to the other methods which are more sophisticated. We have also shown that the similar symptoms are selected by all three types of explainable analysis: feature ranking, decision tree based rule generation and association analysis by Apriori algorithm. The symptoms that were selected mostly are 'frequent_infection', 'loss_of_appetite _or_nausea', 'bone_pain', etc.

We strongly believe that this is a first of the kind work carried out in the context of Bangladesh. This study will provide benefits in early detection and screening of leukemia in research and in practice as well.

One of the future work is to enhance the dataset by incorporating more samples from the relevant hospital wards. This will also initiate the requirements of a study to validate

the results from this pilot dataset and the symptom based prediction model. We strongly believe this pilot model will help building an enhanced dataset which in turn will help strengthen the model after further analysis and model building.

ACKNOWLEDGMENT

The authors sincerely acknowledge the advice and consultation given by Bangabandhu Sheikh Mujib Medical University Hospital (BSMMU). They also acknowledge Dhaka Shishu (Children) Hospital and the National Institute of Cancer Research & Hospital (NICRH) for providing them training and blood sample data.

REFERENCES

[1] (2019). *Medical News Today*. Accessed: Nov. 2, 2020. [Online]. Available: https://www.medicalnewstoday.com/articles/142595

[2] (2019). *Medical News Today*. Accessed: Nov. 2, 2020. [Online]. Available: https://www.medicalnewstoday.com/articles/282929

[3] U. Zelig, S. Mordechai, G. Shubinsky, R. K. Sahu, M. Huleihel, E. Leibovitz, I. Nathan, and J. Kapelushnik, "Pre-screening and follow-up of childhood acute leukemia using biochemical infrared analysis of peripheral blood mononuclear cells," *Biochim. et Biophys. Acta (BBA) Gen. Subjects*, vol. 1810, no. 9, pp. 827–835, Sep. 2011.

[4] K. A. S. A. Daqqa, A. Y. A. Maghari, and W. F. M. A. Sarraj, "Prediction and diagnosis of leukemia using classification algorithms," in *Proc. 8th Int. Conf. Inf. Technol. (ICIT)*, May 2017, pp. 638–643.

[5] N. Mahmood, S. Shahid, T. Bakhshi, S. Riaz, H. Ghufran, and M. Yaqoob, "Identification of significant risks in pediatric acute lymphoblastic leukemia (ALL) through machine learning (ML) approach," *Med. Biol. Eng. Comput.*, vol. 58, no. 11, pp. 2631–2640, Nov. 2020.

[6] E. Fathi, M. J. Rezaee, R. Tavakkoli-Moghaddam, A. Alizadeh, and A. Montazer, "Design of an integrated model for diagnosis and classification of pediatric acute leukemia using machine learning," *Proc. Inst. Mech. Eng., H, J. Eng. Med.*, vol. 234, no. 10, pp. 1051–1069, Oct. 2020.

[7] T. Markiewicz, S. Osowski, B. Marianska, and L. Moszczynski, "Automatic recognition of the blood cells of myelogenous leukemia using SVM," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, vol. 4, Jul. 2005, pp. 2496–2501.

[8] S.-H. Hsieh, Z. Wang, P.-H. Cheng, I.-S. Lee, S.-L. Hsieh, and F. Lai, "Leukemia cancer classification based on support vector machine," in *Proc. 8th IEEE Int. Conf. Ind. Informat.*, Jul. 2010, pp. 819–824.

[9] M. Reiter, M. Diem, A. Schumich, M. Maurer-Granofszky, L. Karawajew, J. G. Rossi, R. Ratei, S. Groeneveld-Krentz, E. O. Sajaroff, S. Suhendra, M. Kampel, and M. N. Dworzak, "Automated flow cytometric MRD assessment in childhood acute B-lymphoblastic leukemia using supervised machine learning," *Cytometry A*, vol. 95, no. 9, pp. 966–975, Sep. 2019, doi: 10.1002/cyto.a.23852.

[10] E. B. Leinoe, M. H. Hoffmann, E. Kjaersgaard, J. D. Nielsen, O. J. Bergmann, T. W. Klausen, and H. E. Johnsen, "Prediction of haemorrhage in the early stage of acute myeloid leukaemia by flow cytometric analysis of platelet function," *Brit. J. Haematol.*, vol. 128, no. 4, pp. 526–532, Feb. 2005.

[11] S. Shafique and S. Tehsin, "Acute lymphoblastic leukemia detection and classification of its subtypes using pretrained deep convolutional neural networks," *Technol. Cancer Res. Treatment*, vol. 17, Sep. 2018, Art. no. 1533033818802789.

[12] B. K. Das and H. S. Dutta, "Infection level identification for leukemia detection using optimized support vector neural network," *Imag. Sci. J.*, vol. 67, no. 8, pp. 417–433, Nov. 2019.

[13] M. Fatma and J. Sharma, "Identification and classification of acute leukemia using neural network," in *Proc. Int. Conf. Med. Imag., m-Health Emerg. Commun. Syst. (MedCom)*, Nov. 2014, pp. 142–145.

[14] J. Rawat, A. Singh, B. Hs, J. Virmani, and J. S. Devgun, "Computer assisted classification framework for prediction of acute lymphoblastic and acute myeloblastic leukemia," *Biocybern. Biomed. Eng.*, vol. 37, no. 4, pp. 637–654, 2017.

[15] K. K. Jha, P. Das, and H. S. Dutta, "FAB classification based leukemia identification and prediction using machine learning," in *Proc. Int. Conf. Syst., Comput., Autom. Netw. (ICSCAN)*, Jul. 2020, pp. 1–6.

[16] S. Mohapatra and D. Patra, "Automated leukemia detection using Hausdorff dimension in blood microscopic images," in *Proc. INTERACT*, Dec. 2010, pp. 64–68.

[17] S. Warnat-Herresthal, K. Perrakis, B. Taschler, M. Becker, K. Baßler, M. Beyer, P. Günther, J. Schulte-Schrepping, L. Seep, K. Klee, T. Ulas, T. Haferlach, S. Mukherjee, and J. L. Schultze, "Scalable prediction of acute myeloid leukemia using high-dimensional machine learning and blood transcriptomics," *iScience*, vol. 23, no. 1, Jan. 2020, Art. no. 100780. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2589004219305255

[18] S.-I. Lee, S. Celik, B. A. Logsdon, S. M. Lundberg, T. J. Martins, V. G. Oehler, E. H. Estey, C. P. Miller, S. Chien, J. Dai, A. Saxena, C. A. Blau, and P. S. Becker, "A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia," *Nature Commun.*, vol. 9, no. 1, pp. 1–13, Dec. 2018.

[19] L. Pan, G. Liu, F. Lin, S. Zhong, H. Xia, X. Sun, and H. Liang, "Machine learning applications for prediction of relapse in childhood acute lymphoblastic leukemia," *Sci. Rep.*, vol. 7, no. 1, pp. 1–9, Aug. 2017.

[20] D. Chen, G. Goyal, R. Go, S. Parikh, and C. Ngufor, "Predicting time to first treatment in chronic lymphocytic leukemia using machine learning survival and classification methods," in *Proc. IEEE Int. Conf. Healthcare Informat. (ICHI)*, Jun. 2018, pp. 407–408.

[21] M. Lin, V. Jaitly, I. Wang, Z. Hu, L. Chen, M. A. Wahed, Z. Kanaan, A. Rios, and A. N. D. Nguyen, "Application of deep learning on predicting prognosis of acute myeloid leukemia with cytogenetics, age, and mutations," Tech. Rep., 2018.

[22] K. Fuse, S. Uemura, S. Tamura, T. Suwabe, T. Katagiri, T. Tanaka, T. Ushiki, Y. Shibasaki, N. Sato, T. Yano, T. Kuroha, S. Hashimoto, T. Furukawa, M. Narita, H. Sone, and M. Masuko, "Patient-based prediction algorithm of relapse after allo-HSCT for acute leukemia and its usefulness in the decision-making process using a machine learning approach," *Cancer Med.*, vol. 8, no. 11, pp. 5058–5067, Sep. 2019, doi: 10.1002/cam4.2401.

[23] A. A. Kashef, T. Khatibi, and A. Mehrvar, "Prediction of cranial radiotherapy treatment in pediatric acute lymphoblastic leukemia patients using machine learning: A case study at MAHAK hospital," *Asian Pacific J. Cancer Prevention*, vol. 21, no. 11, pp. 3211–3219, Nov. 2020.

[24] R. Agius, C. Brieghel, M. A. Andersen, A. T. Pearson, B. Ledergerber, A. Cozzi-Lepri, Y. Louzoun, C. L. Andersen, J. Bergstedt, J. H. von Stemann, M. Jørgensen, M.-H.-E. Tang, M. Fontes, J. Bahlo, C. D. Herling, M. Hallek, J. Lundgren, C. R. MacPherson, J. Larsen, and C. U. Niemann, "Machine learning can identify newly diagnosed patients with CLL at high risk of infection," *Nature Commun.*, vol. 11, no. 1, pp. 1–17, Dec. 2020.

[25] M. Karimi, D. Mehrabani, H. Yarmohammadi, and F. S. Jahromi, "The prevalence of signs and symptoms of childhood leukemia and lymphoma in Fars Province, Southern Iran," *Cancer Detection Prevention*, vol. 32, no. 2, pp. 178–183, Jan. 2008.

[26] M. S. Hossain, M. S. Iqbal, M. A. Khan, M. G. Rabbani, H. Khatun, S. Munira, M. Miah, A. L. Kabir, N. Islam, T. F. Dipta, and F. Rahman, "Diagnosed hematological malignancies in Bangladesh—A retrospective analysis of over 5000 cases from 10 specialized hospitals," *BMC Cancer*, vol. 14, no. 1, pp. 1–7, Dec. 2014.

[27] M. Akter Hossain, M. Islam Sabik, I. Muntasir, A. K. M. Muzahidul Islam, S. Islam, and A. Ahmed, "Leukemia detection mechanism through microscopic image and ML techniques," in *Proc. IEEE REGION 10 Conf. (TENCON)*, Nov. 2020, pp. 61–66.

[28] H. Abedy, F. Ahmed, M. N. Qaisar Bhuiyan, M. Islam, N. Y. Ali, and M. Shamsujjoha, "Leukemia prediction from microscopic images of human blood cell using HOG feature descriptor and logistic regression," in *Proc. 16th Int. Conf. ICT Knowl. Eng. (ICT&KE)*, Nov. 2018, pp. 1–6.

[29] F. T. Zahra, N. A. Nahid, M. R. Islam, M. M. A. Al-Mamun, M. N. H. Apu, Z. Nahar, A. L. Kabir, S. K. Biswas, M. U. Ahmed, M. S. Islam, and A. Hasnat, "Pharmacogenetic variants in MTHFR gene are significant predictors of methotrexate toxicities in Bangladeshi patients with acute lymphoblastic leukemia," *Clin. Lymphoma Myeloma Leukemia*, vol. 20, no. 2, pp. e58–e65, Feb. 2020. [Online]. Available: https://www.sciencedirect.com/

[30] Y. Takahashi, M. Ueki, M. Yamada, G. Tamiya, I. N. Motoike, D. Saigusa, M. Sakurai, F. Nagami, S. Ogishima, S. Koshiba, K. Kinoshita, M. Yamamoto, and H. Tomita, "Improved metabolomic data-based prediction of depressive symptoms using nonlinear machine learning with feature selection," *Transl. Psychiatry*, vol. 10, no. 1, pp. 1–12, Dec. 2020.
[31] Y. R. Park, Y. J. Kim, W. Ju, K. Nam, S. Kim, and K. G. Kim, "Comparison of machine and deep learning for the classification of cervical cancer based on cervicography images," *Sci. Rep.*, vol. 11, no. 1, pp. 1–11, Dec. 2021.
[32] M. A. Hossain, M. I. Sabik, M. M. Rahman, S. N. Sakiba, A. M. Islam, S. Shatabda, S. Islam, and A. Ahmed, "An effective leukemia prediction technique using supervised machine learning classification algorithm," in *Proc. Int. Conf. Trends Comput. Cognit. Eng.*, Cham, Switzerland: Springer, 2021, pp. 219–229.
[33] V. Fonti and E. Belitser, "Feature selection using lasso," *VU Amsterdam Res. Paper Bus. Anal.*, vol. 30, pp. 1–25, Mar. 2017.
[34] W.-Y. Loh, "Classification and regression trees," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 1, no. 1, pp. 14–23, 2011.
[35] T. K. Ho, "Random decision forests," in *Proc. 3rd Int. Conf. Document Anal. Recognit.*, vol. 1, Aug. 1995, pp. 278–282.
[36] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *Amer. Statist.*, vol. 46, no. 3, pp. 175–185, 1992.
[37] R. E. Schapire, "The boosting approach to machine learning: An overview," in *Nonlinear Estimation Classification*, 2003, pp. 149–171.
[38] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. London, U.K.: Pearson, 2002.
[39] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011.

MOHAMMAD AKTER HOSSAIN received the B.Sc. degree in computer science and engineering from United International University (UIU). His honors thesis was on "Towards IoT and machine learning driven blood cancer prediction system using machine learning and IoT." He was working as a Research Assistant (RA) with United International University (UIU) and worked on healthcare-based projects. He is currently working as an Associate Software Engineer at Kaz Software, Bangladesh.

A. K. M. MUZAHIDUL ISLAM (Senior Member, IEEE) received the M.Sc. degree in computer science and engineering from the Kharkiv National University of Radio Electronics, Ukraine, and the D.Eng. degree in the field of computer science and engineering from the Nagoya Institute of Technology, Japan. From January 2011 to January 2017, he has served as a Senior Lecturer with the Malaysia-Japan International Institute of Technology (MJIIT), Universiti Teknologi Malaysia (UTM), Malaysia. He has also served as the Head of the Department for the Department of Computer Science and Engineering (CSE), University of Liberal Arts Bangladesh (ULAB). He is currently a Professor with the Department of Computer Science and Engineering (CSE), United International University (UIU), Bangladesh. His research interests include network architecture, communication protocol, cognitive radio network, wireless sensor networks, the IoT, cloud computing, healthcare, and smart farming. He has published 99 research articles and secured over ten national and international research grants. He also supervised many Ph.D., master's, and B.Eng. students through their graduation. He is a member of the BAETE's Sectoral Committee. He was the General Chair of ICBBDB 2021 and the Program Chair of ICAICT 2016 and 2020 and ETCCE 2020 and 2021 international conferences. He has also served as the Secretariat of ICaTAS 2016 international conference, Malaysia, and the 7th AUN/SEED-Net 2014 International Conference on EEE. He is a Chartered Engineer (CEng) and a fellow of IEB (FIEB).

SALEKUL ISLAM (Senior Member, IEEE) received the Ph.D. degree from the Department of Computer Science and Software Engineering, Concordia University, Canada, in June 2008. He is currently a Professor and the Head of the Department of Computer Science and Engineering (CSE), United International University (UIU), Dhaka, Bangladesh. He also worked as an FRQNT Postdoctoral Fellow with the Énergie, Matériaux et Télécommunications (EMT) Center, Institut national de la recherche scientifique (INRS), Montréal, Canada. He was a Visiting Faculty Member of Anglia Ruskin University, U.K., through the EU-FUSION Project, in 2015. His research areas mainly focus on future internet architecture, blockchain, edge cloud computing, software-defined network (SDN), multicast security, and security protocol validation. He is serving as an Associate Editor for the IEEE Access journal.

SWAKKHAR SHATABDA received the bachelor's degree from the Department of CSE, Bangladesh University of Engineering and Technology (BUET), and the Doctor of Philosophy degree from the Institute for Integrated and Intelligent Systems (IIIS), Griffith University, in 2014. He is currently working as a Professor with the Department of Computer Science and Engineering (CSE), United International University. His research interests are in the field of AI search, optimization, machine learning, and computational biology. His works are published in reputed journals, such as the *Scientific Reports*, *Bioinformatics*, IEEE Access, *Information Sciences*, and *Genomics*. He is currently serving as an Academic Editor for *PLOS One*.

ASHIR AHMED (Member, IEEE) received the Ph.D. degree in information science from Tohoku University, Japan. He is currently an Associate Professor with the Faculty of Information Science and Electrical Engineering, Kyushu University, Japan. His research aims to produce disruptive technologies to solve social problems focusing on digital healthcare systems. His team has developed the portable health clinic (www.portablehealth.clinic) system to provide affordable healthcare services to the unreached community. He worked for Avaya Laboratories and NTT Communications on VoIP systems. His research interests include ICTD, digital health, explainable AI, and social business.