UNIVERSITY OF CAPE TOWN

DEPARTMENT OF COMPUTER SCIENCE

Computer Science Honours FIRST DRAFT Paper

2017

Title: A Comparison of Machine Learning Techniques for Symptom Prediction and Consultation Defaulter Prediction on an Imbalanced Malawian Clinical Dataset

Author: Shaaheen Sacoor

Project Abbreviation: ML4H

Supervisor: Dr. Brian DeRenzi

Category	Min	Max	Chosen
Requirement Analysis and Design	0	20	0
Theoretical Analysis	0	25	0
Experiment Design and Execution	0	20	20
System Development and Implementation	0	15	5
Results, Findings and Conclusion	10	20	20
Aim Formulation and Background Work	10	15	15
Quality of Paper Writing and Presentation	10		10
Adherence to Project Proposal and Quality of Deliverables	10		10
Overall General Project Evaluation (this section allowed only with motivation letter from supervisor)	0	10	0
Total marks		80	

A COMPARISON OF MACHINE LEARNING TECHNIQUES FOR SYMPTOM PREDICTION AND CONSULTATION DEFAULTER PREDICTION ON AN IMBALANCED MALAWIAN CLINICAL DATASET

Shaaheen Sacoor
Computer Science Department, UCT
SCRSHA001@myuct.ac.za

ABSTRACT

We compare the Machine Learning discriminatory power of Logistic Regression, K Nearest Neighbours, Decision Tree Classifier, Gaussian Naïve Bayes, Random Forest, Multi-layer perceptron, Adaptive Boosting and Support Vector Machine on the problem of predicting a patient’s next symptom (symptom prediction problem) and on the problem of predicting if a patient will miss their next medical consultation (consultation defaulter problem). With the ROC metric being used to measure discriminatory power, the Decision Tree Classifier came out as the peak performer for the symptom prediction problem by obtaining a ROC value of 74%. This is after the Near Miss balancing technique was applied to the problem’s imbalanced result classes. The symptom prediction problem thus produced fair results but not significant enough to be considered for the high requirements of a real world medical application yet. The consultation defaulter problem experiment deduced that the highest performing technique in the problem approach was the Support Vector Machine, which obtained a ROC metric value of 84.9% on the All-KNN technique balanced dataset. This result is competitively significant and could be considered for real world medical application use.

CCS CONCEPTS

• Outline of machine learning → Supervised Learning; • Applied Computing → Computational healthcare

KEYWORDS

Machine Learning, Clinical Resource Allocation, Electronic Health Record, Prediction modelling

1. Introduction

Malawi experiences a patient-to-doctor ratio at the alarming level of 50 000:1 [14]. This, when compared to the USA’s 375:1 ratio and the World Health Organisation’s minimum suggested ratio of 450:1, is significantly below the reasonable health care standard for number of medical professionals available for people in a country [47]. Along with this, the Malawian health sector also suffers from shortages of vital drugs being available at health facilities. The appropriate amount of vital drug stocks was only held at 27% of all health centres in Malawi according to Mueller et. al’s study [37]. These ratios emphasize the extent of the limited medical resources available in Malawi. In addition to this, Malawi has one of the highest HIV prevalence rates in world, which was found to be at around 10.6% in 2015 [61]. This is one of the indicators that show the significant medical requirements that the population have on health care in Malawi.

As a result of the limited resources and high medical demands, there is often not enough skilled personnel and time available to efficiently allocate a clinic’s resources [14]. Inefficient allocation of resources can result in wastage of the already limited resources available to Malawian health facilities, and thus further degrading the health care available to Malawian citizens.

Computerised Systems such as OpenMRS[48] have been introduced in the Malawian health sector to streamline data collection and data management so as to make resource allocation decisions easier on medical professionals thus attempting to improve the aforementioned inefficiencies [55]. These systems have been adopted well enough to have produced large amounts of useful medical data that medical professionals can use to make quicker and more informed high-level resource allocation decisions [55]. However, the large scale of the data prevents patient specific knowledge to be acted on i.e. medical professionals still lack the time to go through each patient’s data to allocate the necessary resources to individual patients e.g., acknowledging that a patient stopped retrieving medication and subsequently notifying them to continue their treatments.

There is thus a strong need for automated patient specific clinical resource allocation assistance within the Malawian health care context. Machine learning holds the ability to learn from individual patient histories and provide personal predictions for patients [17]. These personal predictions can provide medical professionals with intelligent information about patients so they are able to make quicker, more informed and more detailed resource allocation decisions for the clinic.

This paper aims to contribute in investigating the effectiveness of Machine Learning at providing intelligent patient information that could be used to assist in clinical resource allocation. The paper is limited in scope to two identified clinical resource allocation problems and hence only comments on the effectiveness of Machine Learning for these problems.

1.1 Symptom prediction problem

The Symptom prediction problem is defined as the problem of predicting the next symptom a patient is likely to report based on their past symptom history, demographics and current treatment. This prediction aims to assist clinics in preparing resources for a patient's future medical needs. The preparation that can be done will be based on the requirements of the specific symptom condition but could involve changing a patient's medication in advance, setting up a check-up appointment, sending a health worker to check on patient, notifying the caretaker and patient of what to expect, ensuring clinic will have enough medication available, etc.

1.2 Consultation defaulter problem

The Consultation defaulter problem involves predicting if a specific patient is likely to miss their next consultation based on their appointment history, demographics and the consultation date details. As previously mentioned, there is significantly limited medical resources available hence a patient missing a consultation is a substantial problem as it wastes these vital resources. Currently, around 8% to 12% of all consultations in Malawian clinics are missed [61]. Clinics cannot afford to waste resources on continuously rescheduling and reallocating medical professionals to postponed consultations. There exist methods to avoid this problem which involve sending mail reminders, sending health worker reminders, providing medication with consultation, etc. However, all these methods require more resources allocated to all consultation patients but not all patients are likely to miss their next consultation. The patients that are not likely to miss their consultation thus do not require these extra reminders and incentives so clinics have the possibility of saving on those resources by only allocating them to those patients that are likely to default on their next consultation. The proposed machine learning model can thus assist clinical resource allocation by providing predictions on which patients are likely to default on their next consultation and subsequently which patients require extra consultation reminder resources.

2. Background Work

2.1 Evolution of Symptom prediction and related work

Accurately predicting symptoms has the additional complexity of having to deal with its inherent temporal characteristic. The next symptom is predicted given the patient's past sequence of symptoms and medications along with the history of other patients reported symptom timelines. The symptom prediction problem therefore involves building a model around a sequence of events (reporting symptoms) that the patient follows and not only on a single set of patient descriptors [33, 45].

Rudin et. al [45] attempts to formalise this sequential event prediction by using association rules between the variables and their sequences. Association rules create rules based on common patterns in the sequences in the data. These rules are then used to assist a model in predicting the next event. For example, in the context of building a recommendation system for online grocery shopping, a possible association rule could be "bought lettuce and carrots"-> recommend tomatoes. The main contribution of Rudin et. al [45] 's paper is the establishment of supervised learning based on association rules. The characteristics of this supervised learning type were heavily considered in this paper's experiment as the interaction between symptoms and their sequences were established to be important considerations when predicting a patient's next symptom. An example of a possible association rule that was found in this paper's experiment is "reported lactic acidosis symptom and on celebrex medication" -> likely to lead to skin rash. This shows how the interaction of variables in their sequences are used in the building of models for symptom prediction.

Sequential event prediction was also brought to focus by Letham et. al [32] and Davis et. al [11] as they applied statistical techniques to many inherent event sequence problems including symptom prediction and future disease risk prediction. The systems acknowledged the theory of association rules and were subsequently built to exploit the interconnection between symptom variables and their timelines. The core method, to include the sequence of symptom reporting events, that was used by Davis et. al [11] was to add time-sensitive features to their model such as length between symptom reports, etc. This was reported to increase the accuracy substantially from models that did not account for the sequence of symptoms. These time-sensitive features allow for the sequences of symptom reports to be exploited thus symptom prediction should include the creation and management of temporal features to be accurate models for prediction.

McCormick et. al [33] delves deeper into detail in the application of symptom prediction such that McCormick et. al [3]'s model, bayesian HARM, utilises the sharing of patient information across the training set in addition to using the sequences of a patient's symptoms. The previous mentioned papers did not fully utilise information from other patient's sequences and histories as they were only used to establish association rules. HARM can draw on similar patient's sequences when lacking information about a specific patient. This leads to a more powerful symptom prediction model which is proven as the HARM system was reported as more accurate than Rudin et. al [45] and Letham et. al [2]'s models. The histories of other patients can therefore be a useful predictor when determining a patient's next symptom i.e., checking if they are following the path of a similar patient.

The HARM and previously mentioned systems have all approached the symptom prediction problem from a bayesian statistical methodology instead of a machine learning approach. This was mainly done due to machine learning technique's limitations when dealing with time-related problems. Machine Learning techniques do not offer a way to track and learn from the sequence of events an object follows. However, time-sensitive features have been further explored in literature and the use of individual temporal features were successfully applied in Davis et. al [1] and McCormick et. al [3]'s bayesian models hence there is space to adapt current generic machine learning techniques to utilise temporal features as a measure of a patient's sequence of symptoms. Machine learning techniques already have the strong characteristic of being able to draw on the history of similar data patterns (i.e., similar patient patterns). The now included temporal features along with the historic capabilities of Machine Learning could make Machine Learning a competitively strong predictor for symptom prediction hence why it is explored in this paper.

2.1 Consultation Defaulters

There is strong literature behind defining the attributes of patients who end up missing medical appointments [5, 8-10, 30, 54]. The majority of work focuses on identifying indicators that can be used to understand why a patient misses an appointment (defaults) and subsequently shows what factors signal a patient who is likely to miss a future appointment.

Earlier papers such as Bickler [5] and Cosgrove [10] 's studies used simple statistical methods within relatively smaller medical datasets to find the defaulter characteristics and situations that cause patients to miss appointments. Their long-term goal was to allow clinical facilities to navigate around situations that encourage defaulting e.g., not setting appointments on Fridays, not scheduling appointments too soon, etc. Following these papers, more literature focused on asserting whether the same features from Bickler [1] and Cosgrove [4]'s studies hold in different contexts and larger datasets [8, 18, 30, 54].

After our review of the subsequent literature, the consultation defaulter features that seem to be the most commonly successful are age, day of the week, number of past missed consultations, number of days scheduled ahead of appointment, occupation and cell phone provision [5, 8, 10, 18, 54].

The more recent papers done by Chariatte et. al [9] and Lee et. al [30] attempt to use the known successful consultation defaulter features to build statistical models to predict the likelihood of a patient missing an appointment. These models have the ability to identify patients more personally hence clinics can focus control measures (e.g., SMS's, sending health workers to home, etc) on the patient's who need it the most. Chariatte et. al [9] uses a Markovian multilevel model and produces statistically significant results ($p<0.05$) while Lee et. al [30] utilises a multiple logistic regression model and achieves an area under the ROC Curve value of 84%. The Area under the ROC Curve metric is a measure of how well a model can distinguish between two classes hence 84% is a significant performance for the model. The approach taken by these papers follow a similar approach set by our paper however the datasets used by Chariatte et. al [9] and Lee et. al [30] are based on significantly developed regions (Switzerland and Singapore respectively) whilst our dataset is based on a developing region (Malawi). This is an important consideration for our paper as there is a stronger need in developing regions for enhancing clinic efficiency due to the lack of funding and staffing existing in many developing countries [7]. Models based on developed regions may not translate accurately to developing regions hence why building the consultation defaulter model on developing region data is explored in this paper. The Markov model and logistic regression were also the only models found to be tested for this problem, however the consultation defaulter problem holds characteristics that are well suited enough to be successful with other machine learning techniques hence the problem is tested with several different machine learning techniques in our experiment [28].

Overall, our experiment aims to contribute the performance of the consultation defaulter problem within the developing region context and the performance of the problem with previously unexplored machine learning techniques.

2.3 Challenges with clinical data

2.3.1 Imbalanced class distribution

Real world medical data maintains a common problem for most machine learning models as it often produces imbalanced class distributions. This occurs when there is a low incidence of one class (e.g., patient has cancer) being overwhelmed another class (e.g., patient does not have cancer) [34]. Machine learning attempts to draw out significant features from both classes but with strong imbalances, the majority class could overwhelm the minority class such that very little weighting and importance is given to predicting the minority class [42]. For example, if cancer is being predicted by machine learning but is given an imbalanced class distribution such that only 5% of patients are classified as having cancer, the machine learning technique may adjust to always predict that a patient does not have cancer. It may do this as machine learning techniques aim to achieve maximum prediction accuracy and by ignoring all features to just predict that every patients does not have cancer, it will achieve a substantial 95% accuracy in this example. This is achieved even though it cannot realistically predict a patient with cancer.

Imbalanced clinical data thus poses the threat of producing inherently inaccurate models [42]. This makes the imbalanced class distribution an important problem to address with medical data. The simplest and most common solution to the imbalance problem is to engage in undersampling (ignoring cases from the majority) and oversampling (duplicating cases from the minority) on the dataset [4]. More complex sampling methods exist that can create artificial minority class data instead of simple duplication. There also exists sampling methods that prioritise ignoring insignificant cases from the majority class instead of randomly excluding possibly important, cases [19]. These sampling methods rebalance the classes to not overpower each other.

In our experiment, their existed imbalanced class distributions amongst both the symptom prediction problem and the consultation defaulter problem. These imbalances were both trialled with many complex sampling techniques to achieve the best possible class distribution.

2.3.2 Inherently temporal variables/trends

Clinical data often contains important patterns within its time-related sequences, however generic machine learning techniques do not contain the capabilities to automatically exploit temporal data [36]. Machine learning techniques thus have to first be adapted through specified time-related features, time-related functions or by changing to an underlying time-related model to be able to successfully exploit temporal trends in the data[1]. The symptom prediction problem was approached through the creation of time-related features as the underlying temporal patterns were seen as important considerations for the model.

2.3.3 Verification needed by clinical experts

Clinical problems are a more sensitive domain than others such that results must be fully logical and understandable by medical professionals to be seen as valid [24]. Clinical data contains a wide array of medical variables that machine learning techniques might utilise, however machine learning could draw on medical variables that are clinically unrelated to the problem. If this occurs, then the results could be based on weak assumptions causing the model to be invalid [12, 24]. Medical consultation is thus needed during the variable selection process to ensure a clinically sound model is produced. Our experiment utilised consultation with medical, machine learning and developing region experts to ensure the validity of the models.

2.3.4 Real world data collection issues

Clinic data capturing is done to primarily increase the efficiency for patients visiting the clinic and not to produce data for use in research and analytics [58]. This leads to less precaution being taken for consistency, clarity and completeness in the data. Many data points within the dataset are subsequently ineffective due to missing entries and information [12]. A core variable in this experiment that suffers from this problem is the location feature. The name of the location was inputted in the dataset instead of the longitude and latitude hence the distance between the clinic and patient could not be utilised in the experiment. In addition to this, lesser known village names were manually inputted instead of being chosen which has led to over 7000 distinct villages being loaded in the database without any support to identify the geographical location of the villages. This limits the features usefulness in its models.

To navigate around the "missing data" points, the common data cleaning approach was completed. Data cleaning involves performing operations on the dataset to remove any outlier and incomplete data points from the data. This ensures the most representative data is used in the machine learning models [12].

3. Experiment Design and Execution

The full experiment methods and approaches are fully described within this section

3.1 Problem exploration with consultation and data analysis

The Malawian clinical dataset in its entirety was explored with medical and data analyst consultations where needed. This exploration was done to be able to identify reasonable problems that could be addressed to improve clinical resource allocation. This exploration and analysis of the dataset was done with the Tableau Desktop Professional (v10.3.2) tool, which had the ability to visualise and manipulate the dataset as needed. Through this analysis and background research, the symptom prediction problem and consultation defaulter problem were identified as plausible and feasible opportunities to improve clinical efficiency. Medical professionals were then again consulted with to verify the problems medical rationality which then supported the claim of being able to improve clinical efficiency through these problem solutions. The problems were then pursued through the application of machine learning.

3.2 Malawi Dataset

3.2.1 General dataset

The dataset used for this experiment was collected at a HIV and TB clinic within Malawi's capital city, Lilongwe. The data was stored using OpenMRS, an open source medical record system, hence the dataset follows the data model set by OpenMRS [57]. The core information stored in this data model falls into the observations table, which is where all patient reportings are stored. The observation table spans from January 2007 to March 2017 and holds 27 887 517 records consisting of many different types of clinical reportings.

3.2.2 Symptom and consultation prediction dataset

Of these observation records, 1 392 122 of them were related to symptoms being reported by patients and with that, 23 258 distinct patients were found to be involved in the aforementioned symptom reportings. These records held 13 distinct symptoms that were stored in the system. These records and patient's information were the key datasets used in the symptom prediction problem. The major symptoms and their reportings are presented in figure 1.

922 167 records were related to consultation reportings on whether a patient attended their consultation or missed it. These records consisted of 86 729 distinct patients which are focussed on in the consultation defaulter problem.

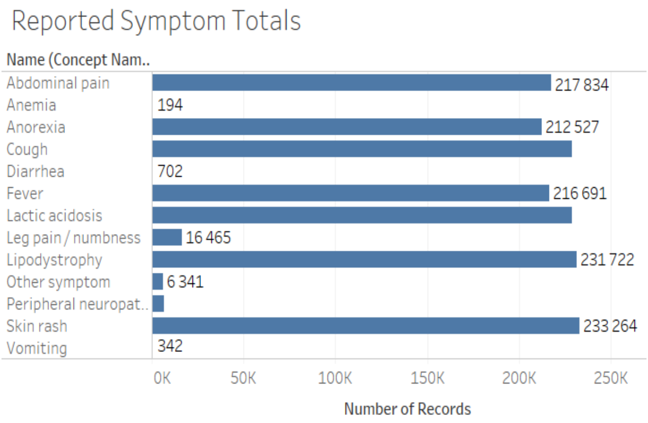

Figure 1: Distinct Symptoms with their total occurrences in the dataset

In addition to the aforementioned dataset information, we also had access to patient demographics such as age, location, sex and occupation along with data on their medical drug use information. These variables were utilised wherever they were deemed relevant by past literature in the symptom prediction and consultation defaulter problems.

Data cleaning methods were also engaged with beforehand to ensure no patient with significantly missing data and no erroneously inputted data was used. The data cleaning was done to ensure that the data used is as representative of the Malawian context as possible.

3.3 Symptom prediction problem approach

This section describes the approach taken to apply machine learning to the symptom prediction problem within the previously described pipeline.

3.3.1 Symptom Prediction Feature and Result Set Building

Feature selection is one of the core factors in machine learning application hence there was a significant focus on the features for this problem. The totals for each reported symptom a patient has done at the clinic was used as a feature as it serves as a descriptor for a patient's total medical history. It allows the model to recognise the types of symptoms a patient has encountered substantially in the past and subsequently weigh in it's likelihood of occurring again.

Since the core aim is to predict a patient's next symptom, the most recently reported symptoms would be significant factors. Therefore, the number of days since the last reporting for each symptom is built as features for the model. These features serve to provide the model with a temporal aspect to a patient's symptom reportings and also aims to exploit the sequence patterns that patient's experience with their symptoms.

In addition to the previously mentioned features, the patient age feature was used due to its prevalence in medical machine learning literature and the last drug dispensed to the patient feature was utilised due to its possible medical relations with specific symptoms.

The result set (i.e. classification trying to predict) is also an important attribute of applying machine learning as it guides the machine learning models in tuning themselves to achieve the best result. The result set was assigned to be the last symptom that was reported by a patient. If the last symptom was reported more than

40 days after the previous reporting then the result changes to “No symptom”. This is due to the fact that the focus in this experiment is on producing a model that could assist in clinical resource allocation and without a time restriction, the model would predict if a patient will, in their clinical lifetime, get a symptom. This is much too broad to be useful in helping the clinic allocate their current resources. 40 days was chosen in particular after consultation with clinical experts on the most reasonable period of time that would be most helpful in assisting clinical resource allocation. Ideally, models should be fitted for many thresholds such as 60 days, 90 days, etc and these varied models should be produced for use in clinics. However, this many thresholds was deemed to be out of scope for this particular experiment. It is also important to mention that since the last symptom reporting is used as the result set, the features were all built using the dataset until the symptom reporting before the last symptom.

The final feature and result set are summarised in Table 1.

Table 1: Features and Result set in symptom prediction problem

Medical history features (Total number of a specific symptom reported)	Cough, Fever, Abdominal pain, Skin rash, Lactic acisdosis, Lipodystrophy, Anemia, Anorexia, Diarrhea, Leg pain, Other, Peripheral neuropathy, Vomiting, Weight loss
Temporal features (how many days ago did the last symptom reporting for that specific symptom occur)	Last Cough, Last Fever, Last Abdominal pain, Last Skin rash, Last Lactic acisdosis, Last Lipodystrophy, Last Anemia, Last Anorexia, Last Diarrhea, Last Leg pain, Last Other, Last Peripheral neuropathy, Last Vomiting, Last Weigh _loss
Patient information features	Age, Last drug dispensed, Total symptom reportings in prev month
Result Set feature (i.e. classification trying to predict)	Last symptom reported if within 40 days of previous symptom reporting. If not, then classified as “No symptom”

3.3.2 Symptom Prediction’s Multiclass and Multilabel complexities

The result set can be classified as any symptom defined in the aforementioned dataset, which makes symptom prediction a multiclass problem. This adds to the complexity of the problem as not all machine learning techniques are built, at their basic form, to be able to be applied to multiclass problems [2, 23]. Another complexity with the symptom prediction problem is that it is plausible that patients will report multiple symptoms in the next 40 days and not just a single reporting e.g., a patient could report cough and fever at the same time, etc. This makes the ideal result set, a list of symptoms that a patient is predicted to experience in the next 40 days. Producing a resultset that is a list is referred to as a mulilabel problem within the machine learning domain and has much literature on methodologies taken to approach this type of problem [52].

The approach taken in this experiment against the multiclass and multilabel challenges was to utilise the One vs All technique [21, 39]. This technique allows each class to be fitted against all the other classes in a resultset while applying a machine learning technique. This thus automatically reduces a machine learning application to many binary class problems instead of a single multiclass problem [21]. This technique consequently produces a model for each individual symptom e.g., Cough, Not Cough; Skin rash, Not Skin rash; Fever, Not Fever, etc. Therefore, for each machine learning technique, there will be a model produced for each individual symptom that will predict whether a patient will experience that specific symptom or not in the next 40 days. For example, a list could be produced about a patient that reads Cough, Not Fever, Not Abdominal Pain, Skin rash, etc. This would indicate that the patient will experience cough and a skin rash but no abdominal pain and fever. Therefore, this technique allows for binary machine learning techniques to be utilised and allows for a multilabel solution that describes a list of symptoms to be produced.

3.3.3 Symptom Prediction’s Class balance

As mentioned in the background work, the balance of the result classes can heavily influence a machine learning technique’s performance hence it is important to address the class imbalances within the dataset so as to achieve the maximum possible performance [42]. The class balance for the symptom prediction problem is presented in table 2.

Table 2: Class balance for the symptom prediction problem. Cough(Cgh), Fever(Fev), Abdominal pain(Abd), Skin rash(Rash), Lipodystrophy(Lipo), Anemia(Anm), Anorexia(Anx), Diarreah(Drha), Leg pain(Leg), Night Sweats(NS), Peripheral neuropathy(PN), Vomitting(Vom, Weight loss(Wgt),Other(Oth), No symptom(None)

Cgh	**Fev**	**Abd**	**Rash**	**Lipo**	**Anm**	**Anx**	**Drha**
368	73	126	10106	34	8	9	51
Leg	**NS**	**PN**	**Vom**	**Wgt**	**Oth**	**None**	
143	5	94	20	167	345	1259	

Table 2 shows the significant imbalance between the many symptom classes hence further justifying the need for sample balancing techniques to be used in this symptom prediction problem.

3.4 Consultation Defaulter problem approach

This section describes the approach taken to apply machine learning to the consultation defaulter problem within the previously described pipeline.

3.4.1 Consultation Defaulter Feature and Result Set Building

The features for the consultation defaulter problem were predominantly determined from the wide range of literature available on significant factors affecting patient appointment attendance [5, 8, 10, 18, 30]. After this literature analysis, the features chosen were Total Consultations Attended, Total Consultations Missed, Sex, Age, Location, Occupation, Day of the Week and If Attended Last Appointment.

The result set was set to a binary boolean value of whether a patient attended their last consultation or missed it. However, since there were very few instances of patients who had missed their last reported consultation (but missed consultations before), the result set rolled back the reportings of those patients to when the patient last missed a consultation. For example, if a patient's consultation reportings were as shown in table 3 then the reportings for patient 182 would be omitted after the second entry. This helps to better understand why a patient defaults on a consultation and significantly improves the class balance of the problem so machine learning techniques can more accurately predict when a patient is likely to miss a consultation.

Table 3: Example of patient attendance reports

patient id	date	attended?
182	04/01/2012	Yes
182	11/01/2012	No
182	20/01/2012	Yes

3.4.2 Consultation Defaulter's Class balance

The class balance for the consultation defaulter problem was a substantially imbalanced result set. This is due to the inherent lower incidence of missed consultations when compared to attended consultations [5]. The class balances shown in table x display the balance using the last reported consultation in the original dataset and the class balance with the aforementioned adjusted rolled back reported consultations. This is displayed to show justification for the adjustment to the original dataset.

Table 4: Class balances for the original and adjusted consultation defaulter problem result set

	Attended	Missed
Original	47716	276
Adjusted	33655	14337

The adjusted class balance, even though improved, still holds a considerable class imbalance hence shows a need for sample balancing techniques to be utilised.

3.5 Pipeline/System Development and Implementation

The experiment was applied through python (v3.6.1) as the core infrastructure due to its ease of use and wide range of libraries available for machine learning. Subsequently. the machine learning application and metrics were done through the popular scikit-learn (v0.19.1) python library for its substantial support and documentation [39]. Finally, the balancing techniques were are all available and applied through another open source python library called imbalanced-learn (v0.3) [31].

The dataset was stored within a MySql database which was accessed through the pymysql python library. This library allowed for specified queries to be called on the dataset thus allowing for the needed records to be brought into the pipeline environment [56]. The records were then manipulated in the python environment to build the necessary features for the problem set. A new MySQL table containing all the necessary problem features (e.g., Cough totals, patient age, etc) was created for each prediction problem. This made the retrieval of features quick and simple for the machine learning techniques

3.6 Machine Learning Techniques Used

Scikit-learn offers a wide range of available machine learning techniques to use. This experiment utilised 8 of these available machine learning techniques due their popularity and past performance in clinical domains [49]. The machine learning techniques used are Logistic Regression, K Nearest Neighbours, Decision Tree Classifier, Gaussian Naïve Bayes, Random Forest, Multi-Layer Perceptron Classifier, Adaptive Boosting and Support Vector Machine. These machine learning techniques are described in more detail, along with their expected performance in table 5.

3.7 Balancing Methods Used

There exists little literature covering the effectiveness of different balancing techniques within the medical context. This creates a difficulty in the choice of balancing method for this experiment. Thus, the approach taken was to compare as many techniques as possible so as to gain insight into which balancing technique should be utilised for the final results. The imbalanced-learn library used in this experiment offers a wide range of balancing methods which were significantly taken advantage of for the broad comparison. However, not all the balancing techniques were built to handle multiclass problems hence the symptom prediction problem underwent less sampler comparisons than the consultation defaulter problem. Undersampling was the main approach focused on in this experiment as it utilises more real data while Oversampling attempts to create artificial data. Prioritising real data is an important consideration due to the sensitivity of the medical context. The consultation problem compared the techniques All-KNN, Near Miss, Condensed Nearest Neighbour, Tomek links, Neighbourhood Cleaning Rule, Instance Hardness Threshold and a uniformly random under sampler as a baseline sampler. The symptom prediction problem only compares Near Miss, RandomUnderSampling and All-KNN balancing techniques.

3.8 Cross-validation and Unseen validation

Overfitting is a common problem in machine learning that occurs when a model learns the detail and noise in a training set to the extent that it negatively affects the model's performance on new data [13]. A common solution to this is k-fold cross-validation, as it is a technique used to evaluate predictive statistical models in such a way that overfitting is significantly less likely to occur [25].

K-fold cross-validation works by partitioning the original dataset into k equal subsets of the dataset. These subsets are named folds and of these folds, one is selected to be used as a validation subset while all other folds act as the training set. The model runs with the allocated training set folds and the one fold validation set, subsequently producing a result of how well the model can predict the validation fold. This process is then iterated with the selection of the validation fold changing every time until each fold has been used as a validation set exactly once. The prediction results from each iteration is then averaged to produce a final cross-validation prediction result. This reduces the likelihood of overfitting as the training set is separated from the test set hence the model is tested against unseen data rather than data from its own training set [25, 43].

Technique	**Summary of technique's algorithm**	**Expected Experiment Performance**
Logistic Regression (LR)	Logistic regression constructs a separating decision boundary (hyperplane) between two datasets, using the logistic function to measure distance from the decision boundary as a probability of being classified as a specific class [22].	Due to the technique's significant performance in past literature with simple models, the technique is expected to perform well with the simple consultation problem but not as well with the more complex symptom problem
K Nearest Neighbours Classifier (KNN)	The K Nearest Neighbours algorithm operates by classifying objects based on the majority vote of its neighbours (plotted in the feature space), with the classification being given to the most common class among the object's k nearest neighbours [46].	KNN is predominantly used as a benchmark measure in medical machine learning literature hence it is not expected to be the peak performer in this experiment [15].
Decision Tree Classifier (DTree)	A Decision Tree is a tree model where leaves represent classifications and branches represent conjunctions of variables that lead to those classifications. The Decision Tree Classifier uses training data to adjust the decision tree to be able to translate observations about an object to a classification [41].	This machine learning technique is very commonly used within medical applications due to its simple human understandable model and performance hence it is expected to produce positive results in this experiment [29].
Gaussian Naive Bayes (Naïve)	Naive Bayes uses Bayes theorem [53] to predict a class from a set of attributes by choosing the class that obtains the highest relative probability. Naive bayes has the 'naive' assumption of independence between features [38].	This is a simple model but often performs significantly well with linear problems hence is expected to perform especially effectively with the consultation defaulter problem.
Random Forest (R Frst)	Random Forests is an ensemble (multiple learning algorithm) technique that constructs multiple varied decision trees with the training data. Each decision tree has the capability of producing its own classification through its tree structure. The technique then outputs the class that occurred most frequently amongst its group of decision trees [59].	Random Forest has seen very strong performance in past literature and holds the advantage of easily attainable variable significance/interpretability [27]. Thus, Random Forests is expected to be one of the peak performing techniques in this experiment.
Multi-Layer Perceptron (MLP)	A Multi-Layer Perceptron Classifier is a type of Artificial Neural Network that consists of at least three layers of nodes which are trained using backpropagation supervised learning. Each neuron uses a nonlinear activation function hence has the ability classify data that is not linearly separable [44].	The MLP classifier is the most complex model used in this experiment but maintains the most flexibility in its learning methods on the data. Hence, the technique is expected to perform positively for both problems [44].
Adaptive Boosting (Ada)	AdaBoost is an ensemble technique that fits many variations of one classifier (Decision Tree in this experiment) and combines them together into a weighted sum that constitutes the final output of the AdaBoost classifier [60].	AdaBoost can thus achieve high accuracy with relatively low amounts of data [50], hence could produce significant results within this experiment's context.
Support Vector Machine (SVM)	The Support Vector constructs a hyperplane (boundary) in a higher dimensional space and classifies objects through this boundary. This is done as decision boundaries may be complex in lower dimensions but easier in a higher dimensional feature space [58].	Support Vector Machines are not as commonly utilised in medical machine learning, but have performed significantly positively when used [16]. This technique thus shows a promising expectation for performance in this experiment.

Table 5: The chosen Machine Learning techniques, their algorithms and their expected performance

Subsequently, 10-fold cross-validation was chosen as the solution to the overfitting problem in this experiment due its simplicity, effectiveness and prevalence of support in medical machine learning literature [9, 12, 30, 34, 58]. A pre-existing cross-validation tool within the python scikit-learn library was utilised to apply 10-fold cross-validation in this experiment [39]. In addition to the cross-validation performance, a 30% portion of the dataset is held back for unseen validation of the performance of the model. This 30% sample undergoes no balancing techniques hence is a better representation of the real-world dataset than the cross-validated balanced results hence is the primary performance metric for the comparison.

3.9 Evaluation Metrics for the techniques

3.9.1 Area under the ROC Curve

Machine learning applications often use the prediction accuracy measure as the main metric for performance. The prediction accuracy metric measures the ratio of correct predictions to the total predictions made. This metric is frequently used due to its simplicity and wide support, however it is only suitable when there is an equal class balance and all predictions are equally important [40]. The problems explored in this experiment do not sufficiently fulfil those requirements as both prediction problems are significantly imbalanced. Therefore, the default prediction accuracy metric is not sufficient for this experiment.

The Area Under the Receiver Operating Characteristic (ROC) Curve could serve as a significantly improved substitute performance metric in this experiment for the machine learning techniques. The ROC curve plots the true positive rate, ratio of correctly classified positive samples (Sensitivity), against the false positive rate, incorrectly classified positive samples (Specificity). Therefore, the area under the ROC curve is a measure of how well a model can distinguish between two classes [6, 20]. This is a more suitable measure as it poses equal importance on classifying the minority (lower incidence) class as the majority class. The ROC metric thus produces a much more appropriate metric for an imbalanced dataset such as the datasets within the problems in this experiment. Hence, the ROC measure was chosen as the primary metric in this experiment for comparison between machine learning techniques.

3.9.2 Interpretability

The proposed use of the machine learning techniques in this experiment are within the sensitive domain of clinical application.

This domain has high requirements of validity needed for models used within its context. This is due its higher human risk and higher requirements for efficient resource usage (e.g., allocation of treatments, doctors, etc). Therefore, there is a strong need for human interpretable machine learning models so that clinical experts can verify the clinical sensibility within a model's prediction.

The approach taken in this experiment to measure a model's interpretability and validity was to showcase the most significant features that a model utilises to produce its prediction. The most significant features were then consulted with literature and medical experts to ensure its clinical sensibility. Hence, this experiment will produce the feature significance for each model and compare this with what literature and medical experts would expect.

It is also important to note that not all chosen machine learning models are built to easily produce their most significant features hence not all model feature significance measures will be compared as this is beyond the scope of this experiment.

4. Results

4.1 Symptom prediction Results

4.1.1 Balancing technique comparison

The sampler balancing techniques are compared and presented in table 6. However, not all selected scikit-learn balancing techniques have the capability to balance a multiclass result set hence some chosen balancing techniques are omitted from this comparison. In addition to this, due to the multiclass nature of the data, undersampling and oversampling techniques were used in combination to improve the final results hence the combinations are compared in this section as well. Samples first underwent uniformly random duplication of the minority sets so the later undersampling of the majority class would not bring the overall class set too low of an amount.

Table 6: Comparison of Cross-validated ROC values with different balancing techniques on the symptom prediction problem

	LR	**KNN**	**Dtree**	**Naïve**	**R Frst**	**MLP**	**Ada**	**SVM**
Random Under	0.501	0.523	0.671	0.427	0.642	0.343	0.626	0.698
Near Miss	0.515	0.561	0.852	0.347	0.831	0.313	0.747	0.734
AllKNN	0.573	0.543	0.740	0.396	0.721	0.272	0.728	0.628

4.1.2 ROC comparison

The Near Miss balancing technique was used in the following results due to its previously seen significant results. The Roc performance metric was used to compare the machine learning techniques as can be seen in table 7.

Table 7: Comparison of the machine learning technique cross-validation roc(Cross Roc) scores and unseen hold set ROC scores(Unseen Roc) on the symptom prediction problem

	LR	**KNN**	**Dtree**	**Naïve**	**R Frst**	**MLP**	**Ada**	**SVM**
Unseen Roc	0.662	0.638	0.744	0.575	0.711	0.601	0.713	0.589
Cross Roc	0.515	0.561	0.852	0.347	0.831	0.313	0.747	0.734

Since the model is based on a multilabel result set which was set to produce a model for each symptom class, the performance for each class can be measured and are presented in table 8. The best three techniques based on the previous Unseen ROC scores (Table 6) are only shown in table 9.

4.1.3 Inter variable significance

The significance of variables when determining certain classes was measured in all machine learning techniques that offered this capability. The average influence of a variable when determining specific symptom classes was then analysed. After this analysis, the variables were then ranked with the most significant variables placed as 1st. This is to show which were the main variables for predicting a specific symptom e.g., Predicting if a patient gets fever is mostly determined from when they last got fever (Last Fever), their total reportings of coughs and when their last cough reporting was. The overall results are presented in in Table 8.

Table 8: Most significant features chosen for each symptom model. Abdominal pain(Abd pain), Skin Rash(Rash), Lipodystrophy(Lipo), Anemia(Anm), Anorexia(Anx), Diarhea(Diar), Leg pain(Leg), Night Sweats(NS), Peripheral Neuropathy(PN), Vomiting(Vom), Weight loss(Wgt), Other(Oth), Last X is the amount of days since the last reporting of X (temporal aspect), Prev month is the amount of symptoms reported in the previous month

	Most Significant Features		
	1st	**2nd**	**3rd**
Cough	Cough Total	Age	Last Cough
Fever	Last Fever	Cough Total	Last Cough
Abd pain	Last Abd	Abd Total	Age
Rash	Age	Last Drug	Last Lactic
Lipo	Last Lipo	Anx Total	Lipo Total
Anm	Last Anm	Anm Total	Age
Anx	Last Anx	Last Lact	Anx Total
Diar	Last Diar	Diar Total	Last Leg
Leg	Last Leg	Leg Total	Age
NS	NS Total	Last NS	Fever Total
PN	Last PN	PN Total	Age
Vom	Last Vom	Vom Total	Age
Wgt	Wgt Total	Last Wgt	Cough Total
Oth	Last Oth	Oth Total	Last Drug
No symp	Last Drug	Age	Prev month

Table 9: Comparison of top three machine learning technique's models for each symptom. Cough(Co), Fever(Fev), Abdominal pain(Abd), Skin Rash(Rash), Lipodystrophy(Lipo), Anemia(Ane), Anorexia(Anx), Diarhea(Diar), Leg pain(Leg), Night Sweats(NS), Peripheral Neuropathy(PN), Vomiting(Vom), Weight loss(Weig)

	Co	**Fev**	**Abd**	**Rash**	**Lipo**	**Ane**	**Ano**	**Diar**	**Leg**	**NS**	**PN**	**Vom**	**Weig**	**Other**	**None**	**Avg**
Dtree	0.59	0.68	0.83	0.53	0.89	1.00	0.50	0.86	0.74	0.50	0.78	0.86	0.97	0.86	0.58	0.74
R Forest	0.66	0.69	0.78	0.53	0.89	0.67	0.50	0.73	0.72	0.50	0.73	0.93	0.98	0.76	0.62	0.71
Ada	0.62	0.60	0.74	0.51	0.88	0.67	0.50	0.91	0.87	0.50	0.82	0.78	0.97	0.78	0.56	0.71

4.2 Consultation Defaulter Results

4.2.1 Balancing technique comparison

The balancing techniques are first compared to establish which sampler balance is the most effective for the problem prediction so more detailed analysis can be done on the machine learning techniques. The Roc values were obtained for each sampling balance and are shown in table 10.

Table 10: Comparison of application of balancing techniques for each machine learning algorithm. Seen cross-validation ROC scores are used to compare performance. Attendance class(Att), Missed Attendance class(Miss), Logistic Regression(LR), K Neighbours (KNN),

Sampler	**Att**	**Miss**	**LR**	**KNN**	**Dtree**	**Naïve**	**R Frst**	**MLP**	**Ada**	**SVM**	**Avg**
Orig	33655	14337	0.81	0.68	0.74	0.81	0.82	0.82	0.84	0.72	0.78
ALLKNN	15281	14337	0.86	0.87	0.81	0.86	0.90	0.87	0.90	0.90	0.87
NearMiss	14337	14337	0.86	0.84	0.78	0.88	0.88	0.87	0.91	0.85	0.86
CNN	12172	14337	0.79	0.57	0.67	0.78	0.78	0.79	0.81	0.53	0.72
Tomek	30296	14337	0.81	0.70	0.75	0.81	0.83	0.82	0.85	0.75	0.79
NCR	16698	14337	0.85	0.82	0.77	0.85	0.88	0.86	0.89	0.84	0.85
Instance	18745	14337	0.84	0.77	0.78	0.84	0.90	0.85	0.89	0.78	0.83
Random	14337	14337	0.81	0.68	0.71	0.81	0.82	0.80	0.85	0.68	0.77

4.2.2 Roc, sensitivity and specificity comparison

The performance metrics ROC, sensitivity and specificity were applied to each machine learning technique with the results presented in table 11.

Table 11: Comparison of performance metrics for the consultation defaulter problem

	LR	**KNN**	**Dtree**	**Naïve**	**R Forest**	**MLP**	**Ada**	**SVM**
Roc	0.856	0.899	0.809	0.857	0.908	0.884	0.900	0.926
Sensitivity	0.951	0.929	0.817	0.950	0.907	0.938	0.885	0.799
Specificity	0.595	0.722	0.806	0.559	0.757	0.683	0.740	0.900
Unseen Roc	0.773	0.825	0.811	0.779	0.832	0.810	0.813	0.849

4.2.3 Feature significance

The significance that each model placed on each feature is presented in table 12. It is to be noted that some models are not built to easily identify their most significant features hence these are omitted from the result table due to them being out of scope for this paper

Table 12: Significance of variables in the consultation defaulter problem. Numerical values represent the weighting of each variable in determining a prediction. All weightings sum up to 1. Features are Total consultations attended (C(A)), Total Consultations Missed(C(M)), Sex, Age, Occupation(Occ), Location(Loc), which day of the week the appointment is(Day), Whether attended last consultation(Prev)

	C(A)	**C(M)**	**Sex**	**Age**	**Occ**	**Loc**	**Day**	**Prev**
R Forest	0.24	0.27	0.01	0.16	0.07	0.17	0.05	0.03
Dtree	0.16	0.40	0.02	0.15	0.04	0.18	0.06	0.00
Ada	0.24	0.04	0.02	0.26	0.14	0.18	0.08	0.04

5. Ethics

Since this experiment operates on the personal medical data of human patients, special ethical clearance was navigated. The dataset was fully anonymised through automation scripts and thus core personal information that could have been used to identify a patient was removed. Special Malawian ethical clearance to be able to operate on a dataset containing Malawian citizens was then requested and obtained. After these two methods were completed, the experiment was allowed to be executed to achieve the best machine learning techniques for the aforementioned problems.

6. Discussion

The three best performers in the symptom prediction problem were all based on tree classifiers. This can be explained by the inherent medical nature of the problem. Clinical problems such as this generally perform well with tree classifiers as the solutions are more systematically rules predominant e.g., if treatment is cotrimoxazole, then more likely cough, etc [12]. However, even the peak performers only reached levels of around 70% which are fair results but are most likely not significant enough to be used within the real world medical context yet. The prediction results were not substantially significant due to the immense complexity underlying the problem i.e., a medical symptom could be the result of any of the numerous amounts of health characteristics a patient has [3]. The features used were based purely on the reported symptoms, however more explanatory power could be needed from more medical characteristics about a patient e.g. blood pressure, CD4 count, etc. The model did show that different symptoms influenced each other in table 8 where symptoms that are medically correlated could be seen interacting with each other to better predict the final result e.g. Night sweats used a patient's fever history to determine its likelihood, fever used a patient's cough history, etc. This shows that the model was able gain explanatory power from having access to full medical histories. The significance of the temporal "Last X" variable was seen to be very powerful as it was predominantly used as a highly ranked and strong feature for most symptom models, as can be seen in table 7. This shows the benefits of modelling the machine learning techniques with temporal characteristics that can take advantage of the temporal nature of the symptom prediction problem. The imbalance in the dataset did also hold back the predictive power of the machine learning models as is evident in Table 8 where the class prediction for the minority classes (Night Sweats, Anorexia, etc) are consistently poor across all techniques.

The balancing techniques produced significantly different results for some machine learning techniques but also majorly performed similarly with other techniques in the consultation defaulter problem. The average performance showed the significant performance of AllKNN which outperformed the weaker balancing techniques by over 10%. This is a very significant improvement as it leads the models to reach a significantly reliable level of performance hence is more likely to be utilised in real world application. The Support Vector Machine obtained the best performance amongst all models, most likely due to its more flexible non-linear approach to classifying the data. The Multi-layer perceptron also hold this non-linear characteristic but require much more data to reach peak performance levels. The Support Vector Machine also managed to attain high sensitivity and specificity metrics indicating the notable ability to not falsely classify objects. Finally, the significance of the features show the insignificance of the sex variable which can easily be accepted as there being no significant difference between the sex's default rates. However, the day of the week and previous appointment attendance also proved to be insignificant which was not expected according the literature. This deviation could be explained by the fact that the literature comments on the variables in terms of the overall patient population. However, machine learning operates on a patient level hence if there is little evidence of an individual patient having the day of the week and last appointment attendance affecting his/her future attendance then the overall machine learning feature significance for those variables are reduced. All other variables proved to be significant in accordance to the majority of literature's comments.

7. Conclusion

We investigated the use of eight machine learning techniques on the problem of predicting a patient's next likely symptom and on the problem of predicting a patient's next likely consultation attendance. The Decision Tree Classifier proved to be the most effective performing machine learning technique at the symptom prediction problem with a resulting ROC value of 74.4%. While this ROC value may be a fair result, it is not significant enough to be utilized within the real world clinical context as yet. This being said, it does provide a good basis for approaching the problem and could be built on in future work to achieve significant enough performance results to be able to be accepted for use in real world application. The consultation defaulter problem managed to achieve the significant ROC value of 85% through the Support Vector Machine as its peak performer. This performance is significant enough to be considered for use in clinical application and hence could be a useful tool to assist clinics in allocating their resources efficiently to intervene on patients who are likely to default on their consultation with the necessary resources. These results were also achieved within a developing country context therefore shows their viability at being used within the developing nation context.

References

[1] Ahmed, N. K., Atiya, A. F., Gayar, N. E. and El-Shishiny, H. An empirical comparison of machine learning models for time series forecasting. Econometric Reviews, 29, 5-6 (2010), 594-621.

[2] Allwein, E. L., Schapire, R. E. and Singer, Y. Reducing multiclass to binary: A unifying approach for margin classifiers. Journal of machine learning research, 1, Dec (2000), 113-141.

[3] Arndt, S., Andreasen, N. C., Flaum, M., Miller, D. and Nopoulos, P. A longitudinal study of symptom dimensions in schizophrenia: prediction and patterns of change. Arch. Gen. Psychiatry, 52, 5 (1995), 352-360.

[4] Batista, G. E., Prati, R. C. and Monard, M. C. A study of the behavior of several methods for balancing machine learning training data. ACM Sigkdd Explorations Newsletter, 6, 1 (2004), 20-29.

[5] Bickler, C. B. Defaulted appointments in general practice. JR Coll Gen Pract, 35, 270 (1985), 19-22.

[6] Bradley, A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit, 30, 7 (1997), 1145-1159.

[7] Bradley, S., Kamwendo, F., Chipeta, E., Chimwaza, W., de Pinho, H. and McAuliffe, E. Too few staff, too many patients: a qualitative study of the impact on obstetric care providers and on quality of care in Malawi. BMC pregnancy and childbirth, 15, 1 (2015), 65.

[8] Challenger, A., Coleman, T. and Lewis, S. Predicting default from smoking cessation treatment following enrolment. Health Educ. J., 66, 1 (2007), 32-43.

[9] Chariatte, V., Berchtold, A., Akr, C., Michaud, P. and Suris, J. Missed appointments in an outpatient clinic for adolescents, an approach to predict the risk of missing. Journal of Adolescent Health, 43, 1 (2008), 38-45.

[10] Cosgrove, M. P. Defaulters in general practice: reasons for default and patterns of attendance. Br. J. Gen. Pract., 40, 331 (1990), 50-52.

[11] Davis, D. A., Chawla, N. V., Christakis, N. A. and Barabsi, A. Time to CARE: a collaborative engine for practical disease prediction. Data Mining and Knowledge Discovery, 20, 3 (2010), 388-415.

[12] Delen, D., Walker, G. and Kadam, A. Predicting breast cancer survivability: a comparison of three data mining methods. Artif. Intell. Med., 34, 2 (2005), 113-127.

[13] Domingos, P. A few useful things to know about machine learning. Commun ACM, 55, 10 (2012), 78-87.

[14] Douglas, G. P., Landis-Lewis, Z. and Hochheiser, H. Simplicity and usability: lessons from a touchscreen electronic medical record system in Malawi. interactions, 18, 6 (2011), 50-53.

[15] Dreiseitl, S., Ohno-Machado, L., Kittler, H., Vinterbo, S., Billhardt, H. and Binder, M. A comparison of machine learning methods for the diagnosis of pigmented skin lesions. J. Biomed. Inform., 34, 1 (2001), 28-36.

[16] Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M. and Haussler, D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics, 16, 10 (2000), 906-914.

[17] Greenes, R. A. *Clinical decision support: the road ahead.* Academic Press, , 2011.

[18] Griffin, S. J. Lost to follow-up: the problem of defaulters from diabetes clinics. Diabetic Med., 15, S3 (1998).

[19] Han, H., Wang, W. and Mao, B. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. Advances in intelligent computing, (2005), 878-887.

[20] Hanley, J. A. and McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143, 1 (1982), 29-36.

[21] Hong, J. and Cho, S. A probabilistic multi-class strategy of one-vs.-rest support vector machines for cancer classification. Neurocomputing, 71, 16 (2008), 3275-3281.

[22] Hosmer Jr, D. W., Lemeshow, S. and Sturdivant, R. X. *Applied logistic regression.* John Wiley & Sons, , 2013.

[23] Hsu, C. and Lin, C. A comparison of methods for multiclass support vector machines. IEEE Trans. Neural Networks, 13, 2 (2002), 415-425.

[24] Jones, R. H., Hannan, E. L., Hammermeister, K. E., DeLong, E. R., O'Connor, G. T., Luepker, R. V., Parsonnet, V. and Pryor, D. B. Identification of preoperative variables needed for risk adjustment of short-term mortality after coronary artery bypass graft surgery. J. Am. Coll. Cardiol., 28, 6 (1996), 1478-1487.

[25] Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In Anonymous *Ijcai.* (). Stanford, CA, , 1995, 1137-1145.

[26] Kononenko, I. Machine learning for medical diagnosis: history, state of the art and perspective. Artif. Intell. Med., 23, 1 (2001), 89-109.

[27] Kononenko, I. Machine learning for medical diagnosis: history, state of the art and perspective. Artif. Intell. Med., 23, 1 (2001), 89-109.

[28] Kotsiantis, S. B., Zaharakis, I. D. and Pintelas, P. E. Machine learning: a review of classification and combining techniques. Artif. Intell. Rev., 26, 3 (2006), 159-190.

[29] Lavanya, D. and Rani, K. U. Ensemble decision tree classifier for breast cancer data. International Journal of Information Technology Convergence and Services, 2, 1 (2012), 17.

[30] Lee, V. J., Earnest, A., Chen, M. I. and Krishnan, B. Predictors of failed attendances in a multi-specialty outpatient centre using electronic databases. BMC health services research, 5, 1 (2005), 51.

[31] Lemaitre, G., Nogueira, F. and Aridas, C. K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of Machine Learning Research, 18, 17 (2017), 1-5.

[32] Letham, B., Rudin, C. and Madigan, D. Sequential event prediction. Mach. Learning, 93, 2-3 (2013), 357-380.

[33] McCormick, T., Rudin, C. and Madigan, D. A hierarchical model for association rule mining of sequential events: An approach to automated medical symptom prediction. (2011).

[34] Mena, L. J. and Gonzalez, J. A. Machine Learning for Imbalanced Datasets: Application in Medical Diagnostic. In Anonymous *Flairs Conference.* (). , 2006, 574-579.

[35] More, A. Survey of resampling techniques for improving classification performance in unbalanced datasets. arXiv preprint arXiv:1608.06048, (2016).

[36] Moskovitch, R. and Shahar, Y. Medical temporal-knowledge discovery via temporal abstraction. In Anonymous *AMIA annual symposium proceedings.* (). American Medical Informatics Association, , 2009, 452.

[37] Mueller, D. H., Lungu, D., Acharya, A. and Palmer, N. Constraints to implementing the essential health package in Malawi. PloS one, 6, 6 (2011), e20741.

[38] Ng, A. Y. and Jordan, M. I. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In Anonymous *Advances in neural information processing systems.* (). , 2002, 841-848.

[39] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R. and Dubourg, V. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, Oct (2011), 2825-2830.

[40] Pencina, M. J., D'Agostino, R. B. and Vasan, R. S. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat. Med., 27, 2 (2008), 157-172.

[41] Polat, K. and Gnes, S. A novel hybrid intelligent method based on C4. 5 decision tree classifier and one-against-all approach for multi-class classification problems. Expert Syst. Appl., 36, 2 (2009), 1587-1592.

[42] Provost, F. Machine learning from imbalanced data sets 101. In Anonymous *Proceedings of the AAAI'2000 workshop on imbalanced data sets.* (). , 2000, 1-3.

[43] Rodriguez, J. D., Perez, A. and Lozano, J. A. Sensitivity analysis of k-fold cross validation in prediction error estimation. IEEE Trans. Pattern Anal. Mach. Intell., 32, 3 (2010), 569-575.

[44] Ruck, D. W., Rogers, S. K., Kabrisky, M., Oxley, M. E. and Suter, B. W. The multilayer perceptron as an approximation to a Bayes optimal discriminant function. IEEE Trans. Neural Networks, 1, 4 (1990), 296-298.

[45] Rudin, C., Letham, B., Salleb-Aouissi, A., Kogan, E. and Madigan, D. Sequential event prediction with association rules. In Anonymous *Proceedings of the 24th annual conference on learning theory.* (). , 2011, 615-634.

[46] Sarkar, M. and Leong, T. Application of K-nearest neighbors algorithm on breast cancer diagnosis problem. In Anonymous *Proceedings of the AMIA Symposium.* (). American Medical Informatics Association, , 2000, 759.

[47] Scheffler, R. M., Mahoney, C. B., Fulton, B. D., Dal Poz, M. R. and Preker, A. S. Estimates of health care professional shortages in sub-Saharan Africa by 2015. Health Aff., 28, 5 (2009), w862.

[48] Seebregts, C. J., Mamlin, B. W., Biondich, P. G., Fraser, H. S., Wolfe, B. A., Jazayeri, D., Allen, C., Miranda, J., Baker, E. and Musinguzi, N. The OpenMRS implementers network. Int. J. Med. Inf., 78, 11 (2009), 711-720.

[49] Szarvas, G., Farkas, R. and Busa-Fekete, R. State-of-the-art anonymization of medical records using an iterative machine learning framework. Journal of the American Medical Informatics Association, 14, 5 (2007), 574-580.

[50] Thongkam, J., Xu, G. and Zhang, Y. AdaBoost algorithm with random forests for predicting breast cancer survivability. In Anonymous *Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on.* (). IEEE, , 2008, 3062-3069.

[51] Tomek, I. An experiment with the edited nearest-neighbor rule. IEEE Trans. Syst. Man Cybern., , 6 (1976), 448-452.

[52] Tsoumakas, G. and Katakis, I. Multi-label classification: An overview. International Journal of Data Warehousing and Mining, 3, 3 (2006).

[53] Vapnik, V. N. and Vapnik, V. *Statistical learning theory.* Wiley New York, , 1998.

[54] Waller, J. and Hodgkin, P. Defaulters in general practice: who are they and what can be done about them? Fam. Pract., 17, 3 (2000), 252-253.

[55] Waters, E., Rafter, J., Douglas, G. P., Bwanali, M., Jazayeri, D. and Fraser, H. S. Experience implementing a point-of-care electronic medical record system for primary care in Malawi. In Anonymous *Medinfo.* (). , 2010, 96-100.

[56] Widenius, M. and Axmark, D. *MySQL reference manual: documentation from the source.* " O'Reilly Media, Inc.", , 2002.

[57] Wolfe, B. A., Mamlin, B. W., Biondich, P. G., Fraser, H. S., Jazayeri, D., Allen, C., Miranda, J. and Tierney, W. M. The OpenMRS system: collaborating toward an open source EMR for developing countries. In Anonymous *AMIA annual symposium proceedings.* (). American Medical Informatics Association, , 2006, 1146.

[58] Wu, J., Roy, J. and Stewart, W. F. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. Med. Care, 48, 6 (2010), S113.

[59] Yang, F., Wang, H., Mi, H. and Cai, W. Using random forest for reliable classification and cost-sensitive learning for medical diagnosis. BMC Bioinformatics, 10, 1 (2009), S22.

[60] Zhu, J., Zou, H., Rosset, S. and Hastie, T. Multi-class adaboost. Statistics and its Interface, 2, 3 (2009), 349-360.

[61] Zungu, L. I., Magombo, T., Chikaonda, T., Thomas, R., Mwenda, R., Kandulu, J., Chilima, B., Chiwaula, M., Mbene, A. and Saka, E. A national quality assurance programme for point-of-care testing in Malawi. African Journal of Laboratory Medicine, 5, 2 (2016), 1-5.