International Journal of Advanced Science and Technology
Vol. 29, No. 4, (2020), pp. 6359 - 6376

Development of a Clinical Predictive Model for Stratification of Cancerous Diseases: A Case Study of Chronic Myeloid Leukemia

Oluwabunmi Omobolanle Olaniyan[1,*], Adewale Opeoluwa Ogunde[2], Toluwase Ayobami Olowookere[3], Idowu Sunday Oyetade[4]
[1,2,3,4] *Department of Computer Science, Redeemer's University, Ede, Osun State, Nigeria*
[1,*] *olaniyano@run.edu.ng, ogundea@run.edu.ng, olowookereta@run.edu.ng, oyetadei@run,edu.ng*

Abstract

Scoring systems are typically used to stratify Chronic Myeloid Leukemia (CML) disease into their risk groups towards cure and survival prolongation. These systems, however, do not computationally handle very large datasets due to noise and overfitting of data. In literature, Machine Learning (ML) algorithms have been used to extract meaningful information from datasets, and their performances measured based on metrics such as accuracy and time to learn, among others. Nevertheless, the loss function (empirical risk) of the ML algorithms has been found not to have been largely considered to determine the risks incurred in adopting the ML algorithms for stratification. The aim of this study was to develop an Empirical Risk Minimization Data Stratification (ERMDS) algorithm that can aid the stratification of Chronic Myeloid Leukemia dataset. The algorithm developed would aid the development of a clinical predictive model using an application called ChroMyL app. A secondary dataset of 1640 CML patients, between 2003 and 2017 was collected from Obafemi Awolowo University Teaching Hospitals Complex, Ile-Ife, Osun State, Nigeria, and mined in WEKA 3.8.0 using basophil count and spleen size values on four ML algorithms (BayesNet, Multilayered perceptron, Projective Adaptive Resonance Theory (PART) and Logistic Regression). The algorithm with the highest performance was used in developing the ERMDS algorithm. Based on the analysis of the four classification algorithms carried out on five performance metrics which are: correctly classified instance, time to learn, kappa statistics, sensitivity and specificity, Logistic Regression had the highest accuracy value of 99.82%. As such, the ERMDS algorithm was developed using L1-regularized logistic regression solver in LibLINEAR 2.20. A Clinical Predictive Model (deployed as, ChroMyL app) was implemented with Javascript scripting language and jQuery on Macromedia Dreamweaver 16.0 to enhance page interactivity. The findings provided better insight into the process of adopting empirical risk minimization techniques in machine learning algorithms to solve disease risk group stratification problems, thus revealing how machine learning algorithms can be applied to real-world problems. The outcome of this study would provide more insight into the theoretical foundations of ML, and the important factors that must be put into consideration in every predictive or stratification models. Future researches can focus more on determining the loss function of other machine learning algorithms used in stratifying the chronic myeloid leukemia disease. Also, the approach to the design of the clinical predictive model application called ChroMyL app could be used for related cases.

Keywords: *Classification, Clinical predictive model, Empirical risk minimization, Logistic regression, Machine learning, Stratification.*

[1] *Oluwabunmi Omobolanle Olaniyan

1. Introduction

Machine Learning (ML) was centered on biologically inspired models and the long term goal is to produce models and algorithms that can process information as well as biological systems. This encompasses many of the traditional areas of statistics with more focus on mathematical models [1, 2]. ML is now central to many areas of interest in Computer Science and related large-scale information processing domains, as such; ML has attained successes due to its strong theoretical foundations and its multidisciplinary approach by integrating aspects of Computer Science, Applied Mathematics, and Statistics, amid others. It is known to have three traditional learning approaches, namely supervised, unsupervised and semi-supervised, which have been explored in many areas due to data explosion [3, 4]. Diverse algorithms in supervised and unsupervised learning have been commonly used for mining data to construct models and make predictions in areas such as the educational system to predict students' academic success; email filtering to detect spam, network intruder detection to detect intrusion, credit-card fraud detection to detect fraud, prediction of the direction of stock market prices, and prediction of disease progression and risk group, among others [5, 6, 7]. ML has been employed in different Health Information Technology (HIT) systems where Clinical Predictive Models (CPM) systems hold a more significant promise for transforming healthcare [8]. Due to this development, tools that can aid effective stratification of Chronic Myeloid Leukemia (CML) risk group are required in a developing country like Nigeria [9, 10].

The successes of the diverse ML algorithms used for classification and prediction were measured based on the performance of the algorithm on a particular problem domain, which is supported by the core property of learning algorithms that is expressed through the "No free lunch" theorem of ML which states that: no given algorithm would have the best possible performance across all problem domains. Studies have shown that accuracy, speed, time to learn, number of features, comprehensibility, robustness, scalability, and interpretability are the benchmarks used for performance evaluation [11, 12] and with these, there are generalizations on the performance(s) of the chosen algorithm (or classifier) to be the "best" or "optimal" one for classification and prediction without considering loss function as a metric for algorithm selection. However, these benchmarks as mentioned earlier do not guarantee the successful adoption of an algorithm and have posed the question of what guarantees the choice of an optimal algorithm that would not result in a loss or cause a risk in prediction? Hence, another vital benchmark is the Empirical Risk Minimization (ERM) technique, which based its philosophy on the possibility of approximating the expectation of the loss functions of a given hypothesis using its empirical means [13] (Yuchen, 2016). This is a useful technique with which a good approximation of globally optimal classifiers can be obtained to provide useful classification [14] and the loss or risk function is determined for minimizing the risk of choosing the hypothesis of a learning algorithm. Utilization of scoring systems for stratification of Chronic Myeloid Leukemia (CML) disease into their risk group was used for cure and survival prolongation, but these systems do not computationally handle large datasets from clinical platforms and are faced with the limitations of transforming the data due to noise and overfitting of the data, which led to the challenge of accurate stratification and prediction. As a result, Machine Learning (ML) algorithms that understand, mimic, and aid the information processing tasks had been applied over the years in diverse areas to extract meaningful information and their performances evaluated. Therefore, the objective of this paper is to: (1) develop an Empirical Risk Minimization Data Stratification (ERMDS) algorithm; and (2) design and implement a clinical predictive model based on the algorithm.

2. Literature review

[15] theory of evolution describes learning as an adjustment to an environment. It was said that living organisms are not static but change and evolve constantly. This concept is brought into Machine Learning that learns from data instances to become adapted (training) and then reproducing what has been learned on other data instances (testing). Learning is considered as a parameter for intelligent machines whereby deep understanding help in decision taking in a more optimized form and efficient method, and it is paramount to the study of data instances for building machines with explicit programming [2]. Classification and prediction algorithms are such that imbibes learning and are used to solve problems [16].

2.1. Classification and prediction

Classification and prediction is a Machine Learning tack of predicting group membership for data instances, and it is known to be one of the most common data mining tasks [17, 18]. Classification identifies categorical (discrete, unordered) labels while prediction models continuous-valued functions. The classification process consists of a training set that is analyzed by a classification algorithm, and the classifier model is represented in the form of classification rules [19]. There are lots of classification approaches for mining knowledge from data such as divide-and-conquer, separate-and-conquer, and covering and statistical approaches [20]. The divide-and-conquer approach commence by selecting an attribute as a root node and then creating a branch for each possible level of that attribute. This would divide the training instances into subsets, one for each likely value of the attribute. The same process will be recurring until all instances that fall in one branch have the same classification, or the residual instances cannot be split any more. The separate-and-conquer approach begins by building up the rules in the greedy approach (one by one). After a rule is established, all instances enclosed by the rule will be deleted [21]. The same process is frequent until the best rule found has a high error rate. Statistical approaches such as Naïve Bayes use probabilistic measures, i.e., likelihood to classify test items while covering and statistical approaches selects each of the accessible classes in turn and looks for a method of covering most of the training objects to that class in order to produce maximum accuracy rules. Several algorithms have resulted from these approaches, such as decision trees, PART, RIPPER, and Prism. Majority of the research carried out on classification problems in data mining has been committed to single-label problems. A traditional classification problem can be defined as follows:

Let $D = \{d_1, d_2, \ldots d_k\}$, Domain of likely training instances

$Y = \{y_1, y_2, \ldots y_k\}$, List of class labels

$H = \{h_1, h_2, \ldots h_k\}$, Set of classifiers for Domain D

As $D \rightarrow Y$, each instance $d \in D$ is assigned a single class y that belongs to Y.

The goal is to discover a classifier hypothesis $h \in H$ that maximizes the probability that $h(d) = y$ for each test case (d, y). In multi-label problems, each instance $d \in D$ can be given multiple labels $y_1, y_2, \ldots, y_k$ for $y_i \in Y$, and is represented as a pair $(d, (y_1, y_2, \ldots, y_k))$ where $(y_1, y_2, \ldots, y_k)$ is a list of ranked class labels from y connected with the instance d in the training data [22].

2.2. Clinical predictive modeling for data stratification

Clinical Predictive Model (CPM) is the application of statistical, mathematical or algorithm-based models to predict an outcome, such as the risk level of a patient in a particular disease, response to treatment, survival or death rate, and potential cost or risk associated with managing a specific patient population [23]. CPM is typically urbanized by fitting a statistical model to existing data; the option of model to be fitted depends on

the nature of the endpoint, and frequent options are logistic regression (for a binary endpoint) and survival models (for a time-to-event endpoint). CPM uses various patient personality to estimate the probability of significant outcomes over a given period (prognostic models), or the probability of a precise diagnosis (diagnostic models). By providing these probabilities, they enable clinicians to personalize medical decisions for individual patients [24]. CPM forms the cornerstone of data stratification for determining patients' risk while undergoing persistent procedures, helping to direct both treatment allocation and the consent process. However, their performances call for testing large datasets independent of those in which the models were developed before they can be used in external populations [25]. CPM has three main practical uses, as discovered by [26]. One, it can be used at an individual patient level to relay risk and aid in the clinical decision-making method by stratifying patients into diverse treatment option groups or to establish whether further testing is warranted to reach an appropriate conclusion. Secondly, it can be used for planning healthcare services by predicting disease occurrence and future demand on services, or by exploring the effects of different local policy options; and thirdly, it can be used in the quality management of healthcare services, where clinical audit processes evaluate observed with expected outcomes, given appropriate adjustments for differences in case-mix.

CPM can generally be categorized into either regression-based or algorithm-based [27]. CPM studies from medicine and epidemiology predominately relate regression-based techniques such as Logistic Regression and the Cox model. In disparity, the studies from Health Informatics, Computer Science and Information System regularly emphasize algorithm-based models such as Decision Trees, Support Vector Machines, among others. Between the various modeling techniques, the Cox model is most frequently used. The popularity of the Cox model stems from its excellent utility for the critical task of time-to-event predictions in the contexts of accepting and assessing disease progression. Literature showed that ANN, SVM, Naïve Bayes, among others, are the commonly used algorithms for classifying and predicting diseases in the healthcare sector. These classifiers had shown their accuracy value of the correctly classified instances of different diseases, as shown in Table 1.

Table 1: Classification Algorithms for Prediction

S/N	Approach	Author	Year	Disease Area	Accuracy Value
1.	Artificial Neural Network	[28]	2006	Chronic Myeloid Leukemia	84.44%
		[29]	2009	Prostate cancer	89.75%
		[30]	2010	Leukemia	96.70%
		[31]	2011	Chronic Myeloid Leukemia	97.66%
		[32]	2014	Breast cancer	92.09%
		[33]	2016	Pancreatic Cancer	97.87%
2.	Naïve Bayes	[28]	2006	Chronic Myeloid Leukemia	78.56%
		[34]	2015	Leukemia	91.17%
3.	Decision Tree	[28]	2006	Chronic Myeloid Leukemia	75.73%
		[29]	2009	Prostate cancer survivability	78.20%
		[35]	2015	Leukemia	91.17%
4.	Support Vector Machine	[29]	2009	Prostate cancer survivability	91.67%
		[35]	2010	Chronic Myeloid Leukemia	60.78%
5.	Lazy Classifier	[34]	2015	Leukemia	82.36%
6.	Logistic Regression	[36]	2019	Heart Disease	98.76%

From this Table, it was observed that the applying ANN and Naïve Bayes for prediction of CML had a predictive value of 84.44% and 78.56% in 2006 [36], and 97.66% in 2011 [31]. It is therefore expected that at the end of this paper, the performance value of correctly classified instances of

BayesNet, Multilayered perceptron, Projective Adaptive Resonance Theory (PART), and Logistic Regression algorithms would be better than previous performances.

Literature on CPMs from numerous domains are summarized and characterized using six dimensions that are related to clinical predictive modeling: (1) author and year of study; (2) the predicting event; (3) prediction models used; (4) model evaluation; (5) performance metric used in the evaluation; and (6) Number of features used as shown in Table 2.

Table 2: Summary of six dimensions pertinent to clinical predictive modeling

S/N	Author (Year)	Predicting event	Prediction models used	Model evaluation	Performance metric used	Number of features used
1.	[37]	Hospitalization among hemodialysis patients	ARM, DT	CV	AC	26
2.	[38]	Risk of heart failure	SOR	CV	AUC	500
3.	[39]	Lung cancer one-year -survival	BN, NB	CV	AUC	9
4.	[40]	Risk of lung cancer	CM, LR	EVD	AUC	11
5.	[41]	Hospitalization among diabetes patients	CM	SF	AUC	79

[Key: AC = Accuracy; ARM = Association Rule Mining; AUC = Area Under the receiver operating characteristic Curve; BN = Bayesian Network; BR = Bootstrap Resample; CI = Concordance Index; CM = Cox Model; CV = Cross Validation; DT = Decision Trees; EVD = External Validation Data; LR = Logistic Regression; NA = Not Available in the paper; NB = Naïve Bayes; RF = Random Forest; SF = Strength of Fit (not based on a holdout set); SOR = Scalable Orthogonal Regression; SVM = Support Vector Machine, YI = Youden's Index.]

2.3. Chronic myeloid leukemia

Chronic Myeloid Leukemia (CML) is a type of leukemia described by the increased and unregulated growth of predominantly myeloid cells in the bone marrow and the buildup of these cells in the blood [42]. It is a cancer of the white blood cells characterized by the expansion of proliferating myeloid cell pool, especially in the bone marrow, spleen, and peripheral blood. The risk of getting CML increases with age as it occurs in the Caucasians from the median age of 65 to 75 years [43] and in the Africans from the median age of 36 years (Range, 13-75) Based on the differences in the median age of occurrence, the Nigerian patients have their prognosis at an early age when compared with the Caucasians [44]. Chronic Myeloid Leukemia is a disease with three phases, i.e., the Chronic-Phase (CP), Accelerated-phase (AP), and Blastic transformation Phase (BP) but emphatically, the interest of this study is the chronic phase of CML because approximately 90% of patients are diagnosed in this phase [45].

In predicting chronic myeloid leukemia-chronic phase (CML-CP), some scoring systems are used for risk stratification, but mainly three (3) of them are widely accepted to stratify the patient into low, intermediate or high-risk groups namely: Sokal, Hasford, and EUTOS (European Treatment and Outcome Study. These scoring systems were long utilized in CML disease prediction till present, and the outcomes are improving [46]. However, nevertheless, the procedure is still accompanied by a high rate of morbidity and mortality due to the long process of stratification, making the risk group selection a crucial issue. Hence, this informs the reason to employ machine learning techniques to improve the accuracy of stratification.

3. Methodology

This section presents the research methodology, which includes the design of the model and the Logistic regression algorithm used in the clinical predictive model for CML stratification. As a requirement for the development of the Clinical Predictive Model (CPM) to stratify CML, five (5) diagrams were used to highlight the design of the model, i.e. (I) architecture of the system, (II) use case diagram, (III) entity-relationship diagram (IV) sequence diagram and (V) activity diagram. Javascript scripting language and jQuery on Macromedia Dreamweaver 16.0 Integrated Development Environment (IDE) was used to design the interface. Javascript and jQuery were used because they are cross-platform programs that enhanced interactive interface pages and easier navigation. At the back end, XAMPP Server which houses MySQL and Apache was used for rules repository and system connection. The ChroMyL App was built using the V-model software development life cycle (SDLC). The V-model SDLC was used because it stresses the need for testing at each phase in the development cycle, and the next phase starts only after the completion of the preceding phase.

3.1. Design of the clinical predictive model

The following five diagrams are used to highlight the design of the CPM model to explain the processes in the system as mentioned earlier:

I. **The Architecture of the ChroMyL App:** This comprised of three parts: (1) the front end; (2) the connector; and (3) the rule repository, as shown in Figure 1.

a. **Front end:** The user interface was designed using Javascript scripting language and jQuery on Macromedia Dreamweaver 16.0. At this point, the input values were supplied to enable the prediction of the risk group and stratification of the data. Hence, decision-making takes place afterward.

b. **Connector:** XAMPP server (Apache) connects the front end to the rules repository where the classifier with minimal risk function was used. At the point where the user supplies the input values, the connector enables the front end to query the rule repository for appropriate rules and decision processes. Apache was used to provide a secure, efficient, and extensible server, which enables the desktop application to execute in an offline machine.

c. **Rule Repository:** XAMPP server (MySQL) database stores the rules generated from the optimal classifier with the minimal loss function during the evaluation phase. The rules generated emerge from the pattern obtained by the optimal model from the historic CML patients' data, and from the new patients' record that was not involved in the model building stage. The ChroMyL App is responsible for mapping the pattern in the rules generated with the new patients' data to predict the likelihood of a risk group.

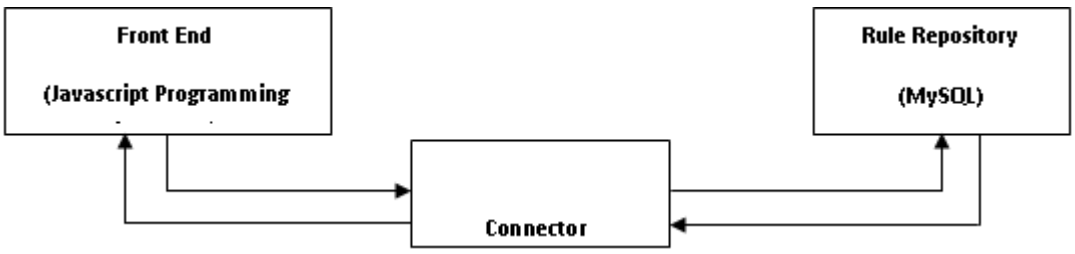

Figure 1. Architecture of the ChroMyL App

II. **Use case diagram:** At the launching of the system, the welcome screen is displayed to show brief information about the ChroMyL App, and a "Continue" button is

clicked to access the next window, which is the prediction window. At the clicking of the "Continue" button, the login page appears for entering username and password for authentication. After supplying the username and password, the "Log-In" button is clicked to submit the query. If it satisfies the conditions in the query, the user continues the process else it asks for the correct details. If the login is successful, the user (i.e., the Haematologist) is asked to supply the patient's card number to enter ChroMyL lab. If the card number is wrong, it indicates that the patient does not exist, otherwise if correct, it displays the patient details. Afterward, the user enters the Basophil count and Spleen size, and automatically, the risk score is predicted, and the group in which the patient belongs is shown. The administrator, on the other hand, logs in using username and password for authentication. After the login, the administrator is directed to "Add New Rule" and "Save Rule" in their windows. The Add New Rule window presents options to input new rules that are not present in the system, and the "Save Rule" update the database with the newly added rules as shown in Figure 2.

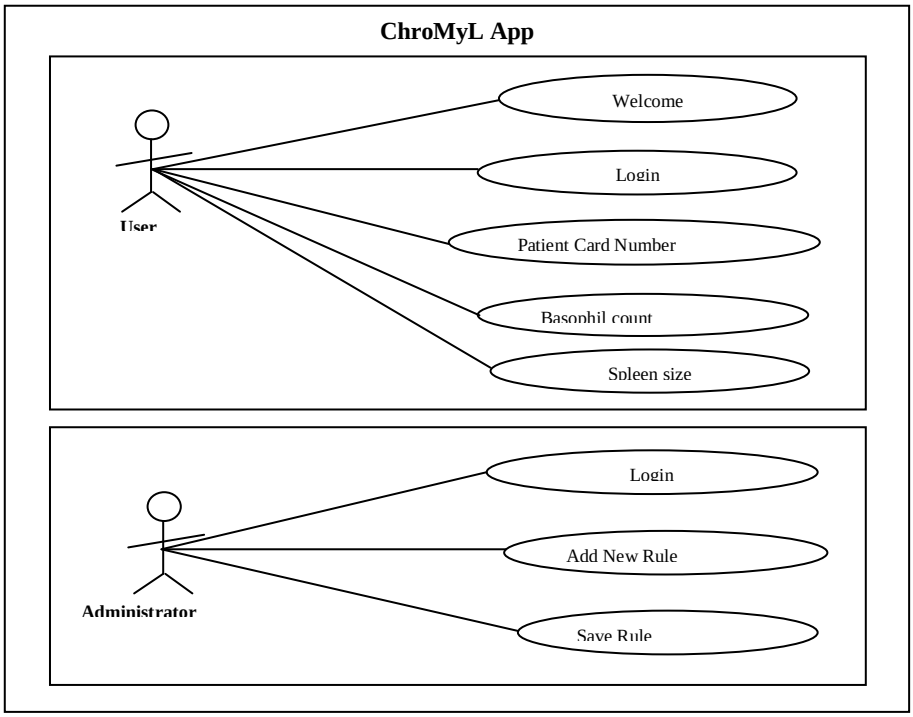

Figure 2. Use Case diagram for ChroMyL App

III. Entity-Relationship Diagram (ERD): This depicts the relationship that exists between the three entities as shown in Figure 3. The PatientT entity has a one-to-many relationship with the StaffT entity since each patient sees many Staff. The StaffT entity has a one-to-one relationship with the HaematologyT entity since it is only one staff that can carry out the test per time. In this case, the Hematologist is a Staff; therefore one Haematologist can access the system per time. Likewise, the PatientT entity has a compulsory one-to-many relationship with the HaematologyT entity represented with the straight line, showing that the Haematologist sees every patient that comes for a test at the lab. Also, there is another one-to-many relationship with the hematologist with the dotted line, which signifies that it is not compulsory to carry out the ChroMyl test but still can see the Haematologist.

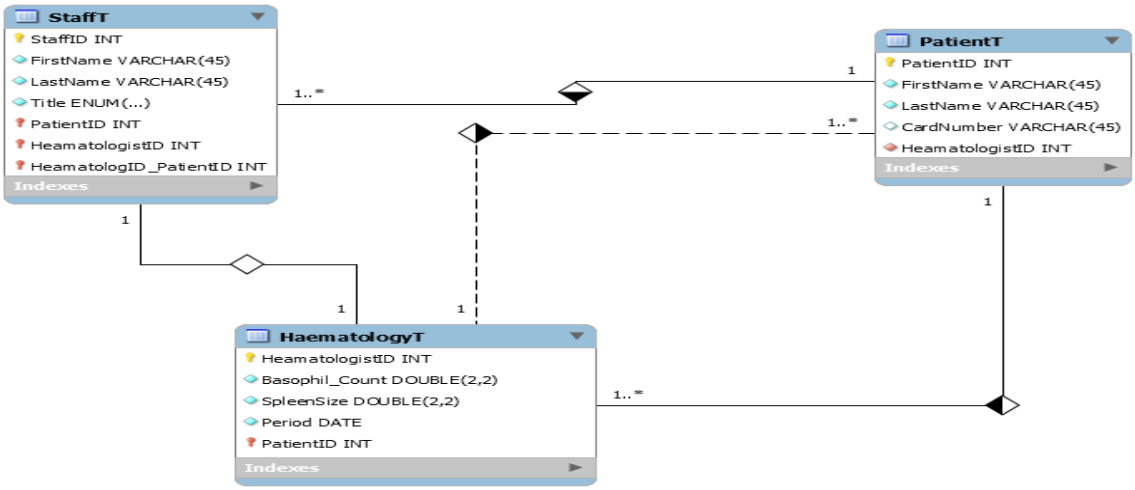

Figure 3. Entity-Relationship Diagram for ChroMyL database

IV. Sequence diagram: Here, the user request login, and supplies username and password in the authentication class; the PatientDetails class has the patient details especially the patient card number required for performing the risk score; the basophil count and the spleen size, and the RiskGroupStratifier class which has the high and low-risk group. The object "user" can then view patient details, get basophil count, get spleen size, and get the risk group.

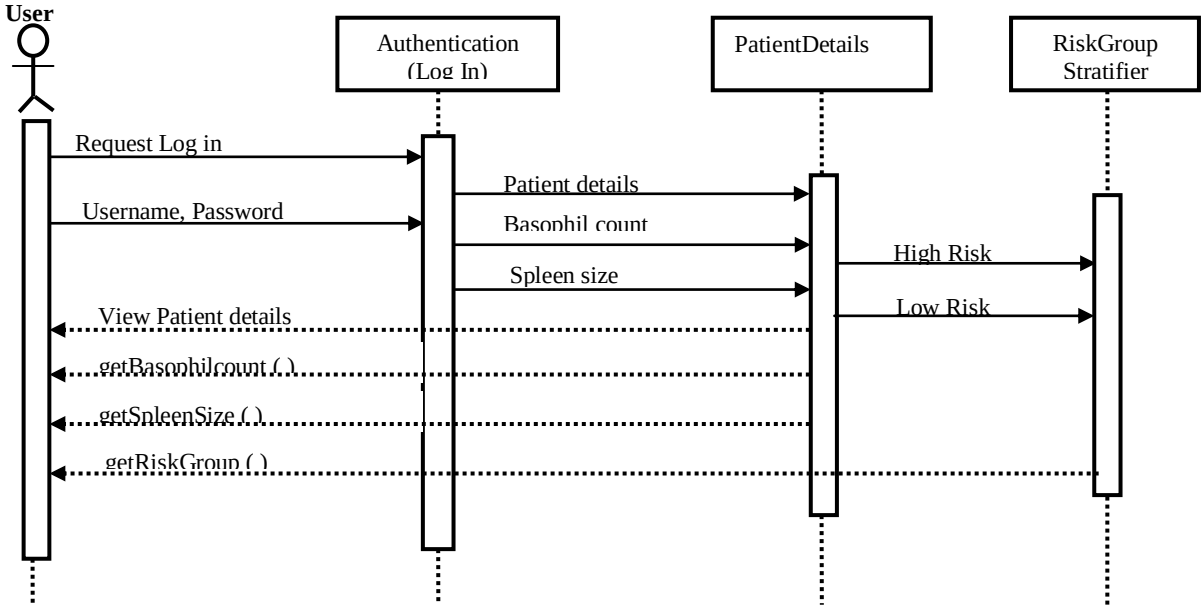

Figure 4. Sequence Diagram for ChroMyLApp

VI. Activity diagram: In the ChroMyL App, the Staff, Patient details, and the Haematologist databases can be retrieved from the Health Information System (HIS) that is hosted in the cloud. At first, the system prompts the user (Haematologist) who is a Staff to login. The login details (username and password) is derived from the StaffT database by extension, and this is validated for successful login. If the details are not correct, then the user returns back to submit the correct login details. When the login is successful, the system asks the Haematologist if the patient is a new patient or an existing one. If it is an existing user, it prompts for the patient card number, which must satisfy the domain constraint 2; otherwise, it prompts for the full name of the patient and the card number. If the card number satisfies constraint 2, it asks for the

basophil count and spleen size to determine the risk score. The moment either of the inputs is entered, the system stratifies and predicts the risk group to which the patient belongs. If the risk score is greater than 87, the patient is stratified to be in the high-risk group of chronic myeloid leukemia, else it belongs to the low-risk group. Then the process terminates as depicted in Figure 5.

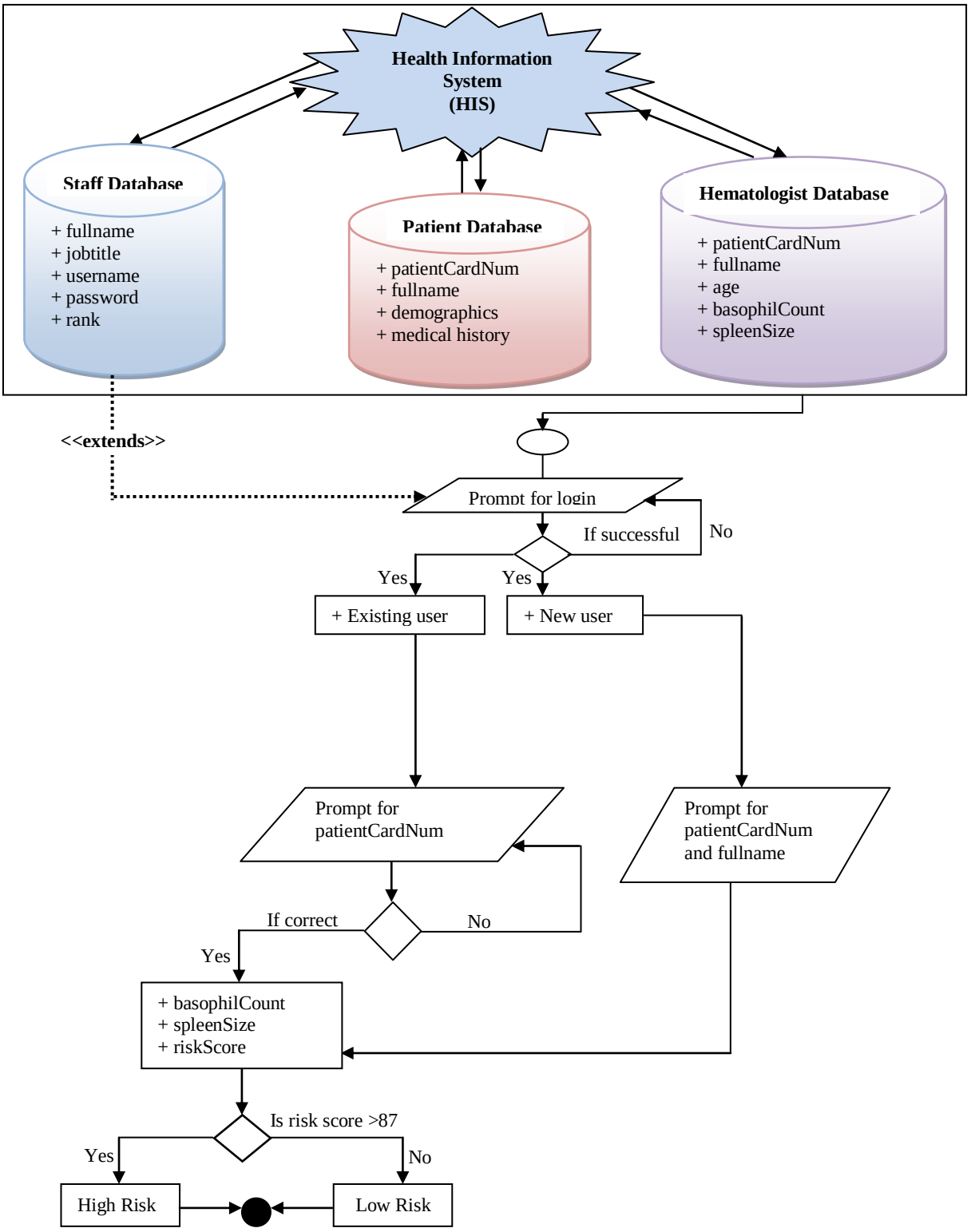

Figure 5. Activity Diagram for ChroMyL App

3.2. Pre-conditions and constraints for the clinical predictive model

The constraints are used to test the data that is entered, however, declaring an attribute of a specific domain acts as a restraint on the values it can take. In SQL, the "check" clause permits the schema design to specify a predicate that must be satisfied by any value allotted to a variable whose type is in the domain. The following pre-conditions and constraints were used:

a. Username and Password Validation

Create table [username] (
[username] VARCHAR (15) IS NULL OR NOT LIKE
CHECK ([username] VARCHAR ("*[!A-Z]* OR *[!a-z]* OR *[!0-9]*"))

b. Patient Number for existing user

Create table [PatientT] (
[PatientT] VARCHAR (45)
CHECK ([CardNumber] ALPHANUMERIC_Input mask ("[@@@@###.PatientT]")
Where @ = alpha
= numeric

c. Basophil count Number

Create table [HaematologyT] (
[HaematologyT] DOUBLE (2,2) IS NULL OR NOT LIKE
CHECK [Basophil_Count] NUMERIC _Input mask ("[##.##.HeamatologyT, ("*[!0-9.]"))
Where # = numeric
. = decimal point

d. Spleen size Number

Create table [HaematologyT] (
[HaematologyT] DOUBLE (2,2) IS NULL OR NOT LIKE
CHECK [SpleenSize] NUMERIC _Input mask ("[##.##.Chromyl, ("*[!0-9.]"))
Where # = numeric
. = decimal point)

3.3. Dataset description

Chronic Myeloid Leukemia dataset was collected from Obafemi Awolowo University Teaching Hospitals Complex (OAUTHC), Osun State, Nigeria, for training and testing of the model. It contained one thousand, six hundred and forty (1640) patients' data between the periods of 2003 and 2017. The input variables of Basophil (x_1) and Spleen size (x_2) were used as the training inputs to generate the risk score (r) as the output, which informed the grouping of the patients to either low risk or high-risk groups. The dataset was converted into Comma Separated Values (.csv) format, and a data repository that interfaces with the Waikato Environment for Knowledge Analysis (WEKA) was created for the data. The grouping of the variables is shown in Table 3. The risk group is the response variable, while other variables are predictors. Each variable is suitably categorized to accommodate all the available information.

Table 3. Description of variables

S/N	Variable Name	Variable format	Variable Type	Data Type
1.	Basophil count (x_1)	——	Continuous	Numeric
2.	Spleen size(x_2)	——	Continuous	Numeric
3.	EUTOS Score	——	Continuous	Numeric
4.	Risk Group (r)	Low Risk, High Risk	Categorical	Nominal

Haven collected the data, performance evaluation of four classification algorithms (BayesNet, Multilayered perceptron, PART and Logistic Regression) was done on WEKA using metrics such as: correctly classified instance, time to learn, kappa statistics, sensitivity and specificity metrics. Among the four algorithms, Logistic regression had the highest accuracy value of 99.82% as previously published in [47]. Hence, an empirical risk minimization data stratification algorithm was developed by using the L1-regularized logistic regression solver.

4. Implementation and results

This section provides the implementation details and the discussion of the results.

4.1. Empirical risk minimization data stratification algorithm

Since the algorithm with minimal risk is logistic regression, the Empirical Risk Minimization Data Stratification (ERMDS) algorithm was developed using LibLINEAR 2.20, an open-source library for linear classification and machine learning tool developed by Rong-En *et al*. (2008) which was modified in 2017. It supports two popular binary linear classifiers: Logistic Regression and linear Support Vector Machine, and provides easy-to-use command-line tools and library calls for users and developers. The Logistic Regression used the L1-regularized logistic regression solver to solve the empirical risk minimization problem. The algorithm was implemented using the NetBeans Integrated Development Environment (IDE) 7.0 environment to implement the interface and to call the L1-regularized logistic regression solver.

From this point, it was assumed that the CML dataset had been processed, and the learning data had the specification of two spaces: $X \equiv$ Input space and $R \equiv$ Output space. In the training set, the differentiating features are basophil count and spleen size (x_1, x_2) for 1640 patients to differentiate between high risk and low risk using the inputs. The CML dataset was set as an approximation drawn independently and identically distributed (i.i.d) from distribution $P(\mathcal{x}, \mathcal{r})$. The loss function was computed by the data points in the CML dataset called empirical risk. An efficient ERMDS algorithm was designed based on the algorithm with minimal loss function. From the learning model built, the hypothesis that minimized the empirical risk was determined by finding the delta (i.e., the percentage difference) between the predicted output $\hat{r}$ and the true output r from the data-points in the dataset selected. Since the two best algorithms are Logistic regression and PART, the approach for computing algorithm to find the hypothesis that minimizes the risk was described, and thus, informs the choice of selecting an algorithm to be optimal.
A D-dimensional data input vector x ∈ R was assumed where the goal of classification is to predict the target class $\hat{r} \in \{-1, 1\}$ for a given input x. Logistic regression classification defines a predictor function in a 0-1 loss problem, and the 0-1 loss problem is equivalent to finding a feasible loss function l with a minimal sum of losses as defined by [48] in equations 1 to 3.

$$f_w(x) = \sum_{j=1}^{D} w_j x_j + w_0 = w^T x + w_0 \quad (1)$$

where $w_j \in R$ and $w_0 \in R$ is a bias.

Then

$$\hat{r} = \begin{cases} 1, & fw(x) \geq 0 \\ -1, & fw(x) < 0 \end{cases} \quad (2)$$

The training dataset contains N data vectors $X = \{x_1, x_2, \ldots, x_N\}$, their corresponding target class $r = \{r_1, r_2, \ldots, r_N\}$ and the "margin of safety" m_i by which the prediction for x_i is accurate. To calculate the confidence of a class prediction for an observation $x_i \in X$, the margin is defined as $m_i(w) = t_i f_w(x_i)$. The margin $m_i(w) < 0$ indicates x_i is misclassified, while $m_i(w) \geq 0$ indicates x_i is correctly classified. The learning objective in

classification is to find the best (homogenous) w to minimize some loss over the training data (X, r), i.e.,

$$\boldsymbol{w^*} = argmin_w \sum_{i=1}^{N} L\ (m_i(w)) + \lambda R(w \quad (3)$$

where loss L($m_i(w)$) is defined as a function of the margin for each data point x_i, $R(w)$ is a regularizer which prevents overfitting (typically $\|w\|_2^2$), and $\lambda > 0$ is the regularization strength parameter.

Algorithm: Loss function computation

Input: Training data $D = \{b, s\}$, parameters: C, E, solvertype.
Set C=1 // Set the cost parameter C
Set E=0.1 // Set the epsilon parameter in loss function
Set solvertype.L1R_LR.getId( ) // L1-regularized logistic regression
Set TrueRisk r = {low, high} = {46.20%, 53.80%}
Output: L, R // Loss function, Risk group

Input Training data (***b***, s)
Output Weights r* minimizing 0-1 loss
1: **function** FIND-MIMIMAL 0-1-LOSS (x_1, x_2)
2: $\overline{w} \leftarrow w^*_{LR}$ from LR solution for (x_1, x_2)
3: $\bar{l} \leftarrow \overline{w}$
4: $loss_{min} \leftarrow \sum_{l=1}^{N} \bar{l}_i$ {Set initial bound to $L(\overline{w})$}
5: **return** w*
6: **function** TRUE-OUTPUT r
7: **if** (r of l are assigned) **then**
8: $\hat{r} \leftarrow$ LR solution for l
9: **function** LOSS-FUNCTION (L)
10: **if** ($\Delta(\hat{r} - r) \leq 1$ && ++ **then**
11: $loss_{min} \leftarrow loss$
12: **else**
13: $\sum_i l_1 (\Delta(\hat{r} - r) \leq 1$ && --
14: **end if**
15: **end if**
16: **end function**
17: **end function**
18: **end function**

4.2. System interface

The section describes the features of the developed ChroMyL App. The system was implemented in javascript and jQuery in an executable format. The application runs on a localhost server – XAMPP where MySQL is the database used for patients' data storage. The goal of App was to provide an environment that allows timely stratification and prediction of the CML patients' into their risk groups and can present the records for faster access and decision-making process.

As described in Figure 5 in section three, the ChroMyL App was designed such that the Staff database is accessed for logging in by extension. This implies that the Staff who can use this system is the Haematologist that already has a username and password created by the administrator. Based on this, the system prompts for login, and the details are submitted for authentication. At the process of authentication, the login query checks on the domain constraint 1 and validates it. After a successful login, the system asks if the user (in this case, the patient) is a new or existing user. If it is an existing user, it prompts for the patient card number, which must satisfy the domain constraint 2; otherwise it prompts for the full name of the patient and the card number. If the card number satisfies the second constraint, (for instance, a card number with sequence; BUTH102), it then

displays the patient detail such as name, gender, age, home address, mobile number, and the date and time last admitted. Afterward, the basophil count and spleen size text boxes become active for the input to be supplied to determine the risk score. The moment either of the inputs is entered, the system stratifies and predicts the risk group to which the patient belongs. If the risk score is greater than 87, the patient is stratified to be in the high-risk group of chronic myeloid leukemia; else it belongs to the low-risk group, as shown in Figures 6 to 14. Figure 6 showed the login page where the Haematologist enters his/her username and password to access the system and for authentication; Figure 7 showed the responsive login page that prompts that a field is required to be filled before logging in. Figure 8 showed the successful login of the Haematologist login, which signified the module to enter patient card number; Figure 9 showed that the patient card number entered is not recognized, and it is advised to check the number and try again. Figure 10 showed the patient card number textbox where BUTH102 is entered, and the patient details such as name, gender, address, age, ICE (mobile) number, and date and time last admitted are displayed. In Figure 11, the basophil count and spleen size textboxes are active for the numbers. When the basophil count or spleen size is entered, the risk score was pre-determined to give the risk score and risk group. In this case, the patient had a risk score of 89.36 and was predicted to be of high risk in chronic myeloid leukemia. Having completed the test process, the test result can be sent to the BUTH database, and another patient card number can be entered for a test, as shown in Figure 11. In Figure 12, the risk score is pre-determined to give a low-risk score and risk group. In this case, the patient had a risk score of 55.04 and was predicted to be of low risk in chronic myeloid leukemia. In situations where the values for the basophil count and the spleen size are not entered, the buttons for sending the test result to the BUTH database and for entering another patient card number would not be active, as shown in Figure 13. When the test process is completed, and the Haematologist is informed of the successful completion of the test on BUTH102 patient, the result is then sent to the Haematologist database, and the Haematologist is re-directed to perform another test as shown in Figure 14.

Figure 6. User Login page

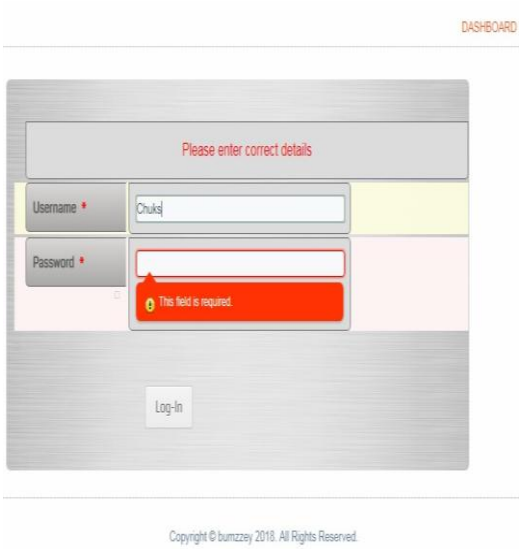

Figure 7. Responsive Login page

Figure 8. A successful login present existing PatientId field

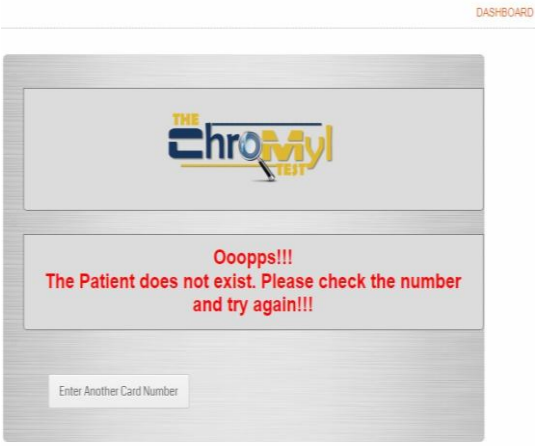

Figure 9. Unverified patient card number

Figure 10. Patient's details and activated ChroMyL test module

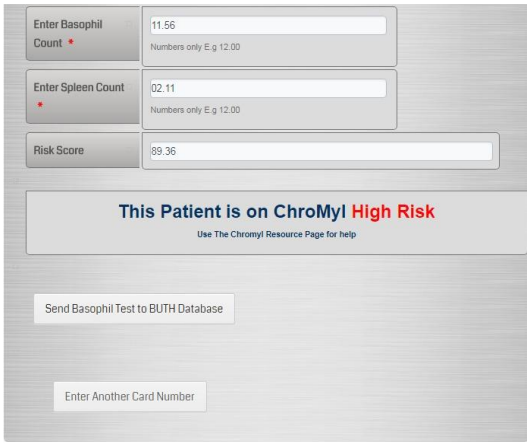

Figure 11. Basophil count and spleen size to determine the CML high risk

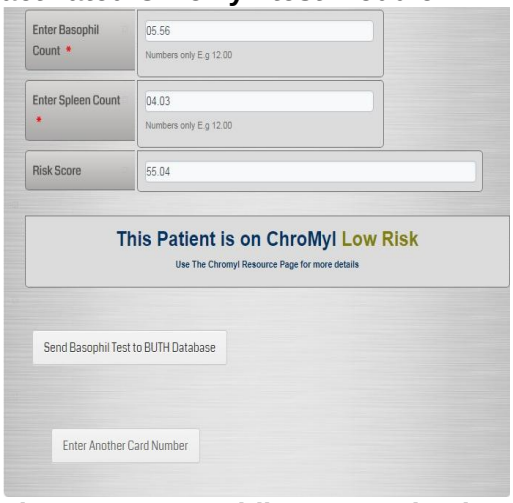

Figure 12. Basophil count and spleen Size to determine the CML low-risk group

Figure 13. Responsive test field and database button disabled

Figure 14. Completion page to redirect for another test

5. Conclusion

The concept of Empirical Risk Minimization is relevant in the world of supervised learning because the actual goal of supervised learning is to find a pattern that solves a problem as opposed to finding a model that best fits the given dataset. In this paper, an algorithm was presented to show how the loss function of a machine learning algorithm can be determined on a training dataset which contained N data vectors $X = \{x_1, x_2, \ldots, x_N\}$, and their corresponding target class $r = \{r_1, r_2, \ldots, r_N\}$, since the goal of classification was to predict the target class of predicted output $\hat{r}$. As such, this algorithm played a big role in producing optimum and faster results for accurate predictions. Also, a Clinical Predictive Model that can stratify CML patient's risk group using the basophil count and spleen size was developed. The developed application would assist Haematologists in proffering fast and easy stratification of patients' risk group, which could aid other diagnostic processes. The findings provided better insight into the process of adopting empirical risk minimization techniques in machine learning algorithms to solve disease risk group stratification problems, thus revealing how machine learning algorithms can be applied to real-world problems.

Acknowledgments

The authors appreciate the Department of Computer Science laboratory, Redeemer's University, Ede, Nigeria for providing a convenient environment for the analysis.

References

[1]. K. P. Atul, P. Prabhat and K. L. Jaiswal, "Classification Model for the Heart Disease Diagnosis", Global Journal of Medical Research Diseases, vol. 14, no. 1, **(2014)**, pp 1-7.
[2]. C. Raffaele, T. Marta, P. Giuseppina, P. Antonella and D. F. Fabio, "Artificial intelligence and machine learning applications in smart production: Progress, trends, and directions", Journal of Sustainability, vol. 12, no. 492, **(2020)**, pp. 1-26.
[3]. S. Shalev-Shwartz and S. Ben-David, "Understanding machine learning: From theory to algorithms (1st ed.)", Cambridge, USA, **(2014)**.
[4]. V. Mounica and B. Srikanth, "A comprehensive study of machine learning mechanisms on big data", International Journal of Recent Technology and Engineering (IJRTE), vol. 7, no. 6S2, **(2019)** pp. 773-779.
[5]. B. Gianluca, "Machine learning strategies for time series prediction", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 10 no. 6, **(2013)**, pp. 78-89.

[6]. T. Jie, L. Rong, Z. Yue-Li, L. Mou-Ze, H. Yong-Fang, S. Ming-Jie, ..., Z. Wei, "Application of machine-learning models to predict tacrolimus stable dose in renal transplant recipients", Statistical Methods in Medical Research, vol. 8 no. 4, **(2017)**, pp. 45-57.
[7]. K. Luckyson, S. Snehanshu and R. D. Sudeepa, "Predicting the direction of stock market prices using random forest", Applied Mathematical Finance, vol. 1 no. 1, **(2016)**, pp. 1-20.
[8]. H. C. Jonathan and M. A. Steven, "Machine Learning and prediction in medicine - Beyond the peak of inflated expectations". Machine Learning Informatics, vol. 20 no. 31, **(2017)**, pp. 2507-2509.
[9]. Z. Meng, L. Zhaoqi, Z. Xiang-Sun and W. Yong, "NCC-AUC: An AUC optimization method to identify multi-biomarker panel for cancer prognosis from genomic and clinical data. Bioinformatics, vol. 31, no. 20, **(2015)**, pp. 3330-3338.
[10]. Y. Safoora, A. Fatemeh, A. Mohamed, D. Coco, E. L. Joshua, S. Congzheng, … A. D. C. Lee, "Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. BioRxiv Journal, doi: http://dx.doi.org/10.1101/131367 **(2017).**
[11]. J. Han, M. Kamber and J. Pei, "Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann, **(2012).**
[12]. H. W. Ian, E. Frank and M. A. Hall, "Data mining-practical machine learning tools and techniques (3rd ed.)", Burlington, USA: Morgan Kaufmann – Elsevier, **(2011)**.
[13]. Z. Yuchen, "Distributed Machine Learning With Communication Constraints (A Published Doctoral Thesis)", California, Berkeley, **(2016)**.
[14]. M. Aryan, "Efficient methods for large-scale empirical risk minimization (A published Doctoral thesis)", Philadelphia, Pennsylvania, **(2017)**.
[15]. C. R. Darwin, "On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life (2nd ed.)", London, UK: John Murray, **(1859)**.
[16]. A. Botchkarev, "A new typology design of performance metrics to measure errors in machine learning regression algorithms. Interdisciplinary Journal of Information, Knowledge, and Management, vol. 14, **(2019)**, pp. 45-79. https://doi.org/10.28945/4184.
[17]. X. Yin and J. Han, "CPAR: Classification based on predictive association rule", In SDM2003, San Francisco, CA, **(2003)**.
[18]. W. G. Stuart, S. C. Gary and A. M. N. Samer, "Statistical Primer: Developing and validating a risk prediction model", European Journal of Cardio-Thoracic Surgery, vol. 54, no. 2, **(2018)**, 203-208.
[19]. M. Thangaraj, & M. Sivakami, "Text classification techniques: A literature review", Interdisciplinary Journal of Information, Knowledge, and Management, vol. 13, **(2018)**, pp. 117-135. https://doi.org/10.28945/4066.
[20]. T. S. Lim, W. Y. Loh and Y. S. Shih, A comparison of prediction accuracy, complexity and training time of thirty-three old and new classification algorithms. Machine Learning, vol. 3, no. 9, **(2012)**, 25-32.
[21]. E. Frank and I. Witten, "Generating Accurate Rule Sets Without Global Optimization. Proceedings of the Fifteenth International Conference, **(2011)**, (pp. 144-151). Madison, San Francisco.
[22]. M. Mehryar, R. Afshin and T. Ameet, "Foundations of Machine Learning (2nd ed.)", The MIT Press Cambridge, Massachusetts Lodon, England **(2012)**.
[23]. C. Evangelia, M. Jie, S. C. Gary, W. S. Ewout, Y. V. Jan and V. C. Ben, "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models", Journal of Clinical Epidemiology, vol. 110, **(2019)**, pp. 12-22.
[24]. P. Carter, "Big data analytics: Future architectures skills and roadmaps for the CIO international data corporation. Proceedings of the 6th International Conference, **(2011)**, (pp. 123-138), Madison, San Francisco.
[25]. R. D. Riley, J. Ensor and K. I. E. Snell, "External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: Opportunities and challenges", BMJ, 353:i3140, **(2016)**.
[26]. S. Ting-Li, J. Thomas, L. H. Graeme, B. Iain and S. Matthew, "A review of statistical updating methods for clinical prediction models", Statistical Methods in Medical Research, vol. 10, no. 1, **(2016)**, pp. 1-16.
[27]. L. Breiman, Statistical modeling: The two cultures. Statistical Science, 16(3), **(2011)**, pp. 199-231.
[28]. P. Andreeva, "Data modelling and specific rule generation via data mining techniques. Proceedings of International Conference on Computer Systems and Technologies – CompSysTech", **(2006)**, (pp. 104-118). Wellington, New Zealand.

[29]. D. Dursun, "Analysis of cancer data: A data mining approach", Journal of Knowledge Engineering, vol. 26, no. 1, **(2009)**, pp. 100-112.
[30]. A. Saeid, A. Fahimeh, V. T. Fereshte, Z. S. Mahin and T. Kobra, "Recognition and prediction of leukemia with artificial neural network (ANN)", MIS Quarterly, vol. 37, no. 2, **(2010)**, pp. 407-424.
[31]. P. Dey, A. Lamba, S. Kumari and N. Marwaha, "Application of an artificial neural network in the prognosis of chronic myeloid leukemia", Anal Quant Cytol Histol, vol. 33, no. 6, **(2011)**, pp. 335-339.
[32]. K. S. Ishwinder, N. Meera, P. A. Ravindra and S. S. Sardul, "Diagnosis of cancer using artificial neural network and cloud computing approach", World Journal of Pharmacy and Pharmaceutical Sciences, vol. 3, no. 6, **(2014)**, pp. 1533-1548.
[33]. M. U. Sanoob, M. Anand, K. R. Ajesh and M. V. Surekha, "Artificial neural network for diagnosis of Pancreatic cancer", International Journal on Cybernetics & Informatics (IJCI), vol. 5, no. 2, **(2016)**, pp. 41-49.
[34]. M. Durairaj and R. Deepika, "Comparative analysis of classification algorithms for the prediction of leukemia cancer", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 5, no. 8, **(2015)**, pp. 787-791.
[35]. E. W. Kangwanariyakul, C. Nantasenamat, T. Tantimongcolwat and T. Naenna, "Data mining of magneto cardiograms for prediction of ischemic heart disease". International Journal of Computer Science Engineering and Information Technology Research, vol. 15, no. 12, **(2010)**, Pp. 120-128.
[36]. E. S. Maren, A. D. Stephen, F. Farhood and L. G. Eric, "Developing prediction models for clinical use using logistic regression: an overview", Journal of Thoracic Disease, vol. 11, no. 4, **(2019)**, pp. S574–S584.
[37]. J. Y. Yeh, T. H. Wu and C. W. Tsao, "Using data mining techniques to predict hospitalization of hemodialysis patients", Decision Support Systems, vol. 50, no.2, **(2011)**, pp. 439-448.
[38]. J. Sun, J. Hu, D. Luo, M. Markatou, F. Wang, S. Edabollahi, …., W. F. Stewart, "Combining knowledge and data driven insights for identifying risk factors using electronic health records", In AMIA Annual Symposium Proceedings **(2012)**, (pp. 442-451). Chicago, IL.
[39]. M. B. Sesen, T. Kadir, R. B. Alcantara, J. Fox and M. Brady, "Survival prediction and treatment recommendation with bayesian techniques in Lung cancer", AMIA Annual Symposium Proceedings, **(2012)**, (838–847). Wellington, New Zealand.
[40]. M. C. Tammemägi, H. A. Katki, W. G. Hocking, T. R. Church, N. Caporaso, P. A. Kvale, … C. D. Berg, "Selection criteria for Lung cancer screening", New England Journal of Medicine, vol. 368, no. 8, **(2013)**, pp. 728-736.
[41]. Y. K. Lin, H. Chen, R. A. Brown, S. H. Li and H. J. Yang, Time-to-event predictive modeling for chronic conditions using electronic health records. IEEE Intelligent Systems, vol. 29, no. 3, **(2014)**, pp. 14–20.
[42]. E. C. Besa, B. Buehler, M. Markman and R. A. Sacher, "Chronic myelogenous leukemia", Krishnan (3rd ed.), **(2013)**, Waterloo, Canada.
[43]. P. Eric, K. Rami and F. L. Alan, "The clinical management of chronic myelomonocytic leukemia", Journal of Clinical Advances in Hematology and Oncology, vol. 12, no. 3, **(2014)**, pp. 172-178.
[44]. A. A. Oyekunle, P. O. Osho, J. C. Aneke, L. Salawu and M. A. Durosinmi, "The predictive value of the Sokal and Hasford scoring systems in chronic myeloid leukaemia in the imatinib era", Journal of Hematological Malignancies, vol. 2, no. 2, **(2012)**, pp. 25-32.
[45]. J. Hasford, M. Baccarani, V. Hoffmann, J. Guilhot and S. Saussele, "Predicting complete cytogenetic response and subsequent progression free survival in 2060 patients with CML on imatinib treatment: The EUTOS score", Blood, vol. 118, **(2011)**, pp. 686-692.
[46]. R. Shouval, O. Bondi, H. Mishan, A. Shimoni, R. Unger and A. Nagler, "Application of machine learning algorithms for clinical predictive modeling: A data-mining approach in SCT". Bone Marrow Transplantation, vol. 49, **(2014)**, pp. 332-337.
[47]. O. O. Taiwo, F. A. Kasali, I. O. Akinyemi, S. O. Kuyoro, O. Awodele, D. D. Ogbaro and T. S. Olaniyan, "Stratification of chronic myeloid leukemia cancer dataset into risk groups using four machine learning algorithms with minimal loss function", African Journal of Management Information System, vol. 1, no. 2, **(2019)**, pp. 1-18.
[48]. B. Matthew, "Advances in empirical risk minimization for image analysis and pattern recognition", **(2014)**, Retrieved from https://tel.archives-ouvertes.fr/tel-01086088

Authors

O. O. Olaniyan, Dr. O. O. Olaniyan is a Lecturer in with the Department of Computer Science, Redeemer's University, Ede, Nigeria. She is a member of the Nigeria Computer Society. She has published widely in both local and international journals. Her research interests are in the area of Artificial Intelligence: data mining, machine learning, big data and health informatics.

A. O. Ogunde, Dr. A.O. Ogunde is a Reader in Computer Science. He is a registered member of the Computer Professionals Registration Council of Nigeria, Nigeria Computer Society, IAENG and several other professional bodies. He has published widely in both local and international journals. His research interests are in the area of Artificial Intelligence: data mining, machine learning, big data with other intelligent and knowledge-based systems.

T. A. Olowookere, Dr. T. A. Olowookere is a Lecturer with the Department of Computer Science, Redeemer's University, Ede, Nigeria. His research interests lie within the areas of Machine Learning, Data Science and Process Mining. He has Publications in reputable peer-reviewed Journals. He is a member of IEEE Computer Society and Nigeria Computer Society.

I. S. Oyetade, Mr. Oyetade is a Senior Technologist at the Department of Computer Science, Redeemer's University, Ede, Osun State. His research interests are in the areas of Artificial Intelligence, Machine Learning, and Data Science. He is a member of Nigeria Computer Society.