Stratification of Chronic Myeloid Leukemia Cancer Dataset into Risk Groups using Four Machine Learning Algorithms with Minimal Loss Function

Article · April 2019

CITATION
1

READS
101

7 authors, including:

Oluwabunmi Omobolanle Olaniyan
Redeemer's University Ede Osun State Nigeria
18 PUBLICATIONS **12** CITATIONS
SEE PROFILE

Funmilayo Kasali
Mountain Top University, Nigeria
21 PUBLICATIONS **11** CITATIONS
SEE PROFILE

Ibidapo Olawole Akinyemi
Mountain Top University, Ibafo, Ogun State, Nigeria
12 PUBLICATIONS **14** CITATIONS
SEE PROFILE

Shade O Kuyoro
Babcock University
53 PUBLICATIONS **451** CITATIONS
SEE PROFILE

Vol. 1, Issue 2, April 2019, pp. 1 - 18
Taiwo, Kasali, Akinyemi, Kuyoro, Awodele, Ogbaro & Olaniyan (2019). Stratification of Chronic Myeloid Cancer Dataset into Risk Groups using Four Machine Learning Algorithms

Stratification of Chronic Myeloid Leukemia Cancer Dataset into Risk Groups using Four Machine Learning Algorithms with Minimal Loss Function

O. O. Taiwo[1], F. A. Kasali[2], I. O. Akinyemi[3], S. O. Kuyoro[4], O. Awodele[5], D. D. Ogbaro[6] and T. S. Olaniyan[7]

[1,2,3]Department of Computer Science and Mathematics, Mountain Top University, Ibafo, Ogun State, Nigeria.

[4,5]Department of Computer Science, Babcock University, Ilishan-Remo, Ogun State, Nigeria.

[6]Department of Haematology and Blood Transfusion, Olabisi Onabanjo University, Ago-Iwoye, Ogun State, Nigeria.

[7]Department of Industrial Relation and Personnel Management, Osun State University, Osogbo, Nigeria.

Email: oltaiwo, fkasali, ioakinyemi{@mtu.edu.ng}, afolashadeng@gmail.com, delealways@yahoo.com, danielsond2002@yahoo.com, toyinlaniyan@gmail.com

ABSTRACT

Chronic Myeloid Leukemia (CML) had been stratified into risk groups using scoring systems but these systems have limitation of overfitting data. Machine Learning (ML) algorithms were used to extract meaningful information from the datasets, but the loss function (empirical risk) of the algorithms was not considered to determine the risk that was incurred in adopting the algorithms for stratification. In this paper, secondary dataset of 1640 CML patients, between 2003 and 2017 was collected from Obafemi Awolowo University Teaching Hospitals Complex, Ile-Ife, Osun Sate, Nigeria. An experimental analysis was performed in Waikato Environment for Knowledge Analysis 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) to determine low and high risk patients. Holdout and 10-fold cross-validation techniques were used to evaluate the performance of the algorithms on correctly classified instances, time to learn, kappa statistics, sensitivity and specificity. Considering the performance metrics, Logistic regression and PART algorithms were the two algorithms with better performance in stratifying patients' risk group as against other algorithms used in this study. Afterwards, the loss functions of the two algorithms were determined by finding the difference between the true output r and the predicted output $\hat{r}$. The results of the loss function of Logistic regression algorithm for low and high risk in holdout and 10-fold cross-validation showed 0.22%, 1.40% and -0.22%, -0.02% respectively. Similarly, PART algorithm yielded -1.58%, 1.40% and -0.22%, -0.26%. From the findings, the Logistic regression algorithm had the minimum non-negative loss function in holdout technique and was used in the developed model to stratify CML into their risk groups. Therefore, the determination of loss function of algorithms minimizes the empirical risk and as such plays a significant role in producing optimum and faster results for accurate stratification.

Keywords: *Classification algorithm, Data stratification, Empirical risk minimization, Loss function, Machine learning*

Reference Format:

Taiwo, O. O., Kasali, F. A., Akinyemi, I. O., Kuyoro, S. O., Awodele, O. Ogbaro, D. D. And Olaniyan, T. S. (2019), Stratification of Chronic Myeloid Leukemia Cancer Dataset into Risk Groups using Four Machine Learning Algorithms with Minimal Loss Function, *Afr. J. MIS*, Vol.1, Issue 2, pp. 1 - 18.

1. INTRODUCTION

The field of Machine Learning (ML) has been employed in different Health Information Technology (HIT) systems where Clinical Predictive Models (CPM) systems hold a greater promise for transforming healthcare. These CPMs have been developed with the aid of known algorithms such as Artificial Neural Network (ANN), Support Vector Machine (SVM), Reinforcement Learning (RL), Decision Trees (DT), k Nearest Neighbour (kNN) and Gaussian Process, to predict the survival, progression, mortality and risk group of both the acute and chronic diseases (Atul, Prabhat & Jaiswal, 2014; Meng, Zhaoqi, Xiang-Sun & Yong, 2015; Stylianou, Akbarov, Kontopantelis, Buchan & Dunn, 2015; Safoora, Fatemeh, Mohamed, Coco, Joshua, Congzheng & Lee, 2017). CPMs have been seen to be successful in its implementation due to ML methods that were integrated in the computer-based systems in the healthcare environment; and as such provide opportunities to facilitate and enhance the work of medical experts, and ultimately improve the efficiency and quality of medical care (Jonathan & Steven, 2017).

The World Health Organization (WHO) reported that there was a high growth rate and prevalence of patients living with chronic conditions in both the developing and developed countries (World Health Organization [WHO], 2016); and in Nigeria there are about 10,000 cancer deaths recorded annually while 250,000 new cases are recorded yearly. Due to the increase in cancer statistics yearly, WHO launched a global action plan in 2013 for the prevention and control of non-communicable diseases for year 2013 to 2020, which aims at reducing premature mortality from cancer, cardiovascular diseases, diabetes and chronic respiratory diseases by 25% on or before year 2025. WHO and International Agency for Research on Cancer (IARC) collaborated with other United Nations (UN) organizations to develop standards and tools to guide the planning and implementation of interventions for prevention, early diagnosis, screening, treatment and palliative and survivorship care; and to provide technical assistance for rapid, effective transfer of best practice interventions to less-developed countries (WHO, 2017). However, it is worrisome that only 17 percent of African countries are said to have sufficiently funded cancer control programmes while less than half of all countries in the world have functional plans to prevent the disease and provide treatment and care to patients. Due to this course, a model that can aid effective stratification of Chronic Myeloid Leukemia (CML) risk group is required. Some stratification models had been developed to stratify related diseases into risk groups; nevertheless, the loss functions of the algorithms used were not considered to minimize the empirical risk. Therefore this paper developed a model that uses the algorithm with minimal loss function to stratify CML into high or low risk group.

2. LITERATURE REVIEW

The field of Machine Learning was centered on biologically inspired models and the long term goals is to produce models and algorithms that can process information as well as biological systems; and encompasses many of the traditional areas of statistics with more focus on mathematical models (David, 2012). Machine learning is now central to many areas of interest in Computer Science and related large-scale information processing domains. ML is broadly defined as computational methods using experience to improve performance or to make accurate predictions; experience refers to the past information available to the learner, which typically takes the form of electronic data collected and made available for analysis. ML entails data-driven methods capable of mimicking, understanding and aiding human and biological information processing tasks; and is closely related with Artificial Intelligence (AI), with ML placing more emphasis on using data to drive and adapt the model from large datasets (Ian & Eibe, 2005). The motivation in ML is majorly to produce an algorithm that can either mimic or enhance human/biological performance (Sepp, 2013). Machine Learning has been successfully deployed in variety of applications areas such as: morphological analysis in natural language processing, Optical Character Recognition (OCR), text or document classification, statistical parsing, named-entity recognition, medical diagnosis, speech recognition, speech synthesis, speaker verification, image recognition, fraud detection, network intrusion, robots, navigation, recommendation systems, search engines, information extraction systems, games, and many more (Mehryar, Afshin & Ameet, 2012). Hina, Syed and Harleen (2018) carried out a comparative survey of machine learning and meta-heuristic optimaization algorithms for sustainable and smart healthcare. The paper reviewed machine learning and its various optimization techniques in disease datasets which would help to avoid any kind of epidemic in algorithm selection. Alongside, an optimized sustainable healthcare framework that can use machine learning techniques and nature-inspired optimization

algorithm was designed to be used in healthcare dataset. In machine learning, more complex models can be searched because the task is more focused to learning only one or few carefully defined models which predict the variable in question. In another paper by Oladejo, Oladele and Saheed (2018), two dimensionality reduction strategies (feature selection and feature extraction) were used to address the problems of highly correlated data and to obtain a robust and efficient dimensional space. Analysis of micro array data was carried out on Leukemia cancer dataset with the goal of finding the smallest quality subsets for precise tumor arrangement. One-way ANOVA algorithm was used to select relevant variables and Principal component analysis (PCA) algorithm was used to remove the most relevant variables out of the ones that were selected. The classification algorithms employed were support vector machine (SVM) and K Nearest Neighbour (KNN) and experimental analysis was performed in matlabR2015a (8.5.0.197613) environment. The result of performance metrics in terms of accuracy attained 90% of SVM and 81.67% of KNN algorithm.

2.1 Empirical Risk Minimization (ERM) Technique and Function

Empirical Risk Minimization (ERM) is a theory in statistical learning that defines a family of learning algorithms and is used to give theoretical bounds on the performance of learning algorithms. It is a natural choice for a learning algorithm that helps to determine a good classification and regression learning function from a bad one (Barnabas, 2012); and it is a common and useful technique with which good approximation of globally optimal classifier can be obtained to give good statistical classification result. ERM is mostly used in determining the loss or risk function in supervised learning problems, and the major interest is to minimize the risk of choosing a hypothesis of a learning algorithm (Liyang, 2016).

The ERM can be computed when the distribution $p(x,y)$ is known to the learning algorithm, and by averaging the loss function on the training set. Considering the situation in which the hypothesis h^* among a fixed class of function $\mathcal{H}$ for which the risk $R(h)$ is minimal. The risk in this hypothesis is to be minimized using the equations 1 to 3 as defined by Vapnik (2000):

$$h^* = \arg min_{h \epsilon H} R(h) \quad (1)$$

In order to minimize the risk, let X and Y be the learning function: $h: X \rightarrow Y$

Training set = $(x_1, y_1), \ldots (x_m, y_m)$ where $x_i \epsilon X$ is an input and $y_i \epsilon Y$ is the corresponding response (output) to give $h(x_i)$. Assuming there is a probability distribution $P(x, y)$ over x and y, and the training set consist of m instances $(x_1, y_1), \ldots (x_m, y_m)$ drawn independently and identically distributed (i.i.d) from distribution $P(x, y)$. This assumption allows the model of uncertainty in predictions. The loss function $L(\hat{y}, y)$ is required to measure the difference between the predicting $\hat{y}$ of a hypothesis and the expected or true outcome y (Ji, Jiang, Wang, Xiong & Ohno-Machado, 2014).

The risk associated with the hypothesis $h(x)$ is the expectation of the loss function:

$$R(h) = \mathrm{E}[L(h(x), y)] = \int L(h(x), y) dP(X, Y) \quad (2)$$

In this case, the learning algorithm chosen for prediction finds the hypothesis h^* among a fixed class of function $\mathbb{H}$ for which the risk $R(h)$ is minimal.

Empirical Risk Minimization Function

The ERM function is computed when the distribution $p(x,y)$ is known to the learning algorithm, and by averaging the loss function on the training set. It is an approximation that replaces $R(h)$. The empirical risk is introduced as:

$$R_{emp}(h) = \frac{1}{m} \sum_{i=1}^{m} L(\mathrm{h}(x_i), y_i) \quad (3)$$

Hence, the principles' interest is to choose a hypothesis $\hat{h}$ that minimizes the empirical risk

The ERM function is important in evaluating the performance of the function $R(h)$ by using non-negative Real valued loss function $L(\hat{y}, y)$, which measures how different the prediction $\hat{y}$ is from the true outcome y. ERM can also be used to compute M-estimators (Chaudhuri, Sarwate & Sinha, 2013) which is obtained as the minima of sums of functions of the data. A regularization term $R(\cdot)$ on R_{emp} can be used to prevent overfitting to give regularized ERM. The regularization term is seen as stabilizer of learning algorithm and it explains the phenomenon that changing a data point in the training set does not affect the performance of output classifier too much. This indicates how to control the trade-off between empirical risk and the difference between the true and empirical risk. Lagrange duality indicates that when we want to find linear classifier f that minimizes ERM with bounded norm $\| f \| \leq C$ for some constant C, we can find f *by* minimizing the regularized ERM for a suitable choice of Lagrange coefficient λ (Mahdavi, et al., 2014; Poline, et al., 2012).

2.2 Chronic Myeloid Leukemia

Chronic Myeloid Leukemia (CML) is a type of leukemia characterized by the increased and unregulated growth of predominantly myeloid cells in the bone marrow and the accumulation of these cells in the blood (Besa, et al., 2013). It is a cancer of the white blood cells characterized by 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 (Eric, et al., 2014) 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 (Oyekunle et al., 2012). 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 (Hasford, et al., 2011).

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) as described in Table 2.1. These scoring systems were long utilized in CML disease prediction till present and the outcomes are improving (Shouval, et al., 2014), but nevertheless, the procedure is still accompanied by high rate of morbidity and mortality due to the longer process of stratification, making the risk group selection a crucial issue. Hence, this calls for the reason to employ machine learning technique to improve the accuracy of stratification. Table 1 is the tradition approach that has been used to define the risk group of a patient based on the input variables that are used for prognosis.

3. METHODOLOGY

A model that can be used for better stratification of CML dataset into their risk groups effectively was developed using the algorithm with minimal loss function. Hold-out (66%) and 10-fold cross-validation evaluation techniques were used to evaluate the performance of the four classification algorithms (supervised learning) to choose the best two algorithms. Hold-out was used on all the data points that are i.i.d (independently and identically distributed) because it is computationally easier to program and cross-validation was used to generate training and validation sets for the hyper-parameter tuning.

3.1 Evaluation of Classification Algorithm Performance

In evaluating the performance of the classification algorithms, the model was built in WEKA 3.8.0 using the hold-out (66% training data) and 10-fold cross-validation evaluation methods on BayesNet, Multilayered Perceptron, PART and Logistic Regression algorithms to train and test the classifiers. After the training process, the values of correctly classified instances, time taken to learn, kappa statistics, sensitivity and specificity were computed to compare their performances in which two algorithms with leading performances were chosen. The value of the Correctly Classified Instances (CCI) was determined by the percentage of correctness of outcome among the test sets, and that compares how close a new test value is to a value predicted by the rules. CCI was determined by dividing the sum of True Positive (TP) and True Negative (TN) values by the sum of TP, TN, False Positive (FP) and False Negative (FN) values multiplied by 100% as defined by Frank and Witten (2011) as shown in equation 4, 5 and 6.

$$\text{CCI} = \frac{TP + TN}{(TP + FP + TN + FN)} \text{ X } 100\% \quad (4)$$

Sensitivity was determined by dividing TP value by the sum of TP and FN values multiplied by 100% as shown in Equation 6. It measures the ability of a test to be positive when the condition is actually present.

$$\text{Sensitivity} = \frac{TP}{(TP + FN)} \text{ X } 100\% \quad (5)$$

Specificity was determined by dividing TN value by the sum of FP and TN values multiplied by 100% as shown in Equation 6. It measures the ability of a test to be negative when the condition is actually not present.

$$\text{Specificity} = \frac{TN}{(FP + TN)} \text{ X } 100\% \quad (6)$$

3.1.1 Model Building

In the machine learning field, there are three basic phases in model building: pre-processing, processing and post-processing in setting up a model. The pre-processing phase involves feature transformation and extraction, the processing phase involves model generation and tuning based on the chosen algorithms; and post-processing phase involves knowledge representation. Waikato Environment for Knowledge Analysis (WEKA) 3.8.0 was

used to build and evaluate the performance of the selected classification algorithms. WEKA was chosen because it is an open source machine learning package developed in Java that contains many machine learning algorithms including Bayesian classifiers, Functions, Lazy classifiers, Rules, Trees and Miscellaneous classifiers for data preprocessing, classification, regression, clustering, association rules and visualization. For the purpose of this study, four classification algorithms in machine learning such as Bayes (BayesNet), Functions (Multilayered Perceptron and Logistic Regression), and Rule (PART) were implemented on Explorer application of WEKA 3.8.0 for easy and fair comparison of each algorithm. These classifiers were selected because they have been used in several researches, established as good classifiers, are "white box" classification models that provide explanation for the classification, and can be used directly for decision making. Since these are from different classifier families, they yielded different models that classify differently on same inputs. Hence, the decision of choosing the algorithms that are most suitable for the stratification problem was made by estimating the values of correctly classified instances, time to learn, kappa statistics, sensitivity and specificity. Afterwards, the empirical minimization technique was performed on the two classifiers to determine their minimum loss function (empirical risk).

3.1.2 Empirical Risk Minimization Computation

While determining the empirical risk minimization function, the learning hypothesis h: $X \rightarrow R$ was set with the training set = $(x_1, r_1), \ldots (x_m, r_m)$ where $x_i \in X$ is the input, and $r_i \in \mathcal{R}$ is the output to give $h(x_i)$, while the probability distribution $P(x, r)$ over x and r is independently and identically distributed (i.i.d). The loss function $L(\hat{r}, r)$ was determined to measure the difference between the true or expected output r and the predicted output $\hat{r}$ of the hypothesis using: $R(h) = E[L(h(x), r)] = \int L(h(x), r) dP(x, r)$ to find the hypothesis h^* among a fixed class of function $\mathbb{H}$ for which the risk $R(h)$ is minimal: $h^* = \arg min_{h \in H} R(h)$; and empirical risk minimization was computed using $R_{emp}(h) = \frac{1}{m} \sum_{i=1}^{m} L(h(x_i), r_i)$ on Logistic regression and PART algorithms, whereby Logistic regression which had the lowest and non-negative loss function value was selected as the optimal one.

Problem Definition

In order to obtain the desired results, some assumptions on the class of the dataset were made. Given a CML dataset $\mathcal{D}$ of n individuals, where each observation d_i lies in this domain, the classifier that had the minimum loss function (empirical risk) was computed by averaging loss function on the training sets. From this point, it was assumed that the CML dataset had been processed, and the learning data has the specification of two spaces: $X \equiv$ Input space and $R \equiv$ Output space.

Training Set of the dataset $\mathcal{D}$

$\mathcal{D}$ = {Input set; Output set} = {Basophil count, Spleen size; Risk score} = {B, S; R}

Input Set

$\mathcal{X} = \{\mathcal{B}, \mathcal{S}\} = \{x_1, x_2\}$

where $x_1 = \mathcal{B} = \{b_1, b_2, \ldots, b_n\}$

$x_2 = \mathcal{S} = \{s_1, s_2, \ldots, s_n\}$

Test Set: Output Set

$\mathcal{R} = \{r_1, r_2, \ldots, r_n\}$

There are two spaces of objects $\mathcal{X}$ and $\mathcal{R}$ which learn a function h: $\mathcal{X} \rightarrow \mathcal{R}$ in which:

Output $r \in \mathcal{R}$ given $x \in \mathcal{X}$

There is the training set $(x_i, r_i), \ldots (x_n, r_n)$, where $x_i \in \mathcal{X}$ is an input and $r_i \in \mathcal{R}$ is the corresponding output were the hypothesis h (x_i) was derived from. It was assumed that there is a joint probability distribution $P(x, r)$ over x and r, and the training set consist of m instances $(x_i, r_i), \ldots (x_m, r_m)$, drawn independently and identically distributed (i.i.d) from distribution $P(x, r)$.

3.1.2.1 Loss Function

The concept of loss function L was introduced and it was assumed that there are non-negative real-valued Loss function $L(\hat{r}, r)$ which measures how the expected or true outcome r is different from predicted $\hat{r}$ of a hypothesis. The risk associated with the hypothesis $h(x)$ is the expectation of the Loss function L using the equations 7 to 12 as defined by Vapnik (2000):

$$R(h) = E[L(h(x), r)] = \int L(h(x), r) dP(x, r) \quad (7)$$

In order to determine the risk of the hypothesis *h* in Equation 8, the integration of the product of the loss function, input *x* and the output *r* on the data distribution was done. However, the ultimate goal was to find a hypothesis *h** among a fixed class of function $\mathcal{H}$ for which the risk *R*(*h*) is minimal:

$$h^* = \arg min_{h \epsilon H} R(h) \quad (8)$$

Therefore Equation 8 gives the minimum hypothesis from the training set that has the minimum loss function.

3.1.2.2 Expected Risk

The expected risk was used to measure the expected performance of the algorithm with respect to L1-regularized logistic regression solver. With a given function *f*, loss function *L*, and a probability distribution $P(x, r)$, the expected risk or true risk of *f* was given to minimize the loss of test data.

$$R_{L,P}(f) = \int_{x*r} L\big(x, r, f(x)\big) dP(x, r), \quad dP(x, r) = P(x, r)\, d\, r dx$$

$$= E[L(X, \mathcal{R}, f(x))] \quad (9)$$

where $L\big(x, r, f(x)\big)$: Loss function

$P(x, r)$: Distribution of the data

The expected risk of the loss function and distribution of the data was computed by finding the integration of the product of input *x* and output *r*, in which the derivative was done to give *dr dx* i.e. the output (risk group) as shown in equation 9.

3.1.2.3 Empirical Risk Minimization Function

The ERM was used to choose the hypothesis (rule) that minimizes the empirical risk $\hat{h} = \arg min_{h \epsilon H} R_{emp}(h)$. In equation 10, the empirical risk of the hypothesis is averaged by $\frac{1}{m}$ where *m* is the number of inputs considered in the training set.

$$R_{emp}(h) = \frac{1}{m} \sum_{i=1}^{m} L(h(x_i), r_i) \quad (10)$$

Where $n(m) = \{\mathcal{B}, \mathcal{S}\} = 2$

$$\therefore m = 2$$

$$R_{emp}(h) = \frac{1}{2} \sum_{i=1}^{2} L(h(x_i), r_i) \quad (11)$$

Therefore, equation 11 gives the empirical risk function on the hypothesis. However, the interest was to choose a hypothesis $\hat{h}$ that minimizes the empirical risk as shown in equation 12.

$$\hat{h} = \arg min_{h \epsilon H} R_{emp}(h) \quad (12)$$

In equation 12, the mathematical operator 'argmin' returns a value minimizing the argument function, hence, the learning algorithm defined by the ERM principle consists in solving the above problem in equation 12. The ERM was tested on the Logistic regression and PART algorithms, and Logistic regression which had the minimum loss function (risk) was identified as the "best fit" algorithms for CML data stratification. Therefore, the ERM function was introduced in the Chronic Myeloid Leukemia Data Stratification model.

3.2 Development of a model for stratifying chronic myeloid leukemia using algorithm with minimal loss function

There are existing models that have been developed using the classification and prediction algorithm that perform best in their respective problem areas. The existing models are: (1) NCC-AUC: An AUC optimization method to identify multi-biomarker panel for cancer prognosis from genomic and clinical data by Meng *et al*. (2015); (2) Prospective stratification of patients at risk for emergency department revisit: Resource utilization and population management strategy implications by *Bo et al*. (2016); and (3) The SurvivalNet framework by Safoora *et al*. (2017). These models served as models leveraged on for the development of a stratification model that used the classifier with minimal loss function.

The model aimed at stratifying CML dataset using the algorithm with the minimal loss function between the two algorithms that satisfied the standards of performance evaluation (i.e. the percentage of correctly classified instances, time to learn, kappa statistics, sensitivity and specificity). The model has nine (9) components namely: Data collection phase (retrospective and prospective data), pre-processing phase, learning phase, classifier selection phase, ERM computation phase, ERMDS algorithm phase, data stratification phase, ERMDS system and predictive score as shown in Figure 1.

1. **Data Collection Phase:** In this phase Chronic Myeloid Leukemia – Chronic Phase patients' data that were treated with imatinib was collected from the Haematology Department, OAUTHC, Ile-Ife, Osun State, Nigeria. The dataset contained both the retrospective and

prospective data between year 2003 and 2017 for the purpose of this study.

2. **Pre-processing Phase:** Pre-processing of the data was done to improve the predictive accuracy of the dataset since it is susceptible to noise, missing and inconsistent data due to the nature of data collected and human error. This process reduced the amount of memory space consumed, computation power and over-fitting of the model.
3. **Learning Phase:** Supervised learning approach of Machine Learning algorithms was applied on the retrospective and prospective data. Four classification algorithms were used where the class label predicted based on Holdout (66% percentile split) and 10-fold cross-validation evaluation performances. The determinants for the performance of the algorithms are correctly classified, time to learn, kappa statistics, sensitivity and specificity.
4. **Algorithm Selection Phase:** Having carried out the learning, comparison of the four algorithms was done to choose the two with leading performances.
5. **ERM Computation Phase:** The ERM technique was computed on the two leading algorithms using $R_{emp}(h) = \frac{1}{m}\sum_{i=1}^{m} L(\mathrm{h}(x_i), r_i)$ to choose one with the minimum loss function.
6. **ERDMS Algorithm Phase:** The $R_{emp}(h)$ function was used to formulate the ERMDS algorithm for data stratification.
7. **Data Stratification Phase:** At this phase, the classifier with minimal risk was chosen to determine the predictive score, thereby stratifying the patients' into either high risk (score > 87) or low risk (score ≤ 87) based on European Treatment and Outcome Study (EUTOS) standard.
8. **ERMDS System:** It is the system that interfaces with the user to give the predictive scores.
9. **Predictive Score:** At this phase, the optimal algorithm employed was used to determine the predictive score, and then the patient was categorized to be in high or low risk group.

3.3 Dataset Description

The dataset used in this study was the Chronic Myeloid Leukemia data obtained from Obafemi Awolowo University Teaching Hospitals Complex (OAUTHC). The dataset contains 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) are 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 Waikato Environment for Knowledge Analysis (WEKA) was created for the data. The grouping of the variables is shown in Table 2. The risk group is the response variable while other variables are predictors. Each variable is suitably categorized to accommodate all the available information.

4. RESULTS

The four classification algorithms namely BayesNet, Multilayered perceptron, PART and Logistic Regression were built in WEKA 3.8.0 and evaluated with the holdout (66%) and 10-fold cross-validation techniques for training and testing on the CML dataset. The performance of the four (4) algorithms were measured based on five (5) existing performance benchmarks: correctly classified instances, time to learn, kappa statistics, sensitivity and specificity. Figures 2 to 9 depict the screen shots of the explorer view of the algorithms implemented in WEKA.

4.1 Summary of Algorithm Performance Compared

The performance of the four classification algorithms (BayesNet, Multilayered perceptron, PART and Logistic Regression) were compared using correctly classified instance, time to learn, kappa statistics, sensitivity and specificity metrics. Logistic regression and PART algorithm had correctly classified instance values of 99.82% and 99.64% in holdout, and 99.76% and 99.58% in cross-validation respectively, which outperformed other algorithms. Considering the time taken in learning in relation to the performance of the algorithms, Multilayered perceptron consumed much time and computation resources, but it can be ascertained that Logistic regression and PART algorithms took shorter time and outperformed other algorithms relative to the accuracy level in this study with 0.02 and 0.09seconds in both holdout and 10-fold cross validation methods. In

kappa statistics, PART and Logistic regression had the greatest result in both holdout and 10-fold cross validation methods. In holdout, PART had 99.64% and Logistic Regression had 99.57% while in 10-fold cross validation method PART had 99.82% and Logistic Regression had 99.71%. In sensitivity, Multilayered Perceptron and PART algorithm had the greatest value of 99.99% and 99.60% holdout method while PART and Logistic Regression have the same value of 99.47% in 10-fold cross validation. The specificity of the algorithms showed that PART and Logistic Regression had the greatest value of 99.99% each in holdout method, while in 10-fold cross validation method the two algorithms had 99.55% and 99.98% respectively. Therefore, it was deduced that Logistic regression and PART algorithms were the two good classifiers for stratifying patients' risk group as against other algorithms used in this study. Table 3 presents the summary of the performance of the models based on the benchmarks.

Thus, from the performance evaluation carried out on the algorithms, Logistic regression and PART algorithms were discovered to have the best performance based on their ability to correctly classify the chronic myeloid leukemia patient dataset within the lowest possible time of 0.02 and 0.09 seconds.

4.2 Determination of Minimum Loss Function of Algorithms

The decision of choosing the minimum loss function (i.e. the empirical risk) of Logistic regression and PART algorithms was defined by finding the difference between the predicted output and true output of the algorithms for high and low risk groups in the holdout and 10-fold cross validation methods as discussed in sections 4.2.1 and 4.2.2.

4.2.1 Loss Function for Logistic Regression Algorithm

In determining the loss function for logistic regression, the delta Δ (i.e. differences) between the values of the true output and the predicted output were computed for both high and low risk patients. The values in the numerator and the denominator are derived from the confusion matrix in logistic regression output. In holdout method the numerators 259 is the total number of patients classified as low risk and 308 patients are classified as high risk, while the denominator 558 is the total number of instances. The predicted output $\hat{r}$ for low risk and high risk patients were determined by finding the percentage of the confusion matrix value divided by the number of instances in both holdout (66% split) and cross-validation methods as shown below. The true output r is 46.20% for low risk patients and 53.80% for high risk patients.

In Holdout Method

Predicted output $\hat{r}$ for low risk = $\frac{259}{558}$ x 100%
= 46.42%

and

Predicted output $\hat{r}$ for high risk = $\frac{308}{558}$ x 100%
= 55.20%

Loss function *L* in holdout method = $\Delta\ (\hat{r} - r)$

For low risk $\boldsymbol{L} = \Delta\ (\hat{r} - r)$ = (46.42 – 46.20)%
= 0.22%

For high risk $\boldsymbol{L} = \Delta\ (\hat{r} - r)$ = (55.20 – 53.80)%
= 1.40%

In 10-fold Cross-validation Method

In 10-fold Cross-validation method, determining the loss function for logistic regression, the delta Δ (i.e. differences) between the values of the true output and the predicted output were computed for both high and low risk patients. The values in the numerator and the denominator are derived from the confusion matrix in logistic regression output. In this method the numerators 754 is the total number of patients classified as low risk and 882 patients are classified as high risk, while the denominator 1640 is the total number of instances. The predicted output $\hat{r}$ for low risk and high risk patients were determined by finding the percentage of the confusion matrix value divided by the number of instances in both holdout (66% split) and cross-validation methods as shown below. The true output r is 46.20% for low risk patients and 53.80% for high risk patients.

Predicted output $\hat{r}$ for low risk = $\frac{754}{1640}$ x 100%
= 45.98%

and

Predicted output $\hat{r}$ for high risk = $\frac{882}{1640}$ x 100%
= 53.78%

Loss function *L* in 10-fold cross-validation method = $\Delta\ (\hat{r} - r)$

For low risk $\boldsymbol{L} = \Delta\ (\hat{r} - r)$ = (45.98 – 46.20)%
= -0.22%

For high risk $\boldsymbol{L} = \Delta\ (\hat{r} - r)$ = (53.78 – 53.80)%
= -0.02%

Hence, the findings from the loss function in holdout method gave the loss function of 0.22% for stratifying patient into low risk and 1.40% for stratifying patient into high risk. In 10-fold cross-validation method, the loss function gave the empirical risk of -0.22% for stratifying patient into low risk and -0.02% for stratifying patient into high risk.

4.2.2 Loss Function for PART Algorithm

In determining the loss function for PART algorithm, the delta Δ (i.e. differences) between the values of the true output and the predicted output were computed for both high and low risk patients. The values in the numerator and the denominator are derived from the confusion matrix in PART output. In holdout method the numerators 249 is the total number of patients classified as low risk and 308 patients are classified as high risk, while the denominator 558 is the total number of instances. The predicted output $\hat{r}$ for low risk and high risk patients were determined by finding the percentage of the confusion matrix value divided by the number of instances in both holdout (66% split) and cross-validation methods as shown below. The true output r is 46.20% for low risk patients and 53.80% for high risk patients.

In Holdout Method

Predicted output $\hat{r}$ for low risk = $\frac{249}{558}$ x 100%

= 44.62%

and

Predicted output $\hat{r}$ for high risk = $\frac{308}{558}$ x 100%

= 55.20%

Loss function *L* in holdout method = $\Delta\ (\hat{r} - r)$

For low risk $\boldsymbol{L} = \Delta\ (\hat{r} - r)$ = (44.62 – 46.20)%

= -1.58%

For high risk $\boldsymbol{L} = \Delta\ (\hat{r} - r)$ = (55.20 – 53.80)%

= 1.40%

In 10-fold Cross-validation Method

In 10-fold Cross-validation method, determining the loss function for logistic regression, the delta Δ (i.e. differences) between the values of the true output and the predicted output were computed for both high and low risk patients. The values in the numerator and the denominator are derived from the confusion matrix in logistic regression output. In this method the numerators 754 is the total number of patients classified as low risk and 878 patients are classified as high risk, while the denominator 1640 is the total number of instances. The predicted output $\hat{r}$ for low risk and high risk patients were determined by finding the percentage of the confusion matrix value divided by the number of instances in both holdout (66% split) and cross-validation methods as shown below. The true output r is 46.20% for low risk patients and 53.80% for high risk patients.

Predicted output $\hat{r}$ for low risk = $\frac{754}{1640}$ x 100%

= 45.98%

and

Predicted output $\hat{r}$ for high risk = $\frac{878}{1640}$ x 100%

= 53.54%

Loss function *L* in cross-validation method = $\Delta\ (\hat{r} - r)$

For low risk, $\boldsymbol{L} = \Delta\ (\hat{r} - r)$ = (45.98 – 46.20)%

= -0.22%

For high risk, $\boldsymbol{L} = \Delta\ (\hat{r} - r)$ = (53.54 – 53.80)%

= -0.26%

The findings from the loss function in holdout method gave the empirical risk of -1.58% for stratifying patient into low risk and 1.40% for stratifying patient into high risk; while in cross-validation method, the loss function gave the empirical risk of -0.22% for stratifying patient into low risk and -0.26% for stratifying patient into high risk.

Interpretation

From the findings discussed in sections 4.2.1 and 4.2.2, the result showed that logistic regression had a minimal loss function with non-negative values in stratifying high and low risk patient in holdout method with values of 0.22% and 1.40% and respectively. In cross-validation method both logistic regression and PART algorithms had negative real valued loss function. In essence, logistic regression is a good classifier with which the risk of the hypothesis was minimized, and that informs the decision of using Logistic regression algorithm in the model to stratify the dataset.

Empirically, the loss functions of logistic regression and PART algorithm were compared, and the result showed that logistic regression in holdout method offered clear advantage in the presence of outliers.

5. CONCLUSION

Minimization of the empirical risk by finding the loss functions of logistic regression algorithm has played a big role in producing optimum and faster results for accurate predictions. The findings of this research in relation to other studies like Kamalika et al. (2011) showed how empirical risk minimization concept was used for privacy-preserving approximations of Logistic regression and Support Vector Machine classifiers to predict whether a network connection was a denial-of-service attack or not. Sensitivity method and objective perturbation algorithms were used by tuning algorithm and Michael and S´ebastien (2015) and Yuchen (2016) agreed with this assertion. The findings of this research agreed with other studies that employed empirical risk minimization technique to determine the loss function of classifiers before choosing and employing an algorithm for stratifying or predicting a dataset from any problem domain. The use of ERM had helped to determine the loss function of the two algorithms (Logistic regression and PART), that had great performance using some metrics (correctly classified instances, time to build, kappa statistics, sensitivity and specificity). Hence, logistic regression had the lowest non-negative loss function in holdout method, and it enhanced the decision of using logistic regression for CML data stratification into their risk group.

Therefore, determining the loss function (empirical risk) of machine learning algorithm is significant when building predictive or prognostic tools. This is important since it would aid the decision of choosing an algorithm for the dataset from the problem domain.

REFERENCES

[1] Atul, K. P., Prabhat, P., & Jaiswal, K. L. (2014). Classification Model for the Heart Disease Diagnosis. *Global Journal of Medical Research Diseases, 14*(1), 1-7.

[2] Barnabas, H. J. (2012). Time-To-Event Predictive Modeling for Chronic Conditions Using Empirical Risk Minimization Technique. *IEEE Intelligent Systems, 29*(3), 14-20.

[3] Besa, E. C., Buehler, B., Markman, M., & Sacher, R. A. (2013). *Chronic Myelogenous Leukemia Krishnan* (3rd ed.). Waterloo, Canada.

[4] Bo, J., Yifan, Z., Shiying, H., Andrew, Y. S., Yue, W., Chunqing, Z., … Xuefeng, B. L. (2016). Prospective Stratification of Patients at Risk for Emergency Department Revisit: Resource Utilization And Population Management Strategy Implications. *BMC Emergency Medicine, 16*(1), 1-10.

[5] Chaudhuri, K., Sarwate, A. D., & Sinha, K. (2013). A Near-Optimal Algorithm for Differentially-Private Principal Components. *The Journal of Machine Learning Research, 14*(1), 2905-2943.

[6] David, B. (2012). *Bayesian Reasoning and Machine Learning* (2nd ed.). Wellington, New Zealand.

[7] Elbedewy, T. A., & Elashtokhy, H. E. (2016). The Utility and Applicability of Chronic Myeloid Leukemia Scoring Systems for Predicting the Prognosis of Egyptian Patients on Imatinib: A Retrospective Study. *J Leuk, 4*(210), 1-9.

[8] Eric, P., Rami, K., & Alan, F. L. (2014). The Clinical Management of Chronic Myelomonocytic Leukemia. *Journal of Clinical Advances in Hematology and Oncology, 12*(3), 172-178.

[9] Frank, E., & Witten, I. (2011). Generating Accurate Rule Sets Without Global Optimization. Proceedings of the Fifteenth International Conference, (pp. 144-151). Madison, San Francisco.

[10] Hasford, J., Baccarani, M., Hoffmann, V., Guilhot, J., & Saussele, S., (2011). Predicting Complete Cytogenetic Response and Subsequent Progression Free Survival in 2060 Patients With CML on Imatinib Treatment: The EUTOS Score. *Blood, 118*: 686-692.

[11] Hina, F., Syed, I. H., & Harleen, K. (2018). A Comparative Survey of Machine Learning and Meta-Heuristic Optimization Algorithms for Sustainable and Smart Healthcare. *Afr. J. Comp. & ICT, 11*, 4, 1 - 17.

[12] Ian, H. W., & Eibe, F. (2005). *Data Mining Practical Machine Learning Tools and Techniques* (2nd ed.). Department of Computer Science, University of Waikato. The Morgan Kaufmann Series in Data Management Systems, Waikato.

[13] Ji, Z., Jiang, X., Wang, S., Xiong, L., & Ohno-Machado, L. (2014). Differentially private Distributed Logistic Regression Using Private and Public Data. *BMC medical genomics*, *7*(1), Suppl 1, S14.

[14] Jonathan, H. C., & Steven, M. A. (2017). Machine Learning and Prediction in Medicine - Beyond the Peak of Inflated Expectations. *Machine Learning Informatics*, *20*(31), 2507-2509.

[15] Kamalika, C., Claire, M., & Anand, D. S. (2011). Differentially Private Empirical Risk Minimization. *Journal of Machine Learning Research*, *12*, 1069-1109.

[16] Liyang, X. (2016). Comparison of Two Models in Differentially Private Distributed Learning (A published M.Sc dissertation). New Brunswick, New Jersey.

[17] Mahdavi, M., Zhang, L., & Jin, R. (2014). Binary Excess Risk for Smooth Convex Surrogates. *arXiv preprint arXiv*:1402.1792.

[18] Mehryar, M., Afshin, R., & Ameet, T. (2012). *Foundations of Machine Learning* (2nd ed.). The MIT Press Cambridge, Massachusetts Lodon, England.

[19] Meng, Z., Zhaoqi, L., Xiang-Sun, Z., & Yong, W. (2015). NCC-AUC: An AUC Optimization Method to Identify Multi-Biomarker Panel for Cancer Prognosis from Genomic and Clinical Data. *Bioinformatics*, *31*(20), 3330-3338.

[20] Michael, C., & S´ebastien, L. (2015). Bandwidth Selection in Kernel Empirical Risk Minimization Via the Gradient. *The Annals of Statistics*, *43*(4), 1617-1646.

[21] Oladejo, A. K., Oladele, T. O., & Saheed, Y. K. (2018), Comparative Evaluation of Linear-SVM and KNN Algorithm. *Afr. J. Comp. & ICT, 11*, 2, 1-10.

[22] Oyekunle, A. A., Osho, P. O., Aneke, J. C., Salawu, L., & Durosinmi, M. A., (2012). The Predictive Value of the Sokal and Hasford Scoring Systems in Chronic Myeloid Leukaemia In The Imatinib Era. *Journal of Hematological Malignancies*, *2*(2), 25-32.

[23] Poline, J.-B., Breeze, J. L., Ghosh, S., Gorgolewski, K., Halchenko, Y. O., Hanke, M., ... Marcus, D. S. (2012). Data Sharing in Neuroimaging Research. *Frontiers in Neuroinformatics*, *6*(10) 87-109.

[24] Safoora, Y., Fatemeh, A., Mohamed, A., Coco, D., Joshua, E. L., Congzheng, S., … Lee, A. D. C. (2017). Predicting Clinical Outcomes from Large Scale Cancer Genomic Profiles with Deep Survival Models. *BioRxiv Journal*, doi: http://dx.doi.org/10.1101/131367.

[25] Sepp, H. (2013). *Theoretical Bioinformatics and Machine Learning* (2nd ed.). Wellington, New Zealand.

[26] Shouval, R., Bondi, O., Mishan, H., Shimoni, A., Unger, R., & Nagler, A. (2014). Application of Machine Learning Algorithms for Clinical Predictive Modeling: A Data-Mining

Approach in SCT. *Bone Marrow Transplantation*, *49*, 332-337, doi:10.1038/bmt.2013.146.

[27] Stylianou, N., Akbarov, A., Kontopantelis, E., Buchan I., Dunn, K. (2015). Mortality Risk Prediction in Burn Injury: Comparison of logistic Regression with Machine Learning Approaches. Annals of Burns and Fire Disasters, 41(5), 925-934.

[28] Vapnik, V. (2000). The nature of statistical learning theory. *Information Science and Statistics*. *Springer-Verlag*. ISBN 978-0-387-98780-4.

[29] World Health Organization. (2016). *World Health Statistics 2016*. Retrieved from http://www.who.int/whosis/whostat/2016/en/

[30] World Health Organization. (2017). *World Health Statistics 2017*. Retrieved from http://www.who.int/whosis/whostat/2017/en/

[31] Yuchen, Z. (2016). Distributed Machine Learning With Communication Constraints (A Published Doctoral Thesis). California, Berkeley.

Table 1: Scoring systems and their calculation methods (Elbedewy & Elashtokhy, 2016)

Scoring system	**Calculation method**	**Risk definition**
Sokal score	Exp [0.0116 x (age in years - 43.4) + 0.0345 x (spleen size cm below costal margin - 7.51) + 0.188 x (platelet count⁄700)2 - 0.563) + 0.0887 x (blast cell % in peripheral blood - 2.10)]	Low risk (score < 0.8) Intermediate risk (0.8 ≤ score ≤ 1.2) High risk (score >1.2)
Hasford score	[0.666 (when age ≥ 50 years) + (0.042 x spleen size cm below costal margin) + 1.0956 (when platelet count >1500 x 109 ⁄ L) + (0.0584 x blast cell % in peripheral blood) + 0.20399 (when basophil % in peripheral blood ≥3%) + (0.0413 x eosinophil % in peripheral blood)] x 1000.	Low risk (score ≤ 780) Intermediate risk (score > 780 ≤ score 1480) High risk (score > 1480)
EUTOS score	(7 x basophils % in peripheral blood) + (4 x spleen size cm below costal margin)	Low risk (score ≤ 87) High risk (score > 87)

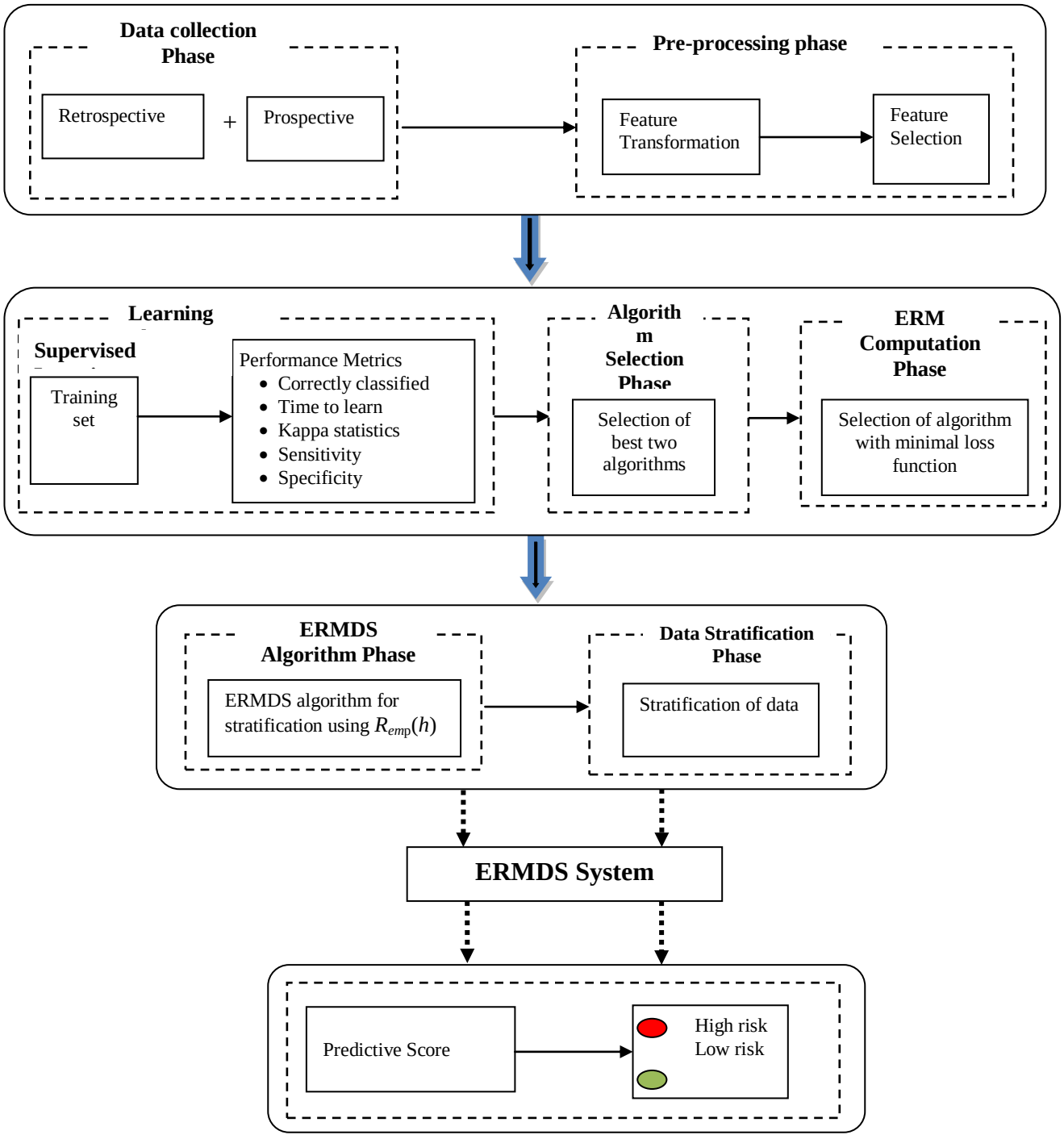

Figure 1: A CML Data Stratification Model

Table 2: 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

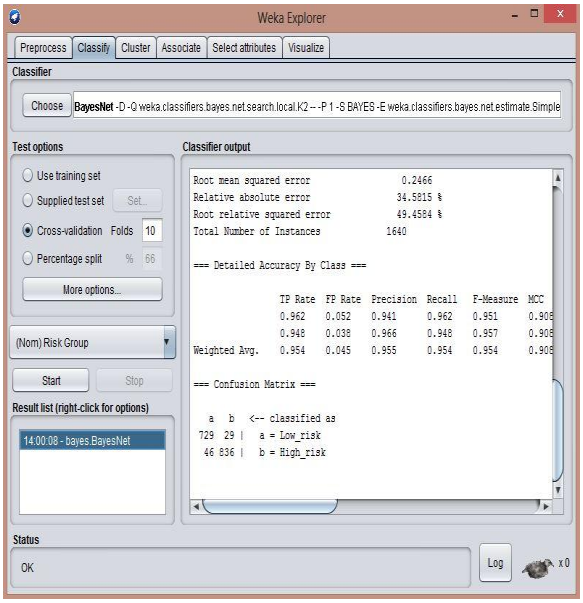

Figure 2: Screenshot of BayesNet in Cross validation

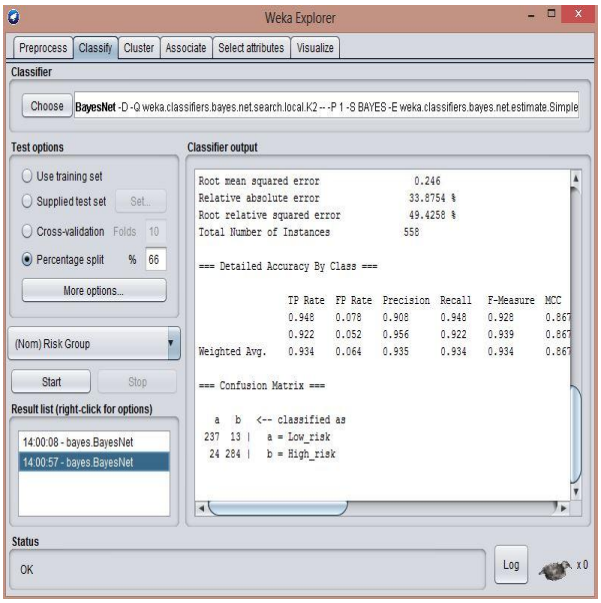

Figure 3: Screenshot of BayesNet in Holdout method

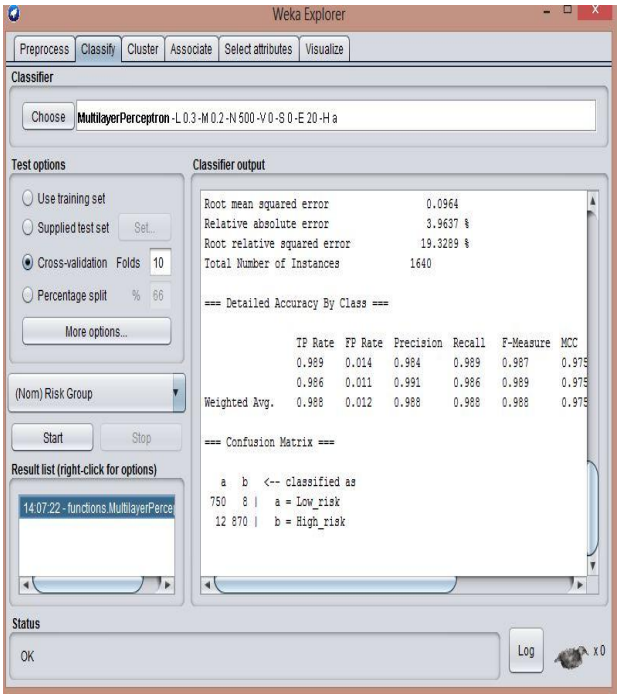

Figure 4: Screenshot of Multilayer perceptron in Cross validation

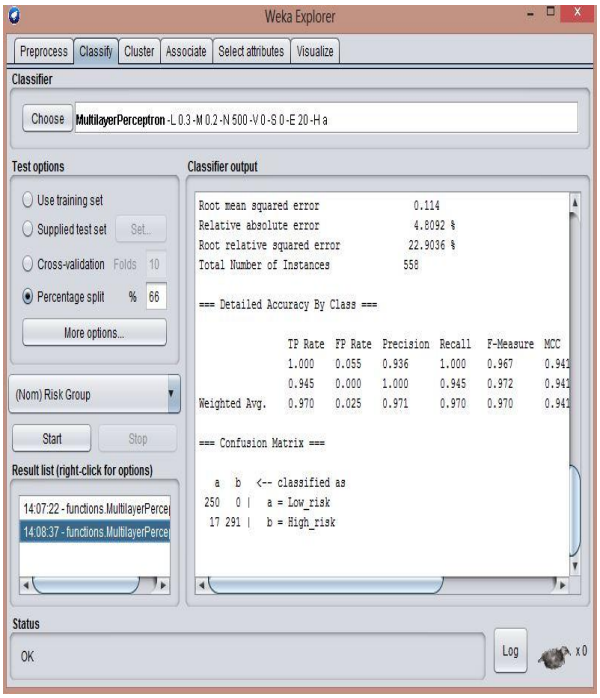

Figure 5: Screenshot of Multilayer perceptron in Holdout method

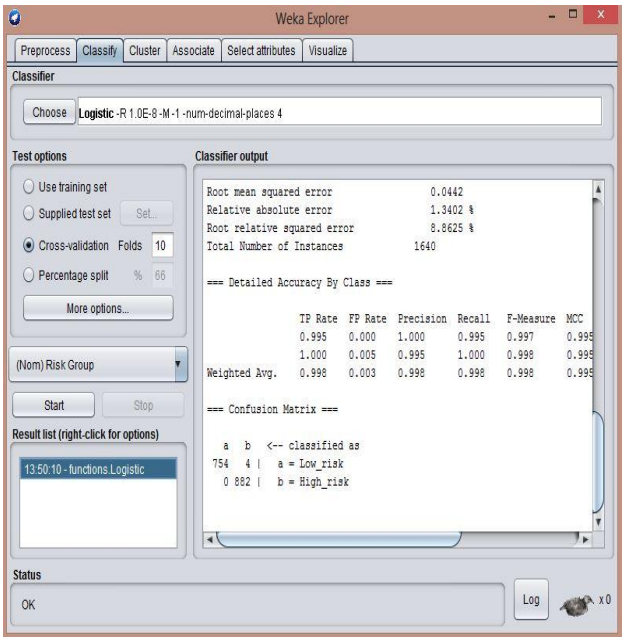

Figure 6: Screenshot of Logistic regression in Cross validation

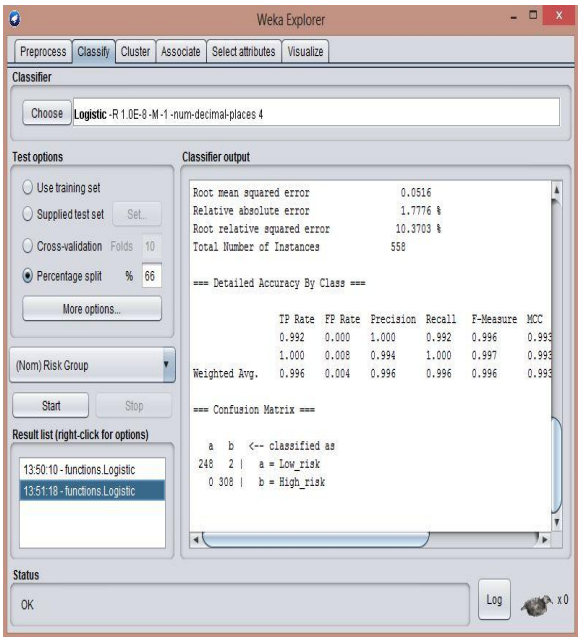

Figure 7: Screenshot of Logistic regression in Holdout method

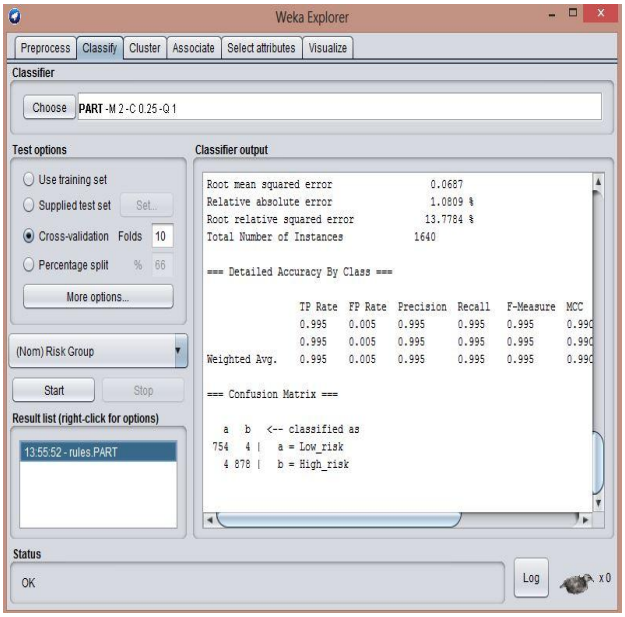

Figure 8: Screenshot of PART in Cross validation

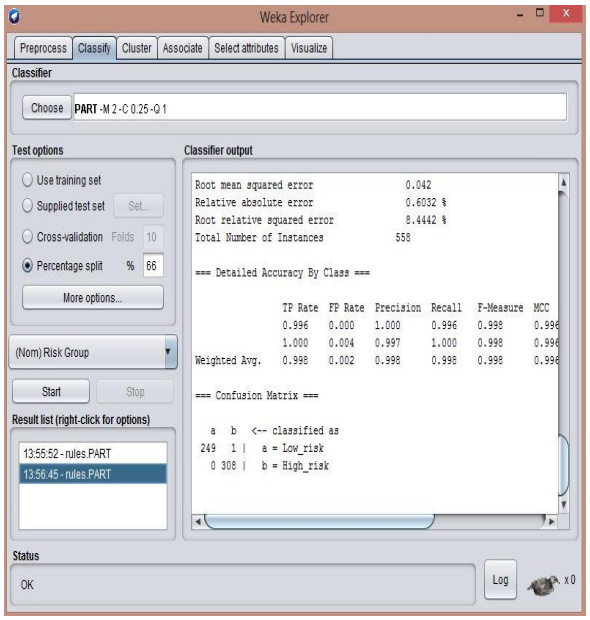

Figure 9: Screenshot of PART in Holdout method

Table 3: Summary of the model performances in holdout and 10-fold cross validation method

	Hold-Out (66% train, remainder test)						10-Cross Validation				
S/N	**Classifier**	**CCI (%)**	**T (s)**	**KS (%)**	**S_e (%)**	**S_p (%)**	**CCI (%)**	**T (s)**	**KS (%)**	**S_e (%)**	**S_p (%)**
1	BayesNet	93.37	0.13	86.65	94.80	92.20	95.43	0.15	90.82	96.17	94.78
2	Multilayered perceptron	96.95	1.83	93.88	99.99	94.48	98.78	1.18	97.55	98.94	98.64
3	PART	99.64	0.09	99.64	99.60	99.99	99.58	0.09	99.82	99.47	99.55
4	Logistic Regression	99.82	0.02	99.57	99.20	99.99	99.76	0.02	99.71	99.47	99.98

Key: CCI = Correctly Classified Instances, T = Time to build, KS = Kappa Statistics, S_e = Sensitivity, S_p = Specificity