Optimality Assessments of Classifiers on Single and Multi-labelled Obstetrics Outcome Classification Problems

Udoinyang G. Inyang[1], Samuel A. Robinson[2], Funebi F. Ijebu[3], Ifiok J. Udo[4], Chuwkudi O. Nwokoro[5]
Department of Computer Science, Faculty of Science, University of Uyo, Nigeria[1, 2, 3, 4, 5]
TEFTFUND Centre of Excellence in Computational Intelligence Research, University of Uyo, Nigeria[1, 2, 3, 4, 5]

***Abstract*—It is indisputable that clinicians cannot exactly state the outcome of pregnancies through conventional knowledge and methods even as the surge in human knowledge continues. Hence, several computational techniques have been adapted for precise pregnancy outcome (PO) prediction. Obstetric datasets for PO determination exist as single label learning (SLL), multi-label learning (MLL) and multi-target (MTP) problems. There is however no single classifier recommended to optimally satisfy the needs of all the classification types. This work therefore identifies six widely used PO classifiers and investigates their performances in all three classification categories; to find the best performing classifier. Obstetric dataset exposed to input rank analysis via Principal component Analysis, produced thirteen (13) significant features for the experiment. Accuracy, F1-measure and build/test time were used as evaluation metrics. Decision tree (DT) had an average accuracy and F1 score of 89.23% and 88.23% respectively, with 1.0 average rank. Under MLL configuration, average accuracy (91.71%) and F1 score (94.28%) were highest in the random forest (RF) which had a 1.0 average test time rank. Using MTP, DT had an average accuracy of 88.80% and average F1 score of 71.13%, the multi-layered perceptron (MLP) had the best time cost with an average rank value of 2.0. From the results, RF is most optimal in terms of accuracy and average rank value, while DT is the most efficient in terms of time cost. The comparative analysis of global averages of the six base classifiers shows that RF is the most optimal algorithm with an average accuracy of 87.3% given all three data setups in the study. MLP on the other hand had an unexpectedly high time cost, making it unsuitable for similar data classifications if time is the main criterion. It is recommended that the choice of the classifier should either be RF or DT depending on the application domain and whether or not time cost is a major consideration.**

Keywords*—Pregnancy outcome; random forest; multi-label learning; comparative analytics; machine learning algorithms; single label learning; maternal outcome prediction; decision tree

I. Introduction

Machine Learning (ML), a fast-rising branch of artificial intelligence (AI), encompasses computer science, engineering, mathematical sciences, cognitive science and many more disciplines [1]. The advancement and wide applications of ML is largely due to the availability of enormous data repositories and the satisfaction and reliability of its performances — accuracy and computational cost. It equips systems with cognitive capability of understanding the concepts of their environments through the building of models and functions, and the communication of their experiences with patterns. These models and patterns are built and implemented through the process called ML. There are two key classes of ML — supervised ML (SML) and unsupervised ML (UML) [1]. Both UML and SML draw inferences by learning, however UML utilizes datasets with input features only while SML depends on datasets having both input and target attributes for mapping and extraction of relationships between input and output feature spaces. Any dataset with target or desired output variable(s) is referred to a labelled dataset. Unlabeled datasets lack response variables therefore do not support model training activity needed by SML techniques [2-4]. In labelled datasets, every record has predefined class label(s) and supports two broad types of data mining applications — regression and classification [5]. In regression tasks, the target variable(s) is in continuous numeric form whereas classification requires class labels or categorical variables as the target. Classification is the most common and widely applied SML approach. It is aimed at identifying and assigning membership class to a new record, from a set of already defined classes [4,6]. Classification approaches are sub-divided into two groups according to the number of labels; single label and multi-label. The conventional single-label classification approach deals absolutely with disjoint classes—each record belongs exclusively to a unique class, whereas in multi-label classification the labels are intertwined and each record is associated with two or more class labels [7]. In single label problems, the categories may comprise of two labels (binary class) or more than two labels (multi-class). For example in medical diagnosis, a laboratory test result might confirm the presence or otherwise of causative organisms in the tested patient's sample while the patient can concurrently suffer from more than two diseases.

In maternal healthcare (MHC), obstetricians are confronted with the tasks ensuring safety of both the mother and baby throughout pregnancy, during delivery, and within a specified period after delivery. This is achieved by providing specialized medical care services while she is expectant, during child delivery and after delivery — antenatal, neonatal and post-natal care services. They are therefore required to obtain clinical factors for the realization of the safety of mother throughout the period during pregnancy and birth, and the newborn in a bid to minimize mortality and morbidity. These involve simultaneous predictions of multiple outcome regarding mother and neonatal status using common baseline

risk factors. Maternal outcome, mother's status during and after delivery, neonatal physiological status, conditions and overall state among others are central in MHC management. Hence, multiple target prediction, multi-label and multi-class predictions are essentially mandatory tasks in the obstetric healthcare domain. However, these maternal decisions are repeatedly made based on doctors' perceptions and experience without utilizing the pieces of vital knowledge concealed in the huge data repositories [8,9]. The author in [10] state that only about 30% of pregnancy outcomes classified by gynecologists and obstetricians concerning pathological fetus or pregnancy turns out to be true. This limitation in current medical practice has led to several complications in deliveries and avoidable deaths from the over 130 million deliveries per year globally. It is therefore expected that a robust computational technique for accurate pregnancy outcome determination will be available to assist medical personnel.

Although solutions from data mining and computational models are laudable and widely accepted methods for medical predictions, none is confirmed as a universal and best-performing model for prediction of diverse maternal outcomes; individually or in a combined target setup. This paper aims at assessing the performances and suitability on obstetrics dataset, classification algorithms under varying maternal outcome target configurations, given that they comprise binary, multi-class and multi-labeled target features. The remaining sections are structured as follows: Section 2 gives a review of related works on medical diagnoses regarding maternal health care management. In Section 3, the dataset acquisition, preprocessing and description are presented while the methodology of the comparative analytics is described in Section 4. The predictive results along with the evaluations of their performances as well as discussions are described in Section 5 while conclusions and further directions are given in Section 6.

II. Related Works

A. Single Label Learning

Classification tasks are broadly categorized into single-label learning (SLL) and multi-label learning (MLL) based on the nature of association existing between target labels and input patterns [11,12]. The goal of SLL is to build a model for the prediction of a distinct class label from a set of non-overlapping labels using input samples. It deals solely with disjoint classes and comprises two types: binary (or filtering in of textual and web-data domian) [13] and multi-class classification [11]. Binary classification has two unique class labels and involves the mapping of input features to only one of the two classes based on an explicit assessment criterion. Examples include disease diagnosis (positive or not), gender discrimination (male or female), email spam detection (spam or not), quality control (pass or fail), maternal status after delivery (alive or death) among others. Some of the famous binary classification datasets are adult dataset (adult.csv) to predict if a person's earnings per annum exceed $50,000 or not, titanic dataset (whose target has passengers who survived or not), diabetes dataset (positive or negative diabetic status), Cleveland heart disease dataset, ionosphere, banknote authentication dataset (authentic or fake). Logistic Regression, k-Nearest Neighbors (KNN), decision trees (DT), support vector machine (SVM), Naive Bayes (NB) and neural networks (NN) are some notable binary classification algorithms. Unlike binary learning problems which have two class labels, multi-class learning is applied to problems involving three or more disjoint class labels. It relies on the assumption that 1) each observation is assigned to only a single label, and 2) each class label is independent of the other [6] For example, a fruit can be one of the following types; apple, mango, orange, pear, a student can graduate with only one class of degree. Iris, zoo, waveform, dermatology, sport, MNIST, ionosphere, glass and wine datasets are some of the examples of widely used multiclass datasets that are available in data repositories and widely reported in the literature. SVM, DT, multinomial logistic regression and multi-layered perceptron are suitable algorithms for multi-class tasks. Widely adopted methodologies for multi-class tasks include; 1) decomposing target label space, via the following methods; one-vs-all, all-vs-all, and error-correcting codes 2) arrangement of the classes in a tree-like structure (hierarchical method) 3) adapting and extending binary classifiers to perform multi-class classification tasks [11,14,15].

B. Multi-label Learning

In real-world scenarios, the same set of input features are often used to concurrently predict more than one target variable. The target feature may consist of binary labels, categorical or continuous values. For binary target features the type of classification is MLL while real-valued target variables are referred to as multi-target regression. However, when the target features are categorical, it becomes a multi-target prediction problem. The MLL problem is a special kind of multi-target learning (MTL) (multi-dimensional or multi-objective), where each label can be associated with more than one values, as opposed to binary labels which have two values depicting relevance(1) or otherwise(0). Recently, MLL has progressively attracted the attention of researchers especially in ML communities and has been extensively applied to solving many problems including image and video analysis, text, bioinformatics, web mining, rule mining, information retrieval, medical diagnosis and prediction and many more [16]. Techniques advanced for MLL classification problems include; algorithm adaptation approach (AAA), problem transformation methods (PTMs) [11,12,17] and ensemble methods [11,18]. The PTMs transform the original MLL problem into multiple SLL (binary or multi-class) or regression tasks while AAAs adapt the base learning algorithms themselves to solve MLL problems rather than transforming them. PTMs adopt the basic SLL classifiers to accomplish the classification task after the transformation stage and thereafter combine the results into an MLL solution. In consideration of the flexibility of the PTMs [12,17], this work performs MLL using classifier chain (CC), bayesian classifier chain (BCC), RAndom k-labEL sets (RAkEL) and Pruned Set (PS) methods and its MTL variant Nearest Set replacement (NSR).

CCs provide a means of combining several binary classifiers into a single multi-label model that is capable of exploiting correlations among targets. It is based on binary

relevance (BR) [12,17,19] approach and beats the weaknesses of BR with an improved performance in addition to the inherited strengths of BR especially low time complexity. The main idea of CC is to incorporate label dependency to BR [7,20]. The BCC [21] uses many classifiers, one per class, linked in a chain to find a joint distribution of the classes C = (C_1, C_2, . . . , C_d) given the attributes X = (x_1, x_2, . . . , x_n). In BCC settings, a CC can be constructed by firstly inducing the classifiers that do not depend on any other class and then proceed with their descendants, according to the dependence structure which can be represented as a Bayesian network. It is an alternative method for MLL that integrates class dependencies while preserving the computational proficiency of the BR technique [21]. The RAkEL algorithm repetitively constructs a cooperative group of Label Powerset (LP) classifiers. That is, it transforms a multi-label problem into one multi-class classification problem where the possible values for the transformed class attribute is a set of distinct subsets of labels present in the original training data. Each LP classifier is trained by relying on label correlations required for ranking of the labels by averaging the zero-one predictions of each model per considered label. RAkEL offers the following advantages [13]: 1) computationally less expensive due to resulting subsets of SLL tasks; 2) improvements in the class-imbalance ratio of the dataset thereby enhancing the accuracy of minority labels; 3) collation of multiple predictions for the same label by the different LP models. The PS method leverages the most significant label relationship within a multi-label dataset by eliminating insignificant and noisy label sets which might distort the performance of the classification. This reduces the complexity originating from the label dependencies without significant information loss [20,22]. The author in [20] report from experimental evidence that the PS approach outperforms LP and other baseline methods and is highly recommended for data sets with diverse concept drifts. The NSR method is the MTL version of PS where the closest sets replace outliers, rather than using subsets.

Researchers have built and used a variety of multi-labeled datasets in disparate formats and have made them available in notable multi-label data repositories including MULAN [13], Multi-label/Multi-target Extension to Weka (MEKA), Library for SVM (LibSVM) [23], Knowledge Extraction based on Evolutionary Learning (KEEL) (Alcala-Fdez *et al*, 2011) and R Ultimate Multilabel dataset repository (RUMDR), each one using two base file formats; comma-separated values (.CSV) and attribute-relation file format (.ARFF) file formats. MULAN, scikit-multi learn, MEKA and the Multi-labelled dataset in R (mldr) package provides exploratory analysis of MLL datasets. While MEKA is a general-purpose MLL software, mldr package is limited to exploratory analysis only [24]. This work therefore adopts MEKA for MLL for comparative analytics of obstetric outcome. The degree to which samples in the dataset have more than one label of datasets (multi-labelness) is estimated with two basic parameters – label cardinality (LC) (1) and Label density (LD) (2) [24]. LC indicates the mean number of labels of the records in the dataset while LD is equivalent to LC divided by the number of labels [14,24].

$$LC = \frac{1}{n}\sum_{i=1}^{n}|Y_i| \quad (1)$$

$$LD = \frac{1}{k}\frac{1}{n}\sum_{i=1}^{n}|Y_i| \quad (2)$$

Where n represents the number of samples in the dataset, Y_i the label set of the *ith* instance, and k the sum of labels in the dataset. The LC level is directly proportional to the number of active labels per sample. Several classifiers have been developed and adapted for binary, multi-class and multi-label classification problems, but there are no classifiers recommended to optimally satisfy the needs of other classification problems. This work investigates the performances of widely used classifiers on all three types of classification with a view of finding the best performing (most suitable) one.

C. Classification Approaches for Medical Diagnostic Problems

Classification is a fundamental and pivotal task of ML and data mining (DM) applications. It is encountered in various areas, such as medicine to identify a disease of a patient, prediction of the effectiveness of surgical procedures, medical tests, and the discovery of relationships among clinical and diagnosis data. The classification of health care data (HCD) for non-faulty diagnosis and appropriate prescriptions is a rising application area of DM that is grabbing the attention of researchers [25, 26]. Several works have utilized various classification methods for diseases' diagnosis and prediction. The proper utilization of classification algorithms significantly improves the analysis, disease prediction and severity level determination in addition to ensuring early detection and effective prevention mechanisms. Over the years, analysis of morbidity and mortality data in maternal-related care has evolved from the traditional to intelligent research approaches with the aim of improving the efficiency of mother and child care during pregnancy. Nonetheless, effective analytical approaches that breed intelligent decisions are dependent on the availability of reliable data collected from the healthcare domain for the purpose of extracting knowledge for informed decision-making. This process is supported by classifiers implemented in binary, multi-class or multi-label approaches. However, a universal and multi-label classification with Extreme Learning Machine (ELM) classification approach capable of performing the functions of the three aforementioned classifiers were proposed by [11] and [14], respectively. The survey conducted by [27], provided information about the association rule, classification and cluster analysis as useful tools in the identification and discovery of risk in maternal care. These tools are developed using a few underlying algorithms that have been used for mining maternal-related care, such as DT, NB, KNN, ANN, SVM, RF, Gaussian NB and so on [28-30]. ML algorithms comprising Logistic Regression (LR), SVM, DT, BPNN, XGBoost and RF, in building predictive models for early pregnancy loss after In vitro fertilization-embryo (IVF) transfer with fetal heart rate. Each of the models experimented on the features associated with on-going pregnancy and early pregnancy loss samples. RF stood out with a high performance of 97% for recall ratio, F_1 and area under the curve (AUC), in addition to an accuracy of 99% especially for those within 10

weeks after embryo transfer. In [31] MLL was performed by adapting and extending three SLL algorithms. The comparative analysis was conducted on Genbase, Yeast and Scene datasets which were evaluated in terms of LD and LC. Genbase dataset which had 27 labels, depicts greatest multi-labelness with LD of 0.05 and LC of 1.35. Four base ML algorithms (SMO, KNN, C4.5 and NB) were used to develop a predictive model which revealed SMO as the best algorithm. However, inclusion of more well-known datasets would have helped in the comparative analysis.

The author in [28] adopted the Gaussian NB classifier-based methodology with four variables obtained from INEGI. These variables were: gender, gestational age, maternal age and fetuses. The result of the classification recorded 96% accuracy in terms of precision, recall and F1-score respectively. Similarly, the NB classifier was used to compare physician-based classification for 21,000 child and adult deaths in India, South Africa and Bangladesh. This comparative study was carried out on the classifier between two different datasets without performance evaluation of any existing analytical methods. To detect gestational diabetes mellitus (GDM) in pregnant women without a visit to the hospital, a decision support system was developed based on MLP with newly designed input [50]. The identification of predictors of in-hospital maternal mortality among women attending referral hospitals in Mali and Senegal was addressed by [51]. Nonetheless, BR, LP and CC methods with different base classifiers were used for classification [12]. Although the work was limited to the phonemes of the Tamil language only, the procedure for evaluation is useful in the classification of maternal care problems. The author in [32] compared SVM and Logistic Regression (LR) to determine their performance efficiency in pregnancy outcome prediction on anonymized dataset of 420 different pregnancy details. Four output categories were defined, and the results show that the average specificity of SVM in all four categories is at least 1% higher than that for LR, except in the case of underweight infant prediction where LR had a higher specificity. On the other hand, the average sensitivity of LR was at least 10% higher than that of SVM. The study failed to compute the classification accuracies of the designed models, although LR was adjudged as a better model. The author in [49], performed a study on the cardiotocography (CTG) dataset of the University of California Irvine machine learning repository. They compared ten machine learning algorithms; focusing on their predictive precision, recall and F1 scores. Submission of the work is that during training; DT learnt better while NB had the least learning accuracy. Conversely, between the MLP, RF, SVM, and NB algorithms; the RF had the best result with an accuracy of 92%. This is followed by MLP with 84% accuracy, then 83% for the SVM classifier with linear kernel and 77% for NB. Moreover, the work reported in [33] compared the classification ability of NB, RF, DT, and SVM on the CTG dataset using the Minimum Reduction Maximum Relevance technique for feature extraction. Their measurement matric comprised of Accuracy, Precision, Recall and F1 Score. After experiments, they report that SVM had the best classification ratings followed by RF with 96%, 88.3%, 91%, and 89.3% respectively. In addition, the work did not consider the MLP classifier even though it has been widely used with interesting results in the literature for pregnancy outcome (PO) prediction. The work reported in [10] proposed an ensemble of One Dimensional Convolutional Neural Network (1DCNN) and MLP for abnormal birth outcome detection. The study performed traced segmentation on CTU-UHB intrapartum cardiotocography dataset with 552 trace observations for class distribution equalization and 1DCNN for learning and automatic feature extraction from segmented CTG data. Classification results from the proposed model were compared with SVM, RF and MLP models trained with random weight initialization. The model evaluation using sensitivity, specificity and AUC showed that the conventional MLP classifier out-performed SVM and RF in two measures, except that it had the lowest specificity. The RF algorithm on the other hand had a higher specificity (69%) and AUC (67%) scores. SVM had 68%, 56% and 62% in sensitivity, specificity and AUC respectively, at a batch size of 500. Considering the sensitivity (80%), specificity (79%) and AUC (86%), the authors concluded that models evaluated in the study failed to produce better classification results compared to the proposed ensemble 1DCNN.

III. Data Acquisition and Feature Selection

Data was acquired from secondary health facilities in Uyo, Nigeria. A total of one thousand six hundred and thirty-two (1,632) records were obtained from archives of retrospective observations of pregnant women recorded while they enrolled for antenatal care, with an input feature space of forty-two (42) features excluding the target variable. A sub-set of the attributes are; maternal age, number of children delivered, previous medical history, abortion, miscarriage, prematurity, previous illness, number of attendances to antenatal care, modal mode of delivery, antenatal registration, and mode of delivery, amongst other features. Cleaning, aggregation and pruning of attributes with only a single domain value was performed. The outcome is a dataset with thirty-five (35) input features, which were exposed to input rank analysis [34,35] via PCA in WEKA software. The selection criterion was based on eigenvalue scores not less than unity [35] regarding PO as target variable. This produced thirteen (13) significant features with a cumulative effect of 67.13%. The distribution of the variance for each factor and rank given in Table I, shows that average maternal blood pressure topped the list with an EV of 3.86 (11.7% percentage of variance), followed by average maternal weight (EV = 2.77, proportion = 8.39%). The 13th ranked attribute, average ascorbic acid level accounted for 3.17% variation with an EV score of 1.05. Target feature description of is also represented in Table I, PO consists of four Death=0) and Neonatal weight (NW) assumes low, normal or overweight as possible values.

TABLE I. RANK AND DESCRIPTION OF SIGNIFICANT INPUT ATTRIBUTES

Rank	Features	Description	EV	Proportion (%)	Cumulative (%)
1	Maternal BP	Average maternal blood pressure	3.86	11.69	11.69
2	Maternal Weight	Average maternal weight	2.77	8.39	20.29
Rank	**Features**	**Description**	**EV**	**Proportion (%)**	**Cumulative (%)**
3	Hemoglobin Level	Average number of red blood cells count	2.37	7.18	27.47
4	PCV level	Average Packed Cell Volume count	1.92	5.82	33.29
5	Pulse Rate	Average number of heart beats per minute	1.54	4.67	37.67
6	Mode of Delivery	Delivery method vaginal delivery =1; caesarean section = 2	1.42	4.30	42.26
7	Malaria Frequency	Number of times maternal malaria Diagnosis	1.39	4.21	46.47
8	Hepatitis C	Indicates history of hepatitis C disease; presence=1, absence=2	1.26	3.82	50.29
9	Diabetes Status	Maternal Diabetic status non-diabetic=0 type1=1; type2=2, others=3	1.18	3.60	53.89
10	Herbal Ingestion	Use of herbal medicinal products during pregnancy	1.15	3.48	57.37
11	Respiratory disorder	Maternal respiratory disease status; presence=1, absence=2	1.12	3.39	60.76
12	Age	Maternal age during pregnancy	1.06	3.20	63.96
13	Ascorbic acid Level	Average amount of ascorbic acid in the body during pregnancy	1.05	3.17	67.13
14	Pregnancy outcome	Maternal delivery outcome miscarriage = 0; pre-term =1; full-term=2, stillbirth=3	-	-	-
15	Maternal status	Records whether mother is alive of death Alive=1, Death=0	-	-	-
16	Neonatal weight	Weight of the newborn low=1, normal=2 overweight=3	-	-	-

IV. MATERIALS AND METHODS

A. Predictive Analytic Models

Widely used and most performing algorithms SML algorithms; NB, SVM, DT, KNN, RF and MLP classifiers are compared. The experiment aims to observe which algorithm is capable of classifying PO in all multiple classification learning scenarios.

- KNN is a supervised classification technique aimed at predicting the target variable $y \in \{1, \dots, c\}$ given a set of features $x \in \mathbb{R}^n$ [36]. It is a type of instance-based learning, or lazy learning approach in which the approximation of functions is performed locally. KNN is based on the principle of determining a fixed number of training examples closest in distance (usually Euclidean distance) to an unknown point, and predict the label from these pieces of information. Although KNN is simple, it does not require categories to be linearly separable in addition flexibility, it is computationally costly although very fast in the training phase and arduous to estimate the optimal value of k [5,15].
- NB is a classifier based on the Bayes theorem. Results from different classification and prediction studies suggests its strength and dynamism. The implementation of NB algorithm computes the posterior probability of a hypothesis given an observed data. Given an observation c_j; NB helps determine the possibility of having d as a component of c_j, using (3):

$$P\left(C_j \middle| d\right) = \frac{P(d|C_j)\,P(C_j)}{P(d)} \quad (3)$$

where $P(d/C_j)$ is the likelihood of finding d in c_j, $P(c_j)$ is the probability of the observation c_j, while $P(d)$ is the probability of observing the data, irrespective of the specified hypothesis. The NB algorithm can often outperform more sophisticated classification methods and ranks among the topmost successful algorithms for text documents classification. It implicitly assumes that all the attributes are mutually independent which violates real-world scenarios and performs poorly on data comprising highly correlated features. It exhibits greater accuracy and speed when applied to large databases, generalizes well even with limited training samples.

- SVM is a non-parametric supervised learning classifier that finds the trade-off between minimizing the training set error and maximizing the margin for optimal classification. It is known to have the best generalization ability and resistant to overfitting [37]. It is a machine learning approach efficient for solving classification and regression problems. It relies on supervised learning models which are trained by learning algorithms and is very effective when confronted with large amount of training samples to identify patterns from them. It is one of the most powerful ML algorithms for optimization, prediction and classification tasks [38,39]. Its efficiency in the prediction of weather, power output, stock market dynamics, bioinformatics, voice and handwriting recognition, image and video analysis, and medical diagnosis, among others has been demonstrated in the literature.

The major strengths of the SVM include: 1) relatively easy training and moderate scaling even with high dimensional data; 2) trade-off between the model complexity and the error are controlled easily; 3) it can handle both continuous and categorical data as well as ability to capture the nonlinear relationships in the data; 4) assumptions regarding data structure are not required because it is a non-parametric technique; 5) provides a good generalization performance with high accuracy. Some of its weaknesses include: 1) comprehensible of results to largely depends on interpretability of the input features; 2) they are computationally costly and need a good kernel function; 3) it lacks transparency in its results because it is a non-parametric method.

- **DT** is a method for approximating discrete-valued functions, in which the learned function is represented by a decision tree. Mathematically, the i^{th} C4.5 DT classifiers solve the following problem that yields the i^{th} decision function as presented in (4).

$$f_i(x) = w_i^T \emptyset(x) + b_i \tag{4}$$

$$\text{Minimize: } L(w, \xi_j^i) = \frac{1}{2}\left||w_i|\right|^2 + C\sum_{l=1}^{N}\xi_j^i$$

$$\text{Subject to: } \underset{\sim}{y_j}\,(w_i^T \emptyset(x_j) + b_i) \geq 1 - \xi_j^i, \xi_j^i \geq 0$$

$$where\ \tilde{y}_j = 1\ if\ y_j = i\ and\ \tilde{y}_j = -1\ otherwise$$

- DT adopts hierarchical design to implement the divide-and-conquer approach. It is a non-parametric technique used for both classification and regression without functional form specification. It can be directly converted to a set of simple if-then rules to enhance human comprehensibility thereby minimizing the ambiguity of complicated decisions. DTs are effective outliers and missing values detection [5]. Because of overfitting the data, additional pruning tasks (pre-pruning and post-pruning) are required, in addition to being computationally expensive. Its performance largely depends on the characteristics of the dataset.

- RF consists of a combination of classifiers where each classifier contributes with a single vote for the assignation of the most frequent class to the input vector (x) [40]. RF is an efficient model for averaging multiple deep DT that has been trained on different parts of the same training set when the goal is to reduce variance in the result. Trees constructed with fixed training data are prone to be overly adapted to the training data. The averaging function of the RF algorithm is described in (5).

$$Y = \frac{1}{N}\sum_{n=1}^{N} Y_n\,(k') \tag{5}$$

where N is the total number of trees created in random subspaces, Y_n is the classification tree, k' represent the instance to be classified, and n is a count of the sub trees which ranges from 1 to N.

- MLP consists of multiple layers of simple, bi-state, sigmoid processing nodes of neurons that interact using weighted connections [41]. The MLP classifier is a neural network that utilizes backpropagation in prediction based on threshold functions comprising a linear combination of weight, bias, and input data, as defined by (6). Each perceptron has an activation threshold; below which the perceptron is inactivated.

$$y = \psi\,(W.X + b) \tag{6}$$

where $\boldsymbol{W}$ denotes the cumulative vector of weights, **X** is the vector of cumulated inputs, $\boldsymbol{b}$ is the bias and $\boldsymbol{\psi}$ is the non-linear activation function.

B. Problem Formulation and Dataset Modelling

The dataset on maternal outcome is modeled in three main data-setups: 1) single label/single target 2) multi-label 3) multi-target. The single label/single target setup has two variants; single target binary class (ST-BC) where each observation is only associated with a single binary class label for modeling MS target attribute; and single-target multi-class (ST-MC) representation where each instance is associated with a single target with multiple class labels (PO and NW target attributes). A record may be associated with more than two binary class labels in the multi-label (MLL) data configuration while in multi-target (MTP), every label can assume many values — nominal attributes. The input vector space $X = \mathbb{R}^k$ consists of k input variables $\{X_1, X_2, \ldots, X_k\}$ representing pregnancy risk factors for PO prediction. The target feature space $Y = \mathbb{R}^m$ has m target variables, $\{Y_1, Y_2, \ldots, Y_m\}$ for the multi-target problem. An instance (x, y), where x= $\{x_1, x_2, \ldots, x_k\}$ is the input feature vector and $y = \{y_1, y_2, \ldots, y_k\}$ is the target vector, together are constituents of X and Y respectively. The input vector space is given in (7) while (8) defines the multi target arrangement.

$$\begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,k} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ x_n & x_{m,2} & \cdots & x_{n,k} \end{bmatrix} \tag{7}$$

$$\begin{bmatrix} y_{1,1} & \cdots & y_{1,m} \\ y_{2,1} & \cdots & y_{2,m} \\ \vdots & \ddots & \vdots \\ y_{n,1} & \cdots & y_{n,m} \end{bmatrix} \quad (8)$$

This paper considers three maternal outcomes (m=3) as target feature; ${Y_1}^T = \{y_{1,1}, y_{2,1}, \dots, y_{n,1}\}$ is defined with an alphabet $\mathcal{Y}_1$, $\{0,1\}^n$ and is associated with the binary-class variable Maternal Status (MS). The alphabet $\mathcal{Y}_2$, $\{1,2,3\}^n$ defines the multi-class target — neonatal weight (NW) vector, represented as ${Y_2}^T = \{y_{1,2}, y_{2,2}, \dots, y_{n,2}\}$ while the vector ${Y_3}^T = \{y_{1,3}, y_{2,3}, \dots, y_{n,3}\}$ defined with alphabet $\mathcal{Y}_3$ $\{1,2,3,4\}^n$ corresponds to PO, another multi-class target arrangement. The multi-target training vector space $D = \{(x_i, y_i)\}_{i=1}^n$ is defined in (9) while the labels, $\mathcal{Y}_i \in Y$, $where\ \mathcal{Y}_i \in \{1,2,\dots,l\}$, of target variables are given in (10) – (12). Equation 13 also depicts the multi-label structure with target vector space defined over alphabet $\mathcal{Y}_4$ $\{0,1\}^n$.

$$\begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,k} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ x_n & x_{m,2} & \cdots & x_{n,k} \end{bmatrix} \begin{bmatrix} y_{1,1} & \cdots & y_{1,m} \\ y_{2,1} & \cdots & y_{2,m} \\ \vdots & \ddots & \vdots \\ y_{n,1} & \cdots & y_{n,m} \end{bmatrix} \quad (9)$$

$$\mathcal{Y}_1 = \begin{cases} 0, & if\ an\ instance\ of\ MS\ is\ \text{"alive"} \\ 1, & if\ an\ instance\ of\ MS\ is\ \text{"death"} \end{cases} \quad (10)$$

$$\mathcal{Y}_2 = \begin{cases} 1, & if\ an\ instance\ of\ NW\ is\ \text{"}under\ weight\text{"} \\ 2, & if\ an\ instance\ of\ NW\ is\ \text{"normal"} \\ 3, & if\ an\ instance\ of\ NW\ is\ \text{"over weight"} \end{cases} \quad (11)$$

$$\mathcal{Y}_3 = \begin{cases} 1, & if\ an\ instance\ of\ PO\ is\ \text{"}miscarriage\text{"} \\ 2, & if\ an\ instance\ of\ PO\ is\ \text{"preterm"} \\ 3, & if\ an\ instance\ of\ PO\ is\ \text{"term"} \\ 4, & if\ an\ instance\ of\ PO\ is\ \text{"stillbirth"} \end{cases} \quad (12)$$

$$\mathcal{Y}_4 = \begin{cases} 1, & if\ an\ instance\ of\ a\ class\ label \\ 0, & otherwise \end{cases} \quad (13)$$

The task is to predict variants of both single-label and multi-labelled data setups. This is followed by the assessment of the weighted accuracies and computational costs of all strategies for optimal predictive power decision making in the domain of obstetric management. Table II gives the specifications of the dataset configurations. In all classification learning types, the input feature dimension is 13×1632 while the target vector for each of the SLL setting (MS, PO and NW) are column vectors. In MLL and MTL, the dimensionality of the target vector is 9 and 3 respectively, with 9 labels each. All variants of SLL setups depict LC of unity and LD of 0.5 for MS, 0.25 and 0.33 for PO and NW respectively. However, MLL and MLT have the same LC (3.00) and LD (0.33).

C. Empirical Setup

The empirical evaluation was performed on some varying experimental setups on the obstetrics outcome dataset. The different configurations were based on SLL and multi-labeled classifications types. The single labeled data configuration comprises ST-BC (where the input features are associated with one of the two class labels of the MS target) and ST-MC (where the input features are mapped to one of the more than two class labels of the PO and NW targets, respectively). All base classifiers were implemented under WEKA [42], in the SLL scenario and MEKA based frameworks [43] with the multi-labeled setting, running under Java JDK 1.7 environment. The following base classifiers: SVM, RF, DT, MLP, KNN and NB were used separately as internal classifiers in WEKA (for the ST-BC and ST-MC configurations) and MEKA (for the MLL and MTL datasets) environments. Implementations were carried out with a train/test mode of 10-fold cross validation [9] on each configuration of the dataset and repeated 20 runs with each classifier–algorithm pair on a 64bit machine of 8GB RAM size with windows 10 operating system.

The WEKA/MEKA default parameters were adopted to implement the base classifiers in both SLL and MLL settings with a batch size of 100. MLP used a learning rate of 0.3 and momentum of 0.2 while the maximum training time was 500 seconds for each iteration. There was no distance weighting associated with KNN while Linear search was used with only a single neighbor. A confidence factor of 0.25 was set for C4.5 DT. John Platt's sequential minimal optimization (SMO) algorithm was adopted for training SVM classifier with RBF Kernel function as well as epsilon value fixed at 1.0×10^{-12}. NB classifier adopted unsupervised discretization without kernel estimator. MLL and MTL setups adopted the following PTMs — classifier chains (CC), random k-label sets (RAkEL) and Bayesian classifier chains (BCC) [14] for optimality evaluations of the six base algorithms. MEKA default parameters were also adopted for the chosen PTMs and base classifiers including a batch prediction size of 100. The BCC employed CC for creating maximum spanning trees based on marginal label dependence, and NB as base classifier [21]. The RAkEL method [31] builds ensembles of Label Powerset (LP) classifiers. The training of LP classifiers relied on label correlations produced through the averaging of zero-one predictions of each model per considered label.

TABLE II. DATASET SPECIFICATIONS

Classification type		**Target feature**	**Target Vector Dimension**	**Number of Labels**	**LC**	**LD**
Single label	ST-BC	MS	1	2	1.0	0.50
	ST-MC	PO	1	4	1.0	0.25
		NW	1	3	1.0	0.33
Multi-Label	MLL	Combined (MS, PO, NW)	9	9	3.0	0.33
	MTL	Combined (MS, PO, NW)	3	9	3.0	0.33

V. Experimental Results and Discussion

A. Single Labelled Learning Results

The results for the SLL settings are presented in Tables III and IV. They represent the mean values, standard deviation (stdev) and the rank given in brackets. A rank of 1 being the highest and indicates the highest performance indicator value while a rank of 6 is the least performance rank value.

From Table III, the computed mean accuracy results show that ST-BC has the highest mean accuracy (0.950 ±0.219) and mean F1 scores (0.964±0.006). This implies that the base classifiers used in this experiment performed better in terms of accuracy and F1 score in the ST-BC configuration. In terms of classifiers, DT, RF and SVM depicts the same mean accuracy (0.964) with a slight upward variation in the *stdev* of DT. The fourth ranked classifier is MLP while NB produced the least mean accuracy (0.896 ±0.023). F1 score produced by DT in the ST-BC (0.982±0.001) is ranked the 1st while MLP yielded the smallest F1 score.

The build and test costs (Table IV), reveal that KNN is the fastest classifier with an average cost of zero during model building in all datasets, this corroborates the findings reported in [5] while MLP consumed the longest average train time in ST-BC(MS) (1.575± 0.149) and ST-MC (1.936±0.135) target setups respectively. SVM model building performance was the worst in the ST-MC (2.403±0.315).

All classifiers showed significant improvements in the testing time, DT and MLP are the top performers with average rank of 1.00 and 1.67 respectively while RF execution time was the highest time and earned a rank of 5.33. The rank of the classifiers based on accuracy and F1 score (Fig. 1) show that DT is the best ranked classifier (rank=2) in both accuracy and F1 score while SVM has a rank of 3 in both metrics. Other classifiers have an average rank greater than 3.0 in both metrics except the accuracy of RF with an average rank of 2.33. NB is the least ranked classifier based on accuracy and second lowest based on F1 score. The average rankings based on train and test time (Fig. 2) are unequal in all the classifiers. However, NB, KNN and RF ranked higher in training than testing while DT yielded the best average ranking.

TABLE III. SLL Accuracy and F1 Score (mean ± STD Deviation) and Rank (in Brackets)

Classifier	Accuracy			Average Rank
	ST-BC (MS)	***ST-MC (NW)***	***ST-MC (PO)***	
NB	0.896 ±0.023 (6)	0.807±0.027 (6)	0.784±0.025 (6)	6.0
SVM	0.964 ±0.002 (3)	0.907±0.014 (3)	0.807±0.019 (3)	3.0
kNN	0.952 ±1.195 (5)	0.909±0.067 (2)	0.738±0.059 (5)	4.0
DT	0.964 ±0.018 (1)	0.893±0.065 (4)	0.820±0.061 (1)	2.0
RF	0.964 ±0.007 (2)	0.936±0.016 (1)	0.789±0.024 (4)	2.33
MLP	0.962 ±0.067 (4)	0.887±0.016 (5)	0.813±0.022 (2)	3.67
F1 –Score				
NB	0.944±0.013 (4)	0.809±0.023 (6)	0.770±0.028 (3)	4.33
SVM	0.982± 0.001(1)	0.890±0.020 (3)	0.765±0.028 (5)	3.0
kNN	0.975±0.006(2)	0.908±0.021 (2)	0.730±0.026 (6)	3.33
DT	0.982±0.001 (1)	0.881±0.022 (4)	0.784±0.021 (1)	2.0
RF	0.955±0.011 (3)	0.930±0.019 (1)	0.766±0.026 (4)	3.33
MLP	0.943±0.004 (5)	0.865±0.021 (5)	0.774±0.026 (2)	4.33

TABLE IV. SLL Build Time and Test Time (mean ± STD Deviation) and Rank (in Brackets)

Classifier	Build Time			Average Rank
	ST-BC(MS)	**ST-MC(NW)**	**ST-MC(PO)**	
NB	0.002±0.001 (2)	0.002±0.003 (2)	0.002±0.004 (2)	2.0
SVM	0.023±0.010 (4)	1.519±0.129 (5)	2.403±0.315 (6)	5.0
kNN	0.000±0.000 (1)	0.000±0.001 (1)	0.000±0.001 (1)	1.0
DT	0.013±0.005 (3)	0.034±0.012 (3)	0.029±0.012 (3)	3.0
RF	0.300 ±0.033 (5)	0.452±0.079 (4)	0.587±0.068 (4)	4.33
MLP	1.575± 0.149 (6)	1.936±0.135 (6)	1.950 ±0.187 (5)	5.66
Test Time				
NB	0.001±0.002 (3)	0.002±0.004 (3)	0.002±0.004 (3)	3.0
SVM	0.000±0.001 (2)	0.027±0.011 (6)	0.032±0.012 (6)	4.67
kNN	0.015±0.003 (5)	0.018±0.007 (5)	0.021±0.009 (5)	5.00
DT	0.000±0.000 (1)	0.000±0.001 (1)	0.000±0.001 (1)	1.00
RF	0.009± 0.002 (4)	0.015±0.015 (4)	0.020±0.003 (4)	5.33
MLP	0.000± 0.000 (1)	0.001±0.001 (2)	0.001±0.001 (2)	1.67

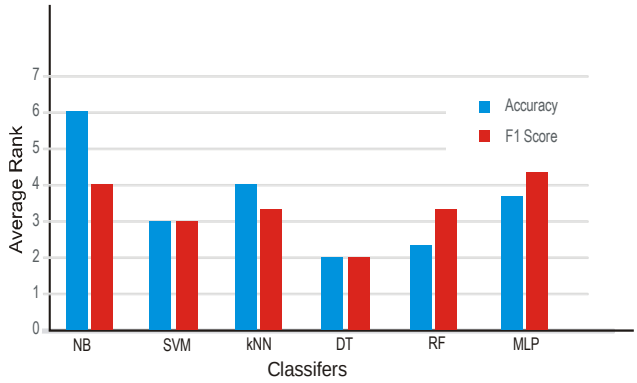

Fig. 1. Average Rank of Algorithms in SLL.

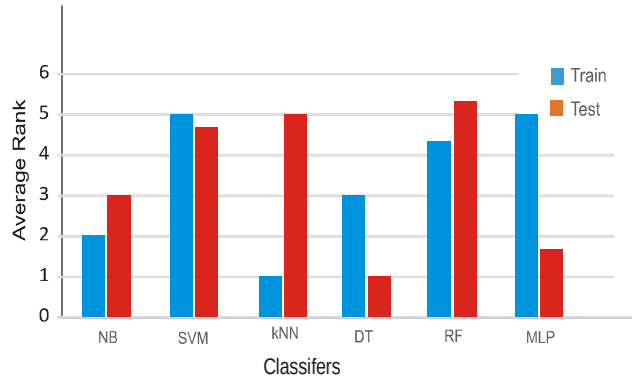

Fig. 2. Average Rank of Time Cost in SLL.

B. *Multi-labelled Learning Performance*

The distribution of average accuracy and F1 score across the PTMs and classifiers (Table V) show that NB earned the lowest accuracy and F1 score (rank=5.50) while RF produced the best performance in both Accuracy and F1 score.

It is observed that the rank of each classifier across the PTMs is the same in both metrics in addition to a marginal variation in their values. Similar results (Table VI), show that top classifies regarding accuracy earned lower ranks for time cost. Although MLP is ranked 6 with outstandingly high build time values, it competes favourably with other classifiers in the test time. KNN and DT had the best performers in the build and test times respectively while the highest execution time is exhibited by KNN followed by RF.

In the MTP scenario, Tables VII and VIII give the accuracy/F1 score and build/test time values, respectively. The F1 scores are the lowest in all dataset configurations and classification types with CC approach producing the highest average performance. The top performers are KNN, SVM and NB, in that order, and with RF having the highest average F1 score and rank of 1.25. NB earned the least rank in both accuracy (5.25) and F1 score (6.0). DT earns the highest rank (1.5) which is slightly higher than that of DT in terms of accuracy. For build and test costs, KNN utilizes an insignificant time during model build and returned as the most expensive algorithm during model execution. The reverse is the case with MLP, although the average rank of KNN is better. The ranking of DT is average in both test and build metrics while RF ranks 4.00 and 4.50 in test and build phases, respectively.

A summary of the ranks of classifiers across the datasets and classification types is given in Table IX and Fig. 3. The result shows that the ranks of classifiers in learning types varies especially between SLL and others. RF earned the best rank in MLL followed by MTP and SLL with an overall best rank of 1.78 for accuracy while depicting the worst rank in terms of time cost. DT is the second best ranked classifier regarding accuracy but is ranked the best regarding time cost while SVM is the second top classifier when considering time cost. In terms of optimality, it implies that RF is capable of producing high accuracy across dataset and classification types although is computationally expensive. This corroborates the findings reported in [9].

In term of both metrics, DT is optimal for consideration followed by KNN. It is therefore necessary to choose between RF and DT depending on the application domain and whether or not time cost should be given consideration. A cursory analysis of the result via statistical significant evaluation is presented in subsequent sections.

TABLE V. MLL ACCURACY, F1 SCORE (MEAN ± STD DEVIATION) AND RANK (IN BRACKETS)

Classifier	**Accuracy**				Ave Rank
	CC	***BCC***	***RAkEL***	***PS/NSR***	
NB	0.834±0.0182 (6)	0.83 ±0.019 (6)	0.825±0.0191 (6)	0.819±0.023 (4)	5.50
SVM	0.881±0.015 (5)	0.881±0.014 (5)	0.883±0.0148 (5)	0.88±0.015 (3)	4.50
kNN	0.889±0.0162 (4)	0.89±0.016 (4)	0.885±0.0166 (4)	0.89±0.015(2)	3.50
DT	0.896±0.0152 (2)	0 .90±0.015 (2)	0.891 ±0.015 (3)	0.89±0.015 (2)	2.25
RF	0.9192±0.013 (1)	0.92±0.013 (1)	0.915±0.0132 (1)	0.914±0.014 (1)	1.00
MLP	0.894±0.016 (3)	0.894 ± 0.16 (3)	0.893±0.015 (2)	0.89±0.015 (3)	2.50
F1 Score					
NB	0.883±0.013 (6)	0.882±0.014 (6)	0.878±0.014 (6)	0.86±0.02 (4)	5.50
SVM	0.916± 0.011 (5)	0.919±0.0103 (5)	0.92±0.0104 (5)	0.92±0.01 (3)	4.50
kNN	0.922±0.012 (4)	0.923±0.012 (4)	0.920±0.0123 (4)	0.92 ±0.01 (3)	3.75
DT	0.928±0.011 (3)	0.93±0.010 (2)	0.927±0.010 (3)	0.92 ±0.01 (3)	2.75
RF	0.944±0.009 (1)	0.945±0.009 (1)	0.942 ±0.010 (1)	0.94±0.010 (1)	1.00
MLP	0.93±0.011 (2)	0.927±0.011 (3)	0.927±0.011 (2)	0.93±0.010 (2)	2.25

TABLE VI. MLL BUILD AND TEST TIMES (MEAN ± STD DEVIATION) AND RANK (IN BRACKETS)

Classifier	Build Time				Ave RanK
	CC	***BCC***	***RAkEL***	***PS/NSR***	
NB	0.060±0.011 (2)	0.061±0.012 (2)	0.101±0.019 (2)	0.011±0.008 (2)	2.00
SVM	1.134±0.36 (4)	1.07±0.229 (4)	6.790 ±1.801 (4)	7.53 ±2.65 (5)	4.25
kNN	0.022±0.001 (1)	0.025±0.008 (1)	0.059±0.018 (1)	0.005±0.004 (1)	1.00
DT	0.24±0.037 (3)	0.34 ±0.062 (3)	0.847±0.13 (3)	0.140±0.028 (3)	3.00
RF	3.927±0.52 (5)	4.59±0.77 (5)	14.76±7.07 (5)	1.415±0.140 (4)	4.75
MLP	85.77±12.16 (6)	67.28 ±8.71 (6)	123.14±17.60 (6)	34.91±5.05 (6)	6.00
	Test Time				
NB	0.023±0.001 (4)	0.021±0.005 (4)	0.083±0.0157 (4)	0.029±0.006 (3)	3.75
SVM	0.008±0.015 (2)	0.005±0.003 (2)	0.025±0.024 (3)	0.047±0.05 (5)	3.00
kNN	0.913±0.13 (6)	0.756±0.120 (6)	0.90±0.2059 (6)	0.077±0.01 (6)	6.00
DT	0.002±0.002 (1)	0.002 ±0.00 (1)	0.006±0.0021 (1)	0.003±0.002 (1)	1.00
RF	0.172±0.032 (5)	0.213±0.035 (5)	0.70 ±0.37 (5)	0.039±0.010 (4)	4.75
MLP	0.012±0.004 (3)	0.009±0.004 (3)	0.0167±0.010 (2)	0.004±0.002 (2)	2.50

TABLE VII. MTP ACCURACY, F1 SCORE (MEAN ± STD DEVIATION) AND RANK (IN BRACKETS)

Classifier	Accuracy				Ave Rank
	CC	***BCC***	***RAkEL***	***PS/NSR***	
NB	0.815±0.018 (6)	0.816±0.017 (4)	0.814±0.0207 (6)	0.814±0.020 (5)	5.25
SVM	0.880±0.013 (3)	0.88±0.014 (2)	0.88 ±0.014 (3)	0.88 ±0.013 (2)	2.75
kNN	0.867±0.014 (4)	0.87±0.014 (3)	0.863±0.013 (5)	0.86 ±0.014 (4)	4.00
DT	0.892 ±0.013 (2)	0.89±0.013 (1)	0.89±0.013 (1)	0.88 ± 0.013 (2)	1.50
RF	0.894 ±0.012 (1)	0.89±0.013 (1)	0.89±0.012 (2)	0.88 ±0.012 (3)	1.75
MLP	0.885±0.013 (3)	0.89±0.013 (1)	0.88±0.013 (4)	0.88 ±0.0135 (1)	2.25
F1 score					
NB	0.600 ±0.032 (6)	0.60±0.032 (6)	0.58±0.035 (6)	0.59±0.036 (6)	6.00
SVM	0.69±0.032 (4)	0.69±0.033 (4)	0.69±0.033 (4)	0.69±0.033 (4)	4.00
kNN	0.648±0.032 (5)	0.65±0.03 (5)	0.64±0.032 (5)	0.64 ±0.032 (5)	5.00
DT	0.715±0.034 (2)	0.72±0.04 (1)	0.702±0.031 (3)	0.70 ±0.035 (2)	1.75
RF	0.716±0.031 (1)	0.72±0.028 (2)	0.704±0.031 (1)	0.705±0.030 (1)	1.25
MLP	0.701±0.031 (3)	0.70±0.03 (3)	0.703±0.032 (2)	0.70 ±0.034 (3)	2.75

TABLE VIII. MTP BUILD AND TEST COSTS (MEAN ± STD DEVIATION) AND RANK (IN BRACKETS)

Classifier	Build Time				Ave. Rank
	CC	***BCC***	***RAkEL***	***PS/NSR***	
NB	0.024 ±0.008 (2)	0.032±0.025 (2)	0.014±0.022 (2)	0.01±0.0047 (2)	2.00
SVM	1.044±0.221 (4)	0.95±0.22 (4)	6.85 ±2.414 (5)	7.141 ± 2.59 (5)	4.50
kNN	0.009±0.004 (1)	0.010±0.004 (1)	0.006±0.003 (1)	0.004 ±0.003 (1)	1.00
DT	0.26 ± 0.050 (3)	0.242±0.046 (3)	0.149 ±0.032 (3)	0.139 ±0.031 (3)	3.00
RF	3.162 ± 0.52 (5)	3.60±0.911 (5)	2.180±0.43 (4)	2.048 ± 0.33 (4)	4.50
MLP	27.89 ±3.78 (6)	25.70±3.41 (6)	35.07 ±4.92 (6)	348.40±15.92 (6)	6.00
Test Time					
NB	0.012±0.004 (4)	0.014±0.003 (4)	0.029±0.006 (3)	0.034±0.007 (3)	3.5
SVM	0.003±0.008 (3)	0.004±0.010 (3)	0.049 ±0.017 (5)	0.059 ±0.049 (4)	3.75
kNN	0.34 ±0.062 (6)	0.28±0.046 (6)	0.036±0.034 (4)	0.078±0.015 (6)	5.5
DT	0.001±0.001 (1)	0.001±0.002 (1)	0.073±0.015 (6)	0.007±0.003 (1)	2.25
RF	0.114 ±0.026 (5)	0.131±0.039 (5)	0.002± 0.002 (1)	0.064±0.059 (5)	4.00
MLP	0.003 ±0.001 (2)	0.003±0.001 (2)	0.004±0.0013 (2)	0.0084±0.003 (2)	2.00

TABLE IX. AVERAGE RANKINGS (AR) OF ALGORITHMS OVER TWO METRICS AND CLASSIFICATION TYPES

	Accuracy/F1 score					Build and Test Time			
Classifier	***SLL AR***	***MLL AR***	***MTP AR***	***Global AR***		***SLL AR***	***MLL AR***	***MTP AR***	***Global AR***
NB	5.17	5.50	5.63	5.43		2.50	2.88	2.75	2.71
SVM	3.00	4.50	3.38	3.63		4.84	3.63	4.13	4.20
kNN	3.67	3.63	4.50	3.93		3.00	3.50	3.25	3.25
DT	2.00	2.50	1.63	2.04		2.00	2.00	2.63	2.21
RF	2.83	1.00	1.50	1.78		4.83	4.75	4.25	4.61
MLP	3.84	2.38	2.5	2.91		3.34	4.25	4.00	3.86

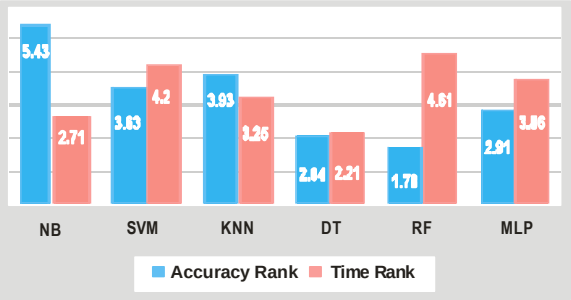

Fig. 3. Global Rank of Classifiers based on Accuracies and Time Cost.

C. Statistical Significance and Rank Validation

The main goal is to ascertain if there is any base classifiers whose performance is significantly different from others and also perform multiple comparison analysis. This was achieved by implementing non-parametric procedures [44,45] individually to each of the four categories of dataset-target setups for informed statistical inferences. Friedman test — a non-parametric variant of the repeated-measures Analysis of Variance, was used to test the null hypothesis that there is no significant difference in the performances (accuracies and time costs) of the classifiers. It compares the average rankings of the six classifiers across each of the four dataset configurations, calculating test statistic which estimates the probability of the observed rankings under the null hypothesis. Nemenyi's test and Bergmann-Hommel's post-hoc procedures implemented in R produced pairwise comparisons of all algorithms. The results are presented in the following sub-sections.

1) SLL Analysis : Friedman test on the performances of the classifiers reveals that there was no statistically significant difference in the accuracies (χ^2=10.071, df =5, p=0.0732) and time cost (χ^2=8.8571, df =5, p=0.1149) of the six classifiers at 95% confidence level (CL). This implies that the null hypothesis that there is no statistically significant difference between performances of classifiers in terms of accuracies and time cost for the SLL dataset setups is accepted. Nemenyi test (Fig. 4) compared all classifiers to each other and obtained the critical difference (CD) value of 3.2853 for both accuracies and time. As shown in Fig. 4, none of the distances separating any two classifiers in terms of their accuracy and time is greater than the CD value, this confirms that the performance of every pair of classifiers is not statistically different. In both cases, DT is the best performing classifier while RF has an average rank (AR) of 2.67 and 4.33 on accuracy and time cost respectively. Although, NB has the lowest accuracy value with an average rank of 5.17 it earned an AR of 2.67 for cost, while SVM is the most computationally expensive classifier in the SLL scenario.

2) MLL Analysis: The results of accuracies (χ^2 = 36.464, df = 5, $p = 7.67 \times 10^{-7}$) and time cost (χ^2 = 10.929, df = 5, p=0.05281) for MLL target configurations signify the existence of statistically significant difference in accuracies of classifies while the average time used by each classifier does not vary significantly at 95% CL. The CD=2.7924 (Fig. 5) is returned for both accuracy and time cost. The top three performing algorithms regarding accuracy; RF, MLP and DT, do not depict statisticaly significant difference between each other while the bottom performing classifiers kNN, SVM and NB are statistically similar. NB is lowest ranked classifier in terms of classification accuracy and is significantly different from values produced by RF, MLP and DT since their respective difference in length is greater than CD (2.7924).

Although RF is the best performing algorithm as evidence by its accuracy, it is the most time consuming algorithm with an AR of 4.75 while DT consumed the smallest amount of time in all dataset configurations, followed by NB.

3) MTL Analysis : In MTL setting, the comparison of the differences in the performance accuracy of the classifiers is statistically significant at a CL of 95% while the time costs across classifiers, statistically, does vary significantly. This is as indicated by their respective p-values and chi-squared values regarding accuracy (χ^2 = 35.125, df = 5, $p = 1.42 \times 10^{-6}$) and time (χ^2 = 5.9107, df = 5, $p = 0.315$). The CD diagram (Fig. 6), depicts the results of Nemenyi test showing the statistical comparison of all classifiers against each other by ARs based on accuracy and time. Classifiers that are not connected by a bold line of length equal to CD have significantly different ARs at 95% CL. In the case of accuracy, the values of NB are significantly different from RF, DT and MLP respectively. RF has the highest AR (1.44) followed by DT (2.12) and MLP (2.69) using accuracy while DT (2.62) and RF(4.25) stand out as the best and worst algorithms respectively when considering computation time.

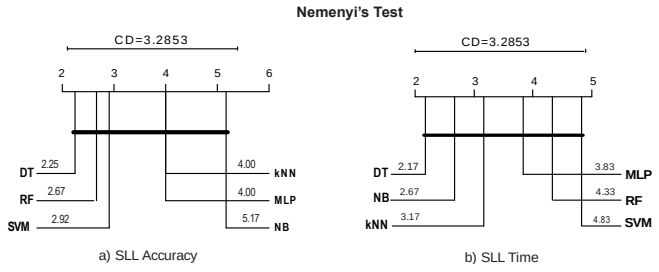

Fig. 4. CD for *Nemenyi* Test at $\alpha = 0.05$ for a) SLL Accuracy b) SLL Time.

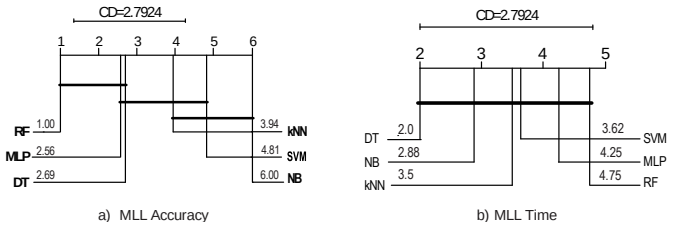

Fig. 5. CD Diagram for *Nemenyi* Test ($\alpha = 0.05$) *a) MLL Accuracy b) MLL Time.*

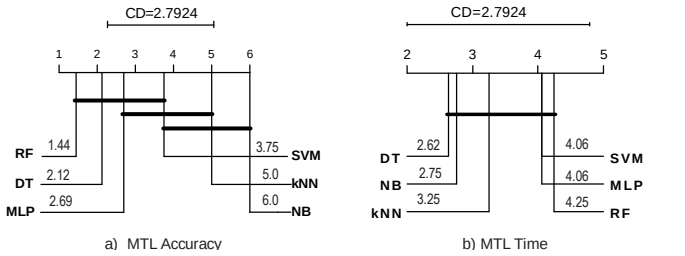

Fig. 6. CD Diagram for Nemenyi Test ($\alpha = 0.05$) a) MTL Accuracy b) MTL Time.

4) Multiple Comparison of Classifiers on all targets setups: Results of multiple comparison analysis on the combined accuracies and time costs obtained from the classifier in four dataset settings are discussed in this section. The Friedman test on aggregated values of the adopted metrics produces accuracy values (χ^2 = 70.019, df = 5, p=1.01×10^{-13}) and time values (χ^2 = 23.123, df = 5, p=3.197×10^{-4}) which depicts a statistically significant difference in performance metrics at $\alpha = 0.05$ significance level. The CD diagram (Fig. 7) obtained from the comparisons for accuracy and time, shows that the accuracy of NB significantly differs from accuracies of other classifiers while the performance of KNN differs significantly from DT and RF. The accuracy of SVM is however equivalent to others except R F and NB. RF is the highest ranked (1.61) and best performing algorithm based on accuracy followed by DT (2.36). MLP earned an AR of 3.0 and returned as the third ranking classifier while the accuracy of NB is the worst. In terms of time cost (Fig. 7b), the worst performing classifier is the RF with an AR of 4.45 and is similar to the accuracies of other classifiers except for NB and DT. DT is best classifier in terms of computational cost closely followed by NB and KNN. This implies that RF yields the highest accuracy across all classification types (dataset configuration) while it is the most computationally expensive algorithm.

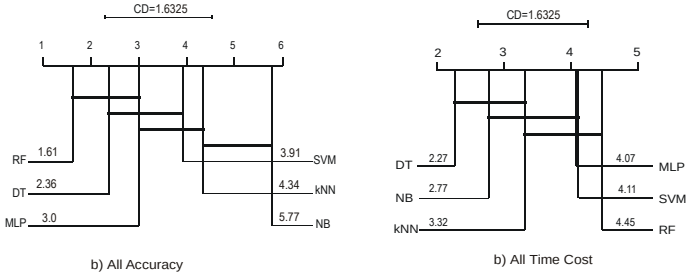

Fig. 7. CD for Nemenyi Test at $\alpha = 0.05$ for a) Accuracy b) Time.

The obtained p-values from the Freidman test specify that the null hypothesis (that all the algorithms perform the same) is reject. This, therefore, serves as the justification for conducting the post-hoc test. Bergmann–Hommel's test procedure is the most powerful, best performing, and most suitable when the number of algorithms is less than nine (9) [46-48], although it is complex and computationally expensive. Statistical pairwise comparison of the six algorithms based on average accuracies and time cost are given in Table X.

As shown in Table X, there are four major heterogeneous pairwise groupings of classifiers based on accuracy, with RF and DT being outstanding and individually significantly different from the rest of the classifiers, except MLP while NB depicts a statistically significant difference from all other classifiers. KNN-SVM and RF-DT pairs, each produced a ρ-value > 0.05, therefore statistically equivalent.

The time cost of RF is significantly different from DT (ρ<0.05) and NB (ρ<0.05) while statistically equivalent with MLP and SVM (ρ=1.0). Although the time used by DT is not statistically different from that of KNN (ρ=0.383), it exhibits a significant difference when compared with MLP (ρ=0.0110) and SVM (ρ=0.0110) in addition to RF. Pairwise comparisons involving KNN yielded no statistically significant difference as well as SVM compared with RF and MLP respectively. The summary of the Bergmann–Hommel's corrected average values (accuracy and time) of each algorithm over all the dataset is given in Table XI and Fig. 8. The results confirm that RF (accuracy=87.3%) is the best performing algorithm followed by DT (accuracy=86.3%) based on accuracy metrics while NB is the least expensive algorithm across all dataset and classification types. The ranking of classifiers considering both performance metric reveals DT (rank=2.0) as the best optimal performing classifier followed by RF (rank=3.0) while MLP (rank=4.5) depicts the worst performance.

TABLE X. CORRECTED P-VALUES USING BERGMANN–HOMMEL'S PROCEDURE FOR ACCURACY (α =0.05)

S/N	Hypothesis	ρ-value	
		Accuracy	*Time*
1	RF vs. NB	2.5×10^{-12}	2.87×10^{-2}
2	RF vs KNN	1.33×10^{-5}	3.08×10^{-1}
3	RF vs SVM	3.3×10^{-4}	1.00
4	RF vs DT	3.67×10^{-1}	1.60×10^{-3}
5	DT vs. NB	1.51×10^{-8}	1.00
6	DT vs. KNN	2.74×10^{-3}	3.83×10^{-1}
7	DT vs. SVM	2.45×10^{-2}	1.10×10^{-2}
8	NB vs. MLP	6.20×10^{-6}	1.05×10^{-1}
9	NB vs. SVM	5.72×10^{-3}	1.05×10^{-1}
10	NB vs. KNN	4.46×10^{-2}	1.00
11	MLP vs. RF	5.59×10^{-2}	1.00
12	MLP vs. KNN	6.98×10^{-2}	6.34×10^{-1}
13	MLP vs. SVM	2.14×10^{-1}	1.00
14	MLP vs. DT	5.18×10^{-1}	1.10×10^{-2}
15	KNN vs. SVM	5.19×10^{-1}	6.34×10^{-1}

TABLE XI. BERGMANN–HOMMEL'S GLOBAL AVERAGE VALUES (α =0.05)

S/N	Algorithm	Accuracy	Time	Global AR
1	NB	0.793	**0.026**	**3.5**
2	SVM	0.854	1.69	4
3	KNN	0.840	0.163	4
4	DT	**0.863**	**0.115**	**2**
5	RF	**0.873**	1.75	3
6	MLP	0.858	34.26	4.5

Fig. 8. Bergmann–Hommel's Global Rank of Classifiers based on Time and Accuracy Scores (α=0.05).

VI. CONCLUSION

Over the years, analysis of morbidity and mortality data in maternal-related care evolved from traditional to intelligent research approaches with the aim of improving the efficiency of mother and child care during pregnancy. For intelligent automated predictive solutions, ML and statistical approaches have been the most popular techniques in the literature; following the increasing clinical and administrative interest in PO determination. Results from both methods have contributed to the research of PO prediction, preconception counseling, antenatal assessment, intrapartum care, postpartum management, and reproductive health education among others. In this paper, six ML-based classifiers, including SVM, RF, DT, MLP, KNN and NB were identified as widely used and highly successful in obstetric outcome prediction. The performances and suitability of these techniques on obstetrics dataset classification under varying maternal outcome target configurations were assessed, positing that they comprise binary, multi-class and multi-labeled target features. Performance efficiency was achieved by empirical evaluation of implemented non-parametric procedures individually for SLL, MLL and MTP to enable informed statistical inferences. Using SLL, three configurations including MS, PO and NW were defined, whereas the MLL and MTP evaluations both used the CC, BCC, RAkEL, PS/NSR PMTs to evaluate performance efficiency. Dataset obtained from archives of secondary healthcare facilities in Uyo, Nigeria, was reduced feature dimension of 13 x 1632. From the results, in the SLL setup, DT had the best accuracy, F1 score and test time with an average rank of 1.0. This was followed by RF in accuracy and SVM in F1 score, while MLP had the second best time cost. NB had the worst accuracy and F1 values, while the worst test time is observed in RF. In MLL, we observed DT was least expensive in terms of time cost; whereas KNN was most expensive. RF performed better with the highest accuracy and F1 scores and was followed by DT and MLP for accuracy and F1 measures, respectively. The accuracy and F1 values obtained for NB suggests that it is the least performing classifier with the MLL setup. With an average rank of 1.50, DT had the highest accuracy in the MTP setup. This was followed by RF, while NB had the worst performance. For F1-measure evaluation, RF, DT and NB had the best, second and least performances respectively. The comparative analysis of global averages of the six base classifiers shows that RF is the most optimal algorithm with an accuracy of 87.3% given all three data setups in the study. The pole position of RF in terms of accuracy measure is in agreement with the submission in [49] (Hoodbhoy et al., 2019) that compared ten machine learning algorithms on PO determination and observed RF had an accuracy of 92% compared to lower scores obtained by MLP, SVM and NB. It also corresponds with the result obtained in [33] where the accuracy of RF was best with a score of 96%, and the work of [9]. In terms of time cost, NB is the least expensive algorithm even though it has the poorest global accuracy score. MLP on the other hand had an unexpectedly high time cost, making it unsuitable for similar data classification if time is the main criterion. Finally, from the comparative analysis, it is recommended that the choice of classifier should either be RF or DT depending on the application domain and whether or not time cost is a major consideration. As further research, the tuning of parameters of the base classifiers using evolutionary computing would be carried out in order to improve performance in terms of accuracy and computational cost.

ACKNOWLEDGMENTS

The funding for this research was provided by Tertiary Education Trust Fund (TETFUND), Nigeria through the Centre of Excellence in Computational Intelligence Research, University of Uyo. The authors would like to appreciate TETFund and the management of the University of Uyo for providing an enabling environment to carry out this research.

REFERENCES

[1] Soofi, Aized Amin, and Arshad Awan. "Classification techniques in machine learning: applications and issues." Journal of Basic and Applied Sciences 13 (2017): 459-465.

[2] Chegini, Mohammad, Jürgen Bernard, Philip Berger, Alexei Sourin, Keith Andrews, and Tobias Schreck. "Interactive labelling of a multivariate dataset for supervised machine learning using linked visualisations, clustering, and active learning." Visual Informatics 3, no. 1 (2019): 9-17.

[3] Inyang, Udoinyang G., Uduak A. Umoh, Ifeoma C. Nnaemeka, and Samuel A. Robinson. "Unsupervised Characterization and Visualization of Students' Academic Performance Features." Computer and Information Science 12, no. 2 (2019): 103-116.

[4] Ekpenyong, Moses, Udoinyang Inyang, and EmemObong Udoh. "Unsupervised visualization of Under-resourced speech prosody." Speech Communication 101 (2018): 45-56.

[5] Mohamed, A. E. "Comparative study of four supervised machine learning techniques for classification." International Journal of Applied 7, no. 2 (2017).

[6] Silva-Palacios, Daniel, Cesar Ferri, and María José Ramírez-Quintana. "Improving performance of multiclass classification by inducing class hierarchies." Procedia Computer Science 108 (2017): 1692-1701.

[7] Ceylan, Zeynep, and Ebru Pekel. "Comparison of multi-label classification methods for pre-diagnosis of cervical cancer." graphical models 21 (2017): 22.

[8] Lashari, S A , Rosziati I , Norhalina Senan, and N. S. A. M. Taujuddin. "Application of Data Mining techniques for medical data classification: A review." In MATEC Web of Conferences, vol. 150, p. 06003. EDP Sciences, 2018.

[9] Inyang, U. G., Osang, F., Eyoh, I.J, Afolorunso, A. A., and Nwokoro, C.O., "Comparative Analytics of Classifiers on Resampled Datasets for Pregnancy Outcome Prediction" International Journal of Advanced Computer Science and Applications(IJACSA), 11(6), 2020. 494-504 http://dx.doi.org/10.14569/IJACSA.2020.0110662.

[10] Fergus P, Chalmer C, Montanez C.C, Reilly D, Lisboa P and Pineles (2019). Modeling Segmented Cardiotocography Time-Series Signals Using One-Dimensional Convolutional Neural Networks for the Early Detection of Abnormal Birth Outcomes. IEEE Transactions in Emerging Topics in Computational Intelligence, arXiv: 1908.02338.

[11] Er, Meng Joo, Rajasekar Venkatesan, and Ning Wang. "An online universal classifier for binary, multi-class and multi-label classification." In 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 003701-003706. IEEE, 2016.

[12] Pushpa, M., and S. Karpagavalli. "Multi-label classification: Problem transformation methods in Tamil phoneme classification." Procedia computer science 115 (2017): 572-579.

[13] Tsoumakas, Grigorios, Ioannis Katakis, and Ioannis Vlahavas. "Random k-labelsets for multilabel classification." IEEE Transactions on Knowledge and Data Engineering 23, no. 7 (2010): 1079-1089.

[14] Venkatesan, R., and Er, M. J. (2014, December). Multi-label classification method based on extreme learning machines. In 2014 13th International Conference on Control Automation Robotics & Vision (ICARCV) (pp. 619-624). IEEE.

[15] Chaitra, P. C., & Kumar, D. R. S. (2018). A review of multi-class classification algorithms. International Journal of Pure and Applied Mathematics, 118(14), 17-26.

[16] Zhang, Min-Ling, and Zhi-Hua Zhou. "A review on multi-label learning algorithms." IEEE transactions on knowledge and data engineering 26, no. 8 (2013): 1819-1837.

[17] Cherman, Everton Alvares, Maria Carolina Monard, and Jean Metz. "Multi-label problem transformation methods: a case study." CLEI Electronic Journal 14, no. 1 (2011): 4-4.

[18] Madjarov, G, Kocev, D., D. Gjorgjevikj, and S. Džeroski, "An extensive experimental comparison of methods for multi-label learning, Pattern Recognition, vol. 45, pp. 3084-3104, 2012.

[19] Zhang, Min-Ling, Yu-Kun Li, Xu-Ying Liu, and Xin Geng. "Binary relevance for multi-label learning: an overview." Frontiers of Computer Science 12, no. 2 (2018): 191-202.

[20] Júnior, Joel D. Costa, Elaine R. Faria, Jonathan A. Silva, João Gama, and Ricardo Cerri. "Pruned Sets for Multi-Label Stream Classification without True Labels." In 2019 International Joint Conference on Neural Networks (IJCNN), pp. 1-8. IEEE, 2019.

[21] Zaragoza, J. C., Enrique Sucar, Eduardo Morales, Concha Bielza, and Pedro Larranaga. "Bayesian chain classifiers for multidimensional classification." In Twenty-second international joint conference on artificial intelligence. 2011.

[22] Read, Jesse, Bernhard Pfahringer, and Geoff Holmes. "Multi-label classification using ensembles of pruned sets." In 2008 eighth IEEE international conference on data mining, pp. 995-1000. IEEE, 2008.

[23] Chang, C.C., and Lin, C.J.: Libsvm: a library for support vector machines. ACMTrans. Intell. Syst. Technol. 2(3), 1–27 (2011).

[24] Charte, Francisco, María J. del Jesus, and Antonio J. Rivera. Multilabel classification: problem analysis, metrics and techniques. Springer, 2016.

[25] Tarle, Balasaheb, Rupali Tajanpure, and Suderson Jena. "Medical data classification using different optimization techniques: A survey." International Journal of Research in Engineering and Technology (IJRET) 5 (2016): 101-108.

[26] Umoh, Uduak A., and Udoinyang G. Inyang. "A FuzzFuzzy-Neural Intelligent Trading Model for Stock Price Prediction." International Journal of Computer Science Issues (IJCSI) 12, no. 3 (2015): 36.

[27] Mehta, R., Bhatt, N., & Ganatra, A. (2016). A survey on data mining technologies for decision support system of maternal care domain. International Journal of Computers and Applications, 138(10), 20-4.

[28] Jurado, I. C., Camarillo, D. R., & Acevedo, E. S. (2020). Problems in pregnancy, modeling fetal mortality through the Naïve Bayes classifier. International Journal of Combinatorial Optimization Problems and Informatics, 11(3), 121-129.

[29] Mathew, N. (2018, April). A Boosting Approach for Maternal Hypertensive Disorder Detection. In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) (pp. 1474-1477). IEEE.

[30] Babu, T. A., & Kumar, P. R. (2018, January). Characterization and classification of uterine magnetomyography signals using KNN classifier. In 2018 Conference on Signal Processing and Communication Engineering Systems (SPACES) (pp. 163-166). IEEE.

[31] Tsoumakas, G, and Ioannis Vlahavas. "Random k-labelsets: An ensemble method for multilabel classification." In European conference on machine learning, pp. 406-417. Springer, Berlin, Heidelberg, 2007.

[32] Guidi G, Adembri G, Vannuccini S and Iadanza E (2014). Predictability of some Pregnancy Outcomes Based on SVM and Dichotomous Regression Techniques. IWAAL. Springer International Publishing, Switzerland. Pp.:163 – 166.

[33] Jayashree J, Harsha T, Anil K.C and Vijayashree (2020). Enhanced Optimal Feature Selection Techniques for Fetal Risk Prediction Using Machine Learning Algorithms, Int'l Journal of Engineering & Advanced Technology, 9(3), Pp.:4364–4370. doi: 10.35940/ijeat.C6502.029320.

[34] Inyang, U. G., and Akinyokun, O. C. "A hybrid knowledge discovery system for oil spillage risks pattern classification." Artificial intelligence Research 3(4), (2014): 77-86.

[35] Akinyokun, O. C., and Inyang, U. G. "Experimental study of neuro-fuzzy-genetic framework for oil spillage risk management." Artif. Intell. Research 2(4), (2013): 13-36.

[36] Losing, V., Hammer, B., & Wersing, H. (2016, December). KNN classifier with self adjusting memory for heterogeneous concept drift. In 2016 IEEE 16th international conference on data mining (ICDM) (pp. 291-300). IEEE.

[37] Chen, H. L., Yang, B., Liu, J., & Liu, D. Y. (2011). A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. Expert systems with applications, 38(7), 9014-9022.

[38] Bhavsar H and Panchal H.M (2012), A Review on Support Vector Machine for Data Classification, International Journal of Advanced Research in Computer Engineering & Technology, 1(10), Pp.: 185 – 189.

[39] Meyer D, Leisch F, and Hornik K (2003). The Support Vector Machine Under Test, Neurocomputing, 55(2), Pp.: 169–186.

[40] Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS Journal of Photogrammetry and Remote Sensing, 67, 93-104.

[41] Pal, S. K., & Mitra, S. (1992). Multilayer perceptron, fuzzy sets, classification.

[42] Read, Jesse, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. "Classifier chains for multi-label classification." Machine learning 85, no. 3 (2011): 333.

[43] Read, Jesse, Peter Reutemann, Bernhard Pfahringer, and Geoff Holmes. "Meka: a multi-label/multi-target extension to weka." The Journal of Machine Learning Research 17, no. 1 (2016): 667-671.

[44] Gardner, J., & Brooks, C. (2017, April). A statistical framework for predictive model evaluation in MOOCs. In Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale (pp. 269-272).

[45] Janicka, Małgorzata, Mateusz Lango, and Jerzy Stefanowski. "Using information on class interrelations to improve classification of multiclass imbalanced data: A new resampling algorithm." International Journal of Applied Mathematics and Computer Science 29, no. 4 (2019): 769-781.

[46] Górecki, T., & Łuczak, M. (2017). Stacked Regression with a Generalization of the Moore-Penrose Pseudoinverse. Statistics in Transition, 18(3), 443.

[47] Calvo, B., & Santafé Rodrigo, G. (2016). scmamp: Statistical comparison of multiple algorithms in multiple problems. The R Journal, Vol. 8/1, Aug. 2016.

[48] García, S., Fernández, A., Luengo, J., & Herrera, F. (2010). Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. Information sciences, 180(10), 2044-2064.

[49] Hoodbhoy Z, Noman M, Shafique A, Nasim A, Chowdhury D and Hasan B (2019). Use of Machine Learning Algorithms for Prediction of Fetal Risk Using Cardiotocographic Data. International Journal of Applied and Basic Medical Research, Vol. 9, Pp.: 226 – 230. doi:10.4103/ijabmr.IJABMR_370_18.

[50] Muller, P. S., Sundaram, S. M., Nirmala, M., & Nagarajan, E. (2015). Application of computational technique in design of classifier for early detection of gestational diabetes mellitus. Applied Mathematical Sciences, 9(67), 3327-3336.

[51] Ndour, Cheikh, Simplice Dossou Gbété, Noelle Bru, Michal Abrahamowicz, Arnaud Fauconnier, Mamadou Traoré, Aliou Diop, Pierre Fournier, and Alexandre Dumont. "Predicting in-hospital maternal mortality in Senegal and Mali." PloS one 8, no. 5 (2013): e64157.