The Application of Data Mining to Predict the Occurrence of Short-Term Adverse Events in NB-UVB Phototherapy Treatments

S. Mohamed, A-M. Tobin, A. D. Irvine, D. R. Wall, N. J. O'Hare, and M-T. Kechadi

***Abstract*—The prediction of the occurrence of short-term adverse events in phototherapy treatments is important for dermatologists who administrate phototherapy to adjust the treatment and standardize the clinical outcomes. Recently, a modeling technique that can detect the potential occurrence of short-term adverse events in phototherapy treatments is required for clinicians. Based on data mining, this study tends to explore the significant features and the class distribution of training data for predicting the occurrence of short-term adverse events in NB-UVB phototherapy treatments. The experimental results highlight that an acceptable prediction accuracy can be achieved using the significant features and the performance of the classifiers can be significantly improved by sampling 40% of the negative class samples in the training data, hyper-parameter tuning of the classifiers and use of stacked classifiers in creating the prediction models.**

***Index Terms*—Adverse events, classification, data mining, dermatology, phototherapy, prediction.**

I. Introduction

Phototherapy involves repeated exposure of the skin to ultraviolet (UV) light to treat various inflammatory skin conditions, such as psoriasis, eczema, and vitiligo. This therapy is one of the oldest treatment modalities in dermatology, dating back to the ancient Egyptians, who used natural light in combination with herbal extracts to treat skin diseases [1], [2]. Phototherapy continues to be a highly preferred treatment by dermatologists [2], [3].

There are three main types of phototherapy used for the treatment of psoriasis: broadband ultraviolet B (BB-UVB), narrowband ultraviolet B (NB-UVB) and psoralen plus ultra-violet A (PUVA) [2]. UVB is the most commonly preferred first-line treatment for moderate-to-severe psoriasis in healthy adults [3]. NB-UVB is the most commonly used phototherapy modality as it has a wider application across various dermatologic conditions, it's easier to use and has fewer adverse events when compared with BB-UVB or PUVA [2], [4]. In addition, these three main types of phototherapy cause some acute short-term adverse events, including erythema and burning, lesional blisters, and pruritus [5]-[7].

Concerns regarding skin cancer development and the occurrence of adverse events in phototherapy have become a common source of legal claims in dermatology and have emphasized the importance of fastidious monitoring of its delivery [8]. It is identified as a yellow flag action recommended service standard by the British Association of Dermatologists to compare and standardize the number of episodes/patient treatments/year for each grade of symptomatic erythema compared to the published standards. The published rates vary but include <0.8% of all treatments result in an acute adverse event (0.6% for NB-UVB [0.5% excluding Grade I Erythema], 1.3% for systemic PUVA and 0.8% for local PUVA); for severe adverse events: 0.05% for NB-UVB and 0.3% for systemic PUVA [9].

If clinicians know the prospectives of the treatment beforehand, they can adjust the treatment and standardize the clinical outcomes. Therefore, a model that can predict whether a treatment may cause acute adverse events is useful for dermatologists who administrate phototherapy.

Recently, data mining techniques have been applied in healthcare areas [10]-[12] and particularly, in dermatology [13].

Artificial neural networks (ANN) have been successfully used for the diagnosis of skin diseases. Yoon, Brobst, Bergstresser and Peterson [14] used ANN with a back-propagation algorithm for the diagnosis of papulosquamous skin diseases. For the diagnosis of erythematous squamous diseases, Übeyli [15] used combined neural networks (CNN) and achieved an accuracy of 97.77%. Chang and Chen [16] achieved a highest accuracy of 80% using a ANN model, outperforming models constructed from a decision tree, the combination of a decision tree and ANN, a decision tree with sensitivity analysis and an ANN with sensitivity analysis. Karlik and Harman [17] used supervised back-propagation with 95% accuracy. Olatunji and Arif (2013) used an ANN and extreme learning machine, Sarhan, Elharir and Zakaria [18] used an ANN Levenberg-Marquardt algorithm with a rough sets attribute reduction technique with 98.8% accuracy.

Support vector machines (SVM) have been used since 2006 to create models for the diagnosis of skin diseases. Nanni [19] proposed an ensemble of linear SVM based on random sub-space and feature selection that improved the average predictive accuracy gained by a standalone SVM or by an RS ensemble of SVMs. Übeyli [20] used a hybrid technique, which combined multi-class SVM with error correcting output code, which achieved an accuracy of 98.32%. Xie and Wang [21] achieved an accuracy of 98.61% for the model they implemented using an SVM and

Manuscript received January 22, 2018; revised April 8, 2018. This work was supported in part by Irish Research Council under Grant EBPPG/2014/84 and Irish Skin Foundation.

S. Mohamed and M-T. Kechadi are with University College Dublin, Belfield, Dublin 4 (sharifa.mohamed@ucdconnect.ie, bing.huang@ucd.ie, tahar.kechadi@ucd.ie).

Anne-Marie Tobin is with Dermatology Department, Tallagh Hospital.

Alan D. Irvine is with School of Medicine, Trinity College Dublin.

Dmitri R Wall is with Irish Skin Foundation.

Neil J. O'Hare is with School of Public Health, Physiotherapy & Sports, University College Dublin.

doi: 10.18178/ijmlc.2018.8.2.671

IFSFFS (improved F-score and sequential forward) feature selection method. Giveki, Salimi, Bitaraf, and Khademian [22] proposed a model based on catfish binary particle swarm optimization (CatfishBPSO), kernelized support vector machines (KSVM) and association rule feature selection method, which gained an accuracy of 99.09%. Abdi and Giveki [23] also proposed a hybrid method of particle swarm optimization, support vector machine and association rules, which achieved an accuracy of 98.91%.

Mroczek, Paja, Piatek, and Wrzesie [24] used an ID3 decision tree as one of the models used for the diagnosis and classification of melanocytic skin lesions. Polat and Güneş [25] used a C4.5 decision tree classifier with a one-against-all approach with 84.48% accuracy and Tran (2008, April) used a Gini index based decision tree for erythematosquamous disease diagnosis.

Manjusha, Sankaranarayanan, and Seena [26] used a naïve bayes classifier to predict eight different dermatological conditions while Aruna, Nandakishore, and Rajagopalan [27] used a hybrid feature selection method with a naïve bayes classifier, which achieved 98.9% accuracy and Danjuma and Osofisan [28] obtained the highest accuracy of 97.4% from a naïve bayes classifier, which outperformed a multi-layer perception and J48 decision tree for the diagnosis of Erythemato-Squamous skin disease.

Cataloluk and Kesler [29] created a diagnostic software tool for skin diseases with basic and weighted K-NN and gained an accuracy of 96.36% when a Manhattan distance was used for the weighted K-NN.

Ensembles made by combining different classification techniques have also been used by many researchers to create skin disease diagnostic models. Elsayad [30] has used an ensemble model created by combining a multi-layer neural network, decision tree and linear discriminant analysis (LDA) techniques and achieved an accuracy of 98.23%. Sharma and Hota [31] proposed a hybrid ensemble model by combining a support vector machine and artificial neural network and obtained a 98.99% test accuracy.

Although there has been a high level of interest in implementing skin disease diagnostic models, there are very few reports in the literature describing the use of data mining techniques in phototherapy data analysis, dermatological treatment outcome prediction or dermatological adverse events occurrence prediction.

This paper introduces a prediction model used to detect the acute adverse events of a treatment using data mining techniques. Based on the NB-UVB phototherapy data set, the proposed prediction model first selects the number of attributes, prepares the data and finally applies classification algorithms to predict the occurrence of adverse events.

II. Methodology

A. Data set Collection

The data set used in this paper was obtained from the Adelaide and Meath Hospital, Dublin, known as Tallaght hospital's PuvaMate UVB phototherapy database. The data set was professionally anonymized by OpenApp Computer Support and Services after obtaining ethical approval from Tallaght hospital research ethics committee.

The UVB phototherapy database consists of 29836 treatment records from 897 patients treated since the end of September 2003. The information on each patient includes the patient personal details (e.g., gender, year of birth, skin type) and treatment details. There were 464 females and 434 males among the patients. These patients were treated for psoriasis, eczema, granuloma annulara, acne, nodular prurigo, mycosis fungoides, ple, urticaria pigmentosa, morphea, lichen spinulosa, pityrasis lichenoideschronicu and vitiligo. Among all these records, the psoriasis, eczema, and nodular prurigo treatment records were studied separately in order to predict the adverse event occurrence of the treatments because these were the top 3 diseases treated. Table I shows the percentage of acute adverse event occurrences for the above-mentioned diseases. If any of the short-term acute adverse events including erythema, burning, lesional blisters or pruritus were noted following the treatment, it was recorded as a positive occurrence of an adverse event and otherwise marked as a negative occurrence. These were used as the prediction classes. As the negative class records were much higher in number than the positive cases, the negative class was the majority class in these cases.

B. Methods

RStudio with R version 3.3.1 was used on a 64-bit Windows operating system to conduct the experiments with the help of the mlr machine learning package.

The prediction model used to detect the acute adverse events of the treatments consists of the pre-processing and classification processes. The pre-processing process first selects the significant attributes from the psoriasis, eczema and nodular prurigo data sets, then filters the noise data and normalizes the data. The classification process applies the modelling algorithms to predict whether each treatment causes an adverse event. These two processes are described as follows:

1) Pre-processing: The goal of this phase was to provide cleaned data for the classification step. In the pre-processing phase, we first derived the new attributes from the data set and applied the information gain technique [32] to select the significant attributes or features. Then, the missing values were added based on the domain knowledge or mode of the attribute in which the missing value was replaced by the value that makes the most sense or in other cases by the value that is most common. Next, we used a local outlier factor technique (LOF) [33] to deal with the local outliers. Finally, the data set was normalized [34]. In this step, new binary attributes were created for the categorical attributes and the numerical attribute values were scaled to fall between 0 and 1.

The features selected by the information gain technique used for the classification process are summarized below. The dosage, dose difference between previous and current treatment, ratio of dosage to med, previous treatment dosage, course cumulative dose, dose difference percentage between previous and current treatment, date difference between previous and current treatment, if the treatment dosage was increased, reduced or repeated compared to previous treatment, total cumulative dose, skin type, and gender were among the attributes, which gave a non-zero information gain for the psoriasis records.

TABLE I: THE OCCURRENCE OF ADVERSE EVENTS IN THE PSORIASIS, ECZEMA AND NODULAR PRURIGO DATA SETS

Dataset	Tot.samples	N.adverse samples	N.normal samples
Psoriasis	23819	3170 (13.3%)	20649 (87%)
Eczema	4238	377(8.9%)	3861 (91%)
Nod. prurigo	791	84(10.6%)	707 (89%)

All of the above-mentioned attributes except the dose difference percentage between previous and current treatment, date difference between previous and current treatment attributes gave a non-zero information gain for the eczema records and except dose difference percentage between previous and current treatment, date difference between previous and current treatment and total cumulative dose attributes gave non-zero information gain for the nodular prurigo records.

2) *Classification:* Patients treated for psoriasis, eczema and nodular prurigo were analyzed separately and 3 experiments were conducted on each data set.

a) Experiment 1: Classification algorithms [35], [36] were used with the default parameters to predict the occurrence of any adverse events.

For each data set, eXtreme gradient boosting (XGB), linear discriminant analysis (LDA) Adaboost, C50 (C5.0 decision tree), generalized boosted regression modeling (GBM), K-nearest neighbors classifier (IBk) J48 (C4.5 decision tree), JRip (propositional rule learner based on association rules with reduced error pruning), naïve bayes (NB), neural network (NNet), OneR (generates one rule for each predictor in the data, then selects the rule with the smallest total error as its "one rule"), PART (uses partial decision trees), random forest (RF: an ensemble learning method for classification, regression, and other tasks, that operate by constructing a multitude of decision trees) and SVM were applied to predict whether each treatment causes any adverse events. The 3-fold cross-validation [37] was used to evaluate each classification model.

The class distribution was imbalanced with only 13.3% positive cases (the occurrence of a short-term adverse event) in the psoriasis data set, 8.9% positive cases in the eczema data set and 10.6% positive cases in the nodular prurigo data set. However, these learning techniques were designed and attempt to find an accurate performance over a full range of samples, based on the balanced classes of training data set. If learning from the data set with the highly imbalanced class distribution, these learning techniques tend to be overwhelmed by the majority class and ignore the minority class, and consequently, provide poor classification results [38].

To solve this problem, we adjusted the class distribution of the training data by under-sampling the majority class samples in the training data. The method of under-sampling data used here was the "farthest distance" technique [39]. When we applied the 3-fold cross-validation for the data set, we under-sampled the majority class samples to 60%, 50% and 40% in the training data of the psoriasis, eczema, and nodular prurigo data sets.

b) Experiment 2: We chose the overall best performing undersampled data sets of psoriasis, eczema, and nodular prurigo patients, and applied parameter tuning of the classifier algorithms to improve the accuracy of the classifiers. The performance was evaluated using 3-fold cross-validation.

c) Experiment 3: We used the parameter tuned classifiers from experiment 3 and created stacked classifiers of size 2 to check if the accuracy could be further improved. An L1-regularized logistic regression classifier was used as the super learner when creating the stacked classifiers. Again, the performance was evaluated using 3-fold cross-validation.

III. RESULTS AND DISCUSSION

This paper uses the area under curve, overall accuracy and f1-score to evaluate the classification models. The area under curve (AUC) is an abbreviation for the area under the ROC (receiver operating characteristic) curve, based on the minor class. It is the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one [40]. The accuracy rate is the proportion of the correctly classified samples in all the samples. The f1-score is the harmonic mean of precision and recall. The details of these terms can be found in [41].

TABLE II: THE PERFORMANCE OF THE CLASSIFICATION ALGORITHMS WHEN THE MAJORITY CLASS INSTANCES ARE UNDERSAMPLED TO 60%, 50% AND 40% OF THE ORIGINAL SIZE IN THE PSORIASIS RECORDS

	Majority Class*60%			**Majority Class*50%**			**Majority Class*40%**		
alg.	***AUC***	***acc***	***f1***	***AUC***	***acc***	***f1***	***AUC***	***acc***	***f1***
XGB	0.84	0.89	0.66	0.86	0.89	0.73	0.90	0.92	0.84
LDA	0.77	0.81	0.39	0.84	0.86	0.69	0.88	0.89	0.80
Adaboost	0.86	0.90	0.68	0.88	0.90	0.74	0.91	0.92	0.84
C50	0.82	0.90	0.68	0.84	0.90	0.74	0.87	0.92	0.83
GBM	0.74	0.79	0.00	0.78	0.76	0.00	0.85	0.71	0.00
IBk	0.76	0.85	0.63	0.79	0.85	0.68	0.84	0.87	0.77
J48	0.82	0.89	0.68	0.83	0.90	0.74	0.87	0.92	0.84
JRip	0.75	0.89	0.66	0.80	0.90	0.74	0.86	0.92	0.84
NB	0.80	0.45	0.40	0.84	0.50	0.46	0.90	0.56	0.55
NNet	0.83	0.89	0.68	0.86	0.90	0.75	0.90	0.92	0.84
OneR	0.53	0.77	0.17	0.71	0.80	0.56	0.85	0.89	0.80
PART	0.83	0.88	0.66	0.86	0.89	0.73	0.90	0.92	0.84
RF	0.85	0.90	0.69	0.87	0.90	0.75	0.90	0.92	0.84
SVM	0.82	0.90	0.70	0.84	0.91	0.76	0.88	0.92	0.85

TABLE III: THE PERFORMANCE OF THE CLASSIFICATION ALGORITHMS WHEN THE MAJORITY CLASS INSTANCES ARE UNDERSAMPLED TO 60%, 50% AND 40% OF THE ORIGINAL SIZE IN THE ECZEMA RECORDS

	Majority Class*60%			**Majority Class*50%**			**Majority Class*40%**		
alg.	***AUC***	***acc***	***f1***	***AUC***	***acc***	***f1***	***AUC***	***acc***	***f1***
XGB	0.77	0.90	0.55	0.83	0.90	0.65	0.85	0.91	0.72
LDA	0.74	0.85	0.01	0.77	0.87	0.44	0.82	0.85	0.64
Adaboost	0.82	0.92	0.63	0.85	0.92	0.70	0.86	0.92	0.76
C50	0.79	0.92	0.63	0.80	0.91	0.67	0.82	0.92	0.75
GBM	0.71	0.86	0.00	0.73	0.83	0.00	0.78	0.80	0.00
IBk	0.77	0.88	0.59	0.79	0.88	0.63	0.82	0.88	0.70
J48	0.78	0.91	0.62	0.80	0.92	0.69	0.81	0.91	0.75
JRip	0.70	0.91	0.57	0.75	0.91	0.64	0.79	0.91	0.73
NB	0.77	0.52	0.34	0.82	0.64	0.42	0.85	0.61	0.49
NNet	0.80	0.91	0.60	0.78	0.90	0.58	0.83	0.90	0.66
OneR	0.51	0.85	0.04	0.60	0.83	0.34	0.73	0.85	0.58
PART	0.78	0.92	0.62	0.78	0.92	0.69	0.83	0.92	0.75
RF	0.81	0.92	0.62	0.85	0.92	0.70	0.87	0.92	0.77
SVM	0.80	0.92	0.64	0.83	0.92	0.70	0.85	0.92	0.77

The classification results of experiment 1 are presented in Tables II, III, and IV. In these tables, the AUC, acc, and f1 represent the average area under the curve, average accuracy and average f1-score of 3-fold cross-validation obtained by the classifier algorithms namely XGB, LDA, Adaboost, C5.0, GBM, IBk, J48, JRip, NB, NNet, OneR, PART, RF and SVM when the majority class instances were under-sampled to 60%, 50%, and 40% of the original size in the psoriasis data set (Table II), eczema data set (Table III) and nodular prurigo data set (Table IV).

TABLE IV: THE PERFORMANCE OF CLASSIFICATION ALGORITHMS WHEN THE MAJORITY CLASS INSTANCES ARE UNDERSAMPLED TO 60%, 50% AND 40% OF THE ORIGINAL SIZE IN THE NODULAR PRURIGO RECORDS

	Majority Class*60%			**Majority Class*50%**			**Majority Class*40%**		
alg.	***AUC***	***acc***	***f1***	***AUC***	***acc***	***f1***	***AUC***	***acc***	***f1***
XGB	0.79	0.88	0.57	0.79	0.85	0.57	0.82	0.88	0.72
LDA	0.88	0.88	0.55	0.89	0.89	0.70	0.91	0.92	0.80
Adaboost	0.87	0.90	0.61	0.93	0.91	0.72	0.93	0.91	0.77
C50	0.76	0.90	0.60	0.86	0.91	0.74	0.88	0.92	0.80
GBM	0.80	0.83	0.00	0.79	0.81	0.00	0.81	0.77	0.00
IBk	0.84	0.91	0.73	0.89	0.92	0.81	0.91	0.92	0.84
J48	0.78	0.90	0.61	0.87	0.93	0.78	0.89	0.92	0.80
JRip	0.71	0.89	0.57	0.73	0.89	0.61	0.80	0.89	0.72
NB	0.85	0.60	0.46	0.89	0.58	0.47	0.89	0.60	0.51
NNet	0.87	0.89	0.68	0.87	0.92	0.77	0.91	0.92	0.83
OneR	0.64	0.82	0.40	0.66	0.82	0.45	0.79	0.89	0.72
PART	0.78	0.90	0.61	0.85	0.92	0.76	0.89	0.93	0.82
RF	0.87	0.91	0.64	0.92	0.91	0.71	0.95	0.92	0.81
SVM	0.92	0.93	0.71	0.92	0.94	0.81	0.92	0.93	0.84

As shown in Table II, when the majority class was under-sampled to 60% of the original size in the psoriasis data set, the Adaboost classification algorithm gave the highest AUC of 0.86, the highest accuracy of 0.90 was recorded by the Adaboost, C5.0, RF, and SVM algorithms and the highest f1-score (70%) was recorded using the SVM algorithm. The lowest AUC was recorded as 0.53 using the OneR algorithm. The lowest accuracy was obtained using the NB algorithm (0.45) and lowest f1-score (0) was obtained using the GBM algorithm. When the majority class instances were under-sampled to 50% and 40% of the original size, the overall performance of the classification algorithms improved. The highest AUC of 0.91 was recorded using the Adaboost algorithm, the highest accuracy (0.92) was recorded using Adaboost, XGB, C50, J48, JRip, NNet, PART, RF, and SVM.

The best f1-score (0.85) was recorded using SVM when the majority class was under-sampled to 40% of the original size. When we consider the overall performance, Adaboost, RF, and SVM were the top 3 algorithms used for the psoriasis data set and GBM, NB and OneR performed poorly.

For the eczema data set shown in Table III, Adaboost recorded the highest AUC (0.82), SVM, RF, Adaboost, PART and C5.0 recorded the best accuracy (0.92) and SVM recorded the highest f1-score (0.64) when the majority class instances are under-sampled to 60% of the original size. Likewise, in the psoriasis data set. The classifier performance in the eczema data set improved when majority class instances are under-sampled to 50% and 40%. RF recorded the highest AUC (0.87), Adaboost, RF, SVM, PART, C50, and J48 recorded the highest accuracy (0.92) and highest f1-score (0.77) was obtained using RF and SVM when majority class instances were under-sampled to 40% of the original size. The overall best performers in the eczema data set were Adaboost, RF, and SVM, which are also the top 3 performers in the psoriasis data set. Likewise, GBM, OneR, and NB performed poorly.

TABLE V: THE HYPER-PARAMETER TUNING SETTINGS

Classifier	**Param name**	**Param type**	**Param descrption**	**Tuned values**
XGB	booster	discrete	booster type	gbtree,gblinear
	eta	numeric	step size of each boosting step	0.1–0.7
	nthread	integer	number of thread used in training	1–20
LDA	tol	numeric	A tolerance to decide if a matrix is singular	0.0001–0.001
Adaboost	loss	discrete	loss type	exponential,logistic
	type	discrete	type of boosting algorithm to perform	discrete,real,gentle
	iter	integer	number of boosting iterations to perform	10–100
C50	winnow	logical	should predictor winnowing (i.e., feature selection) be used?	True, False
	noGlobalPruning	logical	should global pruning step used?.	True, False
GBM	distribution	discrete	the distribution type	bernoulli,Adaboost,huberized
	n.trees	integer	the total number of trees to fit.	10–100
IBk	K	integer	number of nearest neighbors to be used	1–250
J48	C	numeric	confidence threshold for pruning.	0.1–0.5
	M	integer	minimum number of instances per leaf	1–20
JRip	N	numeric	minimal weights of instances	2–10
	O	integer	number of runs of optimizations	2–20
NB	laplace	numeric	provides a smoothing effect	0–10
NNet	maxit	integer	maximum number of iterations	10–100
	size	integer	number of units in the hidden layer	2–50
OneR	B	integer	minimum number of objects in a bucket	2–10
PART	C	numeric	confidence threshold for pruning.	0.1–0.5
	M	integer	minimum number of instances per leaf	1–10
RF	ntree	integer	number of trees to grow	2–500
	mtry	integer	no. of variables used as candidates at each split	2–9
SVM	kernel	discrete	kernel function used in training and predicting	rbfdot, polydot, tanhdot, laplacedot, besseldot, anovadot
	scale	numeric	used with “tanhdot” and “polydot” kernels	1–10
	offset	numeric	used with “tanhdot” and “polydot” kernels	1–10
	sigma	numeric	used with “besseldot,” “anovadot,” “rbfdot,” and “laplacedot” kernels	1–10
	degree	integer	used with “besseldot,” “anovadot,” and "Polydot” kernels	1–6
	order	integer	used with “besseldot” kernel	1–10

TABLE VI: THE BEST HYPER-PARAMETER VALUES AND ACCURACY OF THE CLASSIFIER ALGORITHMS WHEN THE PARAMETERS TUNED ON THE MAJORITY CLASS ARE UNDERSAMPLED TO 40% IN THE PSORIASIS DATA SET

Algorithm	Best hyperparameter values	Accuracy
XGB	booster = gbtree, eta = 0.6333333, nthread = 20	0.92
LDA	tol = 8e-04	0.89
Adaboost	loss = logistic, type = gentle, iter = 60	0.92
C50	winnow = true, noGlobalPruning = true	0.92
GBM	distribution = huberized, n.trees = 100	0.77
IBk	*K* = 56	0.92
J48	*C* = 0.1444444, *M*=1	0.92
JRip	*N* = 7.333333, *O* = 6	0.92
NB	laplace = 4.444444	0.56
NNet	maxit = 40, size = 50	0.92
OneR	*B* = 7	0.89
PART	*C* = 0.4111111, *M* = 10	0.92
RF	ntree = 500, mtry =3	0.92
SVM	kernel=laplacedot, sigma=0.1	0.92

TABLE VII: THE BEST HYPER-PARAMETER VALUES AND ACCURACY OF THE CLASSIFIER ALGORITHMS WHEN THE PARAMETERS TUNED ON THE MAJORITY CLASS ARE UNDERSAMPLED TO 40% IN THE ECZEMA DATA SET

Algorithm	Best hyperparameter values	Accuracy
XGB	booster = gbtree, eta = 0.2333333, nthread = 3	0.91
LDA	tol = 0.001	0.85
Adaboost	loss = exponential, type = gentle, iter = 90	0.92
C50	winnow = false, noGlobalPruning = false	0.92
GBM	distribution = bernoulli,n.trees = 100	0.80
IBk	*K* = 29	0.91
J48	*C* = 0.3222222, *M* = 5	0.92
JRip	*N* = 2, *O* = 10	0.92
NB	laplace = 1.111111	0.67
NNet	maxit = 50, size = 29	0.93
OneR	*B* = 6	0.85
PART	*C* = 0.4111111, *M* = 1	0.92
RF	ntree = 223, mtry = 2	0.92
SVM	kernel = laplacedot, sigma = 1.2	0.92

When we consider the results obtained for the nodular prurigo data set shown in Table IV, we can see that when the majority class instances were undersampled to 60% of the original size, SVM recorded the highest AUC of 0.92 and the highest accuracy of 0.93. The best f1-score was obtained using IBK and recorded as 0.73. Similar to both the psoriasis and eczema data sets, the performance of the classifiers improved when the majority class instances are under-sampled to 50% and 40%. RF recorded 0.95 as the best AUC, with SVM and PART scoring 0.93 as the best accuracy and SVM and IBK scored 0.84 as the best f1-score when the majority class instances are undersampled to 40% of the original size. Unlike for the psoriasis and eczema data sets, the overall best performers in the nodular prurigo data set were SVM, IBk and NNet. However, GBM, NB and OneR were the worst performers in the Nodular prurigo data set, which was the same as that observed for the psoriasis and eczema data sets.

Like psoriasis, the eczema and nodular prurigo records gave the best performance when the majority class instances are undersampled to 40% of the original size; these data sets were used to check if we can further improve the accuracy in experiment 2 and 3.

The results of experiment 2 are presented in Tables VI, VII, and VIII. The hyper-parameters of the 14 classifiers namely XGB, LDA, Adaboost, C50, GBM, IBk, J48, JRip, NB, NNet, OneR, PART, RF, and SVM were tuned to obtain the highest accuracy using the parameter settings shown in Table V.

TABLE VIII: THE BEST HYPER-PARAMETER VALUES AND ACCURACY OF CLASSIFIER ALGORITHMS WHEN THE PARAMETERS TUNED ON THE MAJORITY CLASS ARE UNDERSAMPLED TO 40% IN THE NODULAR PRURIGO DATA SET

Algorithm	Best hyperparameter values	Accuracy
XGB	booster = gbtree, eta = 0.6333333, nthread = 1	0.90
LDA	tol = 0.0006	0.92
Adaboost	loss = exponential, type = discrete, iter = 60	0.92
C50	winnow = false, noGlobalPruning = true	0.93
GBM	distribution = bernoulli, n.trees = 40	0.77
IBk	*K* = 1	0.93
J48	*C* = 0.5, *M* = 1	0.93
JRip	*N* = 2,*O* = 10	0.93
NB	laplace = 5.555556	0.64
NNet	maxit = 80, size = 2	0.93
OneR	*B* = 7	0.90
PART	*C* = 0.5, *M* = 1	0.92
RF	ntree = 334, mtry = 3	0.92
SVM	kernel = laplacedot, sigma = 0.1	0.95

TABLE IX: THE ACCURACY OF THE HYPER-PARAMETER TUNED STACKED CLASSIFIERS ON THE MAJORITY CLASS UNDERSAMPLED TO 40% IN THE PSORIASIS DATA SET

	XGB	LDA	Adaboost	C50	GBM	IBk	J48	JRip	NB	NNet	OneR	PART	RF	SVM
XGB	-	-	-	-	-	-	-	-	-	-	-	-	-	-
LDA	0.92	-	-	-	-	-	-	-	-	-	-	-	-	-
Adaboost	0.92	0.92	-	-	-	-	-	-	-	-	-	-	-	-
C50	0.92	0.92	0.92	-	-	-	-	-	-	-	-	-	-	-
GBM	0.92	0.92	0.92	0.92	-	-	-	-	-	-	-	-	-	-
IBk	0.92	0.92	0.92	0.92	0.92	-	-	-	-	-	-	-	-	-
J48	0.92	0.92	0.92	0.92	0.92	0.92	-	-	-	-	-	-	-	-
JRip	0.92	0.92	0.92	0.92	0.92	0.92	0.92	-	-	-	-	-	-	-
NB	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	-	-	-	-	-	-
NNet	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	-	-	-	-	-
OneR	0.92	0.92	0.92	0.92	0.89	0.92	0.92	0.92	0.91	0.92	-	-	-	-
PART	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	-	-	-
RF	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	-	-
SVM	0.92	0.92	0.92	0.92	0.92	0.92	0.93	0.92	0.92	0.92	0.92	0.92	0.92	-

TABLE X: THE ACCURACY OF THE HYPER-PARAMETER TUNED STACKED CLASSIFIERS ON THE MAJORITY CLASS UNDERSAMPLED TO 40% IN THE ECZEMA DATA SET

	XGB	LDA	Adaboost	C50	GBM	IBk	J48	JRip	NB	NNet	OneR	PART	RF	SVM
XGB	-	-	-	-	-	-	-	-	-	-	-	-	-	-
LDA	0.91	-	-	-	-	-	-	-	-	-	-	-	-	-
Adaboost	0.92	0.92	-	-	-	-	-	-	-	-	-	-	-	-
C50	0.92	0.92	0.92	-	-	-	-	-	-	-	-	-	-	-
GBM	0.91	0.86	0.92	0.91	-	-	-	-	-	-	-	-	-	-
IBk	0.92	0.92	0.93	0.92	0.92	-	-	-	-	-	-	-	-	-
J48	0.92	0.92	0.92	0.92	0.92	0.92	-	-	-	-	-	-	-	-
JRip	0.92	0.92	0.93	0.92	0.92	0.93	0.92	-	-	-	-	-	-	-
NB	0.92	0.88	0.92	0.92	0.89	0.92	0.92	0.92	-	-	-	-	-	-
NNet	0.92	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	-	-	-	-	-
OneR	0.91	0.89	0.92	0.92	0.85	0.91	0.91	0.92	0.87	0.92	-	-	-	-
PART	0.92	0.92	0.92	0.93	0.92	0.92	0.92	0.92	0.92	0.92	0.92	-	-	-
RF	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.93	0.92	0.92	-	-
SVM	0.93	0.93	0.93	0.92	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	-

TABLEXI: THE ACCURACY OF THE HYPER-PARAMETER TUNED STACKED CLASSIFIERS ON THE MAJORITY CLASS UNDERSAMPLED TO 40% IN THE NODULAR PRURIGO DATA SET

	XGB	LDA	Adaboost	C50	GBM	IBk	J48	JRip	NB	NNet	OneR	PART	RF	SVM
XGB	-	-	-	-	-	-	-	-	-	-	-	-	-	-
LDA	0.93	-	-	-	-	-	-	-	-	-	-	-	-	-
Adaboost	0.92	0.93	-	-	-	-	-	-	-	-	-	-	-	-
C50	0.91	0.93	0.91	-	-	-	-	-	-	-	-	-	-	-
GBM	0.91	0.92	0.93	0.90	-	-	-	-	-	-	-	-	-	-
IBk	0.93	0.94	0.95	0.94	0.93	-	-	-	-	-	-	-	-	-
J48	0.92	0.92	0.91	0.94	0.94	0.94	-	-	-	-	-	-	-	-
JRip	0.92	0.93	0.92	0.92	0.90	0.94	0.92	-	-	-	-	-	-	-
NB	0.90	0.92	0.92	0.92	0.87	0.95	0.93	0.91	-	-	-	-	-	-
NNet	0.92	0.93	0.91	0.93	0.92	0.93	0.93	0.92	0.93	-	-	-	-	-
OneR	0.90	0.93	0.91	0.91	0.89	0.93	0.91	0.92	0.90	0.93	-	-	-	-
PART	0.93	0.92	0.92	0.91	0.92	0.94	0.92	0.91	0.93	0.92	0.91	-	-	-
RF	0.93	0.94	0.92	0.93	0.92	0.94	0.93	0.93	0.93	0.93	0.92	0.92	-	-
SVM	0.94	0.95	0.94	0.95	0.94	0.95	0.95	0.95	0.94	0.94	0.95	0.95	0.94	-

When we compare the accuracy of the psoriasis records where the majority class instances are undersampled to 40% under the default classifier settings shown in Table II and when the hyper-parameters were tuned in the classifiers as shown in Table VI, we can see an improvement in the accuracy of two classifiers, namely GBM and IBk. The accuracy of the GBM classifier was improved from 0.71 to 0.77 when the hyper-parameters were tuned. Likewise, the accuracy of IBk improved from 0.87 to 0.92. None of the other classifiers showed an improvement in the accuracy when the hyper-parameters were tuned for the psoriasis records, where the majority class instances are undersampled to 40% of the original size.

By comparing Table III and VII we can see that the IBk, J48, JRip, NB and NNet algorithms improved the accuracy when the hyper-parameters were tuned and the classifiers applied on the majority class undersampled to 40% of the original size in the eczema data set.

The LDA, GBM, and RF algorithms did not improve the performance, while the PART classifier showed a drop in accuracy from 0.93 to 0.92 when the hyper-parameters were tuned and the classifiers applied on the majority class undersampled to 40% of the original size in the nodular prurigo data set. (See Table IV and VIII).

Table IX, X and XI illustrate the results of experiment 3 in which the hyper-parameter tuned stacked classifiers were used in the psoriasis, eczema and nodular prurigo data sets, where the majority class instances were undersampled to 40% of the original size. An L1-regularized logistic regression classifier was used as the super learner.

As shown in Table IX, we can see that all the stacked classifier combinations except the stacked classifiers made of the OneR and GBM base learners, and the OneR and NB base learners recorded an accuracy of 0.92 for the psoriasis data set. The OneR and GBM combination recorded an accuracy of 0.89. The OneR and NB combination recorded an accuracy of 0.91, which was an improvement when compared to 0.89 and 0.56 recorded for experiment 2 in Table VI.

Most of the stacked classifiers scored an accuracy of higher than 0.9 for the eczema data set, as shown in Table X. The GBM and LDA, NB and LDA, OneR and LDA, NB and GBM, and NB and OneR classifier combinations recorded an accuracy of 0.86, 0.88, 0.89, 0.89 and 0.87, respectively, which show an improvement when compred to using them as single classifiers (Table III). The stacked classifier made up of OneR and GBM recorded an accuracy of 0.85 and did not show an improvement when compared to Table III.

Also for the nodular prurigo data set, we can see that most of the stacked classifiers scored an accuracy of 0.9 or higher, as shown in Table XI. The OneR and GBM combination showed a drop in performance from 0.9 to 0.89 when used in the stacked classifier when compared to results shown in Table IV. However, the NB and GBM combination could improve the accuracy to 0.87.

IV. Conclusions and Future Work

This paper introduces the prediction of short-term adverse events in NB-UVB phototherapy treatments using data mining techniques. We identified the significant feature sets for psoriasis, eczema, and nodular prurigo data sets and used 14 learning algorithms to classify the occurrence of short-term adverse events in the data sets in experiment 1. Then, we tried to improve the accuracy of these classifiers by tuning the hyper-parameters in experiment 2. Experiment 3 made use of these hyper-parameter tuned classifiers to create stacked classifiers of size 2. The findings of this paper are:

1) The most effective features that models the occurrence of adverse events.
2) When only 40% of the negative classes with the farthest distance to the positive classes were used to train the models, we could significantly improve the performance of the classifiers.
3) Adaboost, RF, and SVM performed the best in the psoriasis and eczema data sets, while SVM, IBk, and NNet performed well for the nodular prurigo data set. The GBM, OneR and NB algorithms were the worst performers in all 3 data sets.

However unlike in the PuvaMate data set, if all the necessary features that are required to represent a phototherapy record have been captured, in the future, we may be capable of building a more generalized model that will enable a better prediction of the occurrence of an adverse event. When the important attributes that are currently missing from the data sets, such as the psoriasis area and severity index (PASI) [42], Dermatology Life Quality Index (DLQI) [43], have been collected for a considerable amount of records, experiments need to be carried out in the future to check if there's an impact on the performance of the prediction. In order to explore the relationship among patients, social network analysis techniques with clustering algorithms can be used for these applications.

References

[1] E. Racz and E. P. Prens, "Phototherapy and photochemotherapy for psoriasis" *Dermatologic Clinics*, vol. 33, no. 1, pp. 79-89, 2015.

[2] M. Nakamura, B. Farahnik, and T. Bhutani, "Recent advances in phototherapy for psoriasis" *F1000Research*, vol. 5, 2016.

[3] J. Wan, K. Abuabara, A. B. Troxel *et al.*, "Dermatologist preferences for first-line therapy of moderate to severe psoriasis in healthy adult patients," *Journal of the American Academy of Dermatology*, vol. 66, no. 3, 2012, pp. 376-86.

[4] H. Honigsmann, "Phototherapy for psoriasis," *Clinical and Experimental Dermatology*, vol. 26, no. 4, 2001, pp. 343-50.

[5] S. Laube and S. A. George, "Adverse effects with PUVA and UVB phototherapy," *Journal of Dermatological Treatment*, vol. 12, no. 2, pp. 101-105, 2001.

[6] Y. Matsumura and H. N. Ananthaswamy, "Toxic effects of ultraviolet radiation on skin," *Toxicology and Applied Pharmacology*, vol. 195, pp. 298-308, 2004.

[7] Health Quality Ontario, "Ultraviolet phototherapy management of moderate-to-severe plaque psoriasis: An evidence-based analysis," *Ontario Health Technology Assessment Series*, vol. 9, no. 27, pp. 1-66, 2009.

[8] S. Ong and I. Coulson, "Legal claims in English dermatological practice," *British Journal of Dermatology*, vol. 164, pp. 217-219, 2011.

[9] Phototherapy Service Guidance. (October 2016). [Online]. Available: http://www.bad.org.uk/shared/get-file.ashx?itemtype=documentandid=4151

[10] P. Sheenal and H. Patel, "Survey of data mining techniques used in healthcare domain," *International Journal of Information*, vol. 6, 2016.

[11] J. Neesha and W. Husain, "Data mining in healthcare-A review," *Procedia Computer Science*, vol. 72, pp. 306-313, 2015.

[12] A. Parvez, S. Qamar, and S. Q. A. Rizvi, "Techniques of data mining in healthcare: a review," *International Journal of Computer Applications*, vol. 120, no.15, 2015.

[13] R. Madhura, W. Deepanker, and N. Pathak, "A survey on implementation of machine learning techniques for dermatology diseases classification," *International Journal of Advances in Engineering and Technology*, vol. 8, no. 2, p. 194, 2015.

[14] Y. Yoon, R. W. Brobst, P. R. Bergstresser, and L. L. Peterson, "Automatic generation of a knowledge-base for a dermatology expert system," in *Proc. Third Annual IEEE Symposium on Computer-Based Medical Systems*, June 1990, pp. 306-312.

[15] E. D. Übeyli, "Combined neural networks for diagnosis of erythemato-squamous diseases," *Expert Systems with Applications*, vol. 36, no. 3, pp. 5107-5112, 2009.

[16] C. L. Chang and C. H. Chen, "Applying decision tree and neural network to increase quality of dermatologic diagnosis," *Expert Systems with Applications*, vol. 36 no. 2, pp. 4035-4041, 2009.

[17] B. Karlik and G. Harman, "Computer-aided software for early diagnosis of eerythemato-squamous diseases," in *Proc.* I*EEE XXXIII International Scientific Conference on Electronics and Nanotechnology*, April 2013, pp. 276-279.

[18] S. Sarhan, E. Elharir, and M. Zakaria, "A hybrid rough-neuro model for diagnosing erythemato-squamous diseases," *IJCSI Int. J. Comput. Sci. Issues*, vol. 11 no. 1, 2014.

[19] L. Nanni, "An ensemble of classifiers for the diagnosis of erythemato-squamous diseases," *Neurocomputing*, vol. 69 no. 7, pp. 842-845, 2006.

[20] E. D. Übeyli, "Multiclass support vector machines for diagnosis of erythemato-squamous diseases," *Expert Systems with Applications*, vol. 35, no. 4, 2008, pp. 1733-1740.

[21] J. Xie and C. Wang, "Using support vector machines with a novel hybrid feature selection method for diagnosis of erythemato-squamous diseases," *Expert Systems with Applications*, vol. 38, no. 5, 2011, pp. 5809-5815.

[22] D. Giveki, H. Salimi, A. A. Bitaraf, and Y. Khademian, "Detection of erythemato-squamous diseases using AR-CatfishBPSO-KSVM," *Signal and Image Processing*, vol. 2, no. 4, 2011, p. 57.

[23] M. J. Abdi and D. Giveki, "Automatic detection of erythemato-squamous diseases using PSO–SVM based on association rules," *Engineering Applications of Artificial Intelligence*, vol. 26, no. 1, 2013, pp. 603-608.

[24] T. Mroczek, W. Paja, L. Piatek, and M. Wrzesie, "Classification and synthesis of medical images in the domain of melanocytic skin lesions," *Human System Interactions*, pp. 705-709, May 2008.

[25] K. Polat and S. Güneş, "A novel hybrid intelligent method based on C4. 5 decision tree classifier and one-against-all approach for multi-class classification problems," *Expert Systems with Applications*, vol. 36, no. 2, 2009, pp. 1587-1592.

[26] K. K. Manjusha, K. Sankaranarayanan, and P. Seena, "Prediction of different dermatological conditions using Naïve Bayesian classification," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 4, no. 1, 2014.

[27] S. Aruna, L.V. Nandakishore, and S. P. Rajagopalan, "A hybrid feature selection method based on IGSBFS and naïve bayes for the diagnosis of erythemato-squamous diseases," *International Journal of Computer Applications*, vol. 41, no. 7, 2012.

[28] K. Danjuma and A. O. Osofisan, "Evaluation of predictive data mining algorithms in erythemato-squamous disease diagnosis," arXiv preprint arXiv:1501.00607, 2015.

[29] H. Cataloluk and M. Kesler, "A diagnostic software tool for skin diseases with basic and weighted K-NN," *Innovations in Intelligent Systems and Applications (INISTA)*, pp. 1-4, July 2012.

[30] A. M. Elsayad, "Diagnosis of erythemato-squamous diseases using ensemble of data mining methods," *ICGST-BIME Journal*, vol. 10, no. 1, 2010, pp. 13-23.

[31] D. K. Sharma and H. S. Hota, "Data mining techniques for prediction of different categories of dermatology diseases," *Journal of*

Management Information and Decision Sciences, vol. 16, no. 2, p. 103, 2013.

[32] Mitchell and M. Tom, “Machine learning,” The Mc-Graw-Hill Companies, Inc, 1997

[33] M. M. Breunig, H. P. Kriegel, and R. T. Ng, J. Sander, “LOF: identifying density-based local outliers,” *ACM Sigmod Record*, 2000.

[34] L Al Shalabi, Z Shaaban, and B Kasasbeh, “Data mining: A preprocessing engine,” *Journal of Computer Science*, 2006

[35] R. Kumar and R. Verma, “Classification algorithms for data mining: A survey,” *International Journal of Innovations in Engineering and Technology (IJIET)*, vol. 1, no. 2, pp.7-14, 2012.

[36] Y. Huang and T. Kechadi. “An effective hybrid learning system for telecommunication churn prediction,” *Expert Systems with Applications*, vol. 40, no. 14, pp. 5635-5647, 2013.

[37] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” *IJCAI*, vol. 14, no. 2, pp. 1137-1145, August 1995.

[38] B. Huang, M.T. Kechadi, and B. Buckley, “Customer churn prediction in telecommunications,” *Expert Systems with Applications*, vol. 39, no. 1, pp. 1414-1425, 2012.

[39] Y. M. Chyi, “Classification analysis techniques for skewed class distribution problems,” Department of Information Management, National Sun Yat-Sen University, 2003.

[40] J. A. Hanley and B. J. McNeil, “The meaning and use of the area under a receiver operating characteristic (ROC) curve,” *Radiology*, vol. 143, no. 1, pp. 29-36, 1982.

[41] C. Goutte and E. Gaussier, “A probabilistic interpretation of precision, recall and F-score, with implication for evaluation,” *ECIR*, vol. 5, pp. 345-359, March 2005.

[42] P. C. M. Van de Kerkhof, “The psoriasis area and severity index and alternative approaches for the assessment of severity: Persisting areas of confusion,” 1997.

[43] A. Y. Finlay and G. Khan, "Dermatology life quality index (DLQI) - a simple practical measure for routine clinical use,” *Clinical and Experimental Dermatology*, vol. 19, no. 3, pp. 210-216, 1994.

S. Mohamed is currently a research master student at University College Dublin (UCD). She obtained her BSc in computer science international offered by UCD at National School of Business Management (NSBM) Sri Lanka. She was awarded with a higher diploma in computer based information systems and a diploma in computer system design by National Institute of Business Management (NIBM) Sri Lanka.

She has worked as a software developer at NIBM Sri Lanka in 2014 and as a researcher for the Irish Skin Foundation for the last two years. Her current research interests include application of data mining and machine learning techniques on healthcare data.

Anne-Marie Tobin is a consultant dermatologist in Tallaght Hospital and clinical senior lecturer in Trinity College Dublin. She is dual-qualified in medicine and pharmacy from Trinity College Dublin and completed a PhD in translational medicine from University College Dublin. Dr Tobin’s clinical interests are in inflammatory skin disease (psoriasis and hidradenitis suppurativa) and skin cancer. She runs a systemic clinic for patients with psoriasis and eczema and a hidradenitis suppurativa clinic. She also provides skin cancer screening for renal transplant patients and a pigmented lesion clinic.

Dr Tobin is the clinical lead for the National Clinical Programme in Dermatology and clinical lead for the Photodermatology Group Dr Tobin is involved in clinical research and clinical trials in psoriasis, hidradenitis suppurativa and eczema and has received the Jacobs Medal(Royal Academy of Medicine Ireland) and the Burrows Cup (Irish Association of Dermatologists) for laboratory research.

Alan Irvine graduated in medicine from Queen’s University Belfast in 1991. He completed dermatology training in Belfast in July 1999, followed by fellowships in Great Ormond Street Children’s Hospital and Children’s Memorial Hospital Chicago, where he was a fulbright scholar. He was appointed as an attending dermatologist in Our Lady’s Children’s Hospital and St. James’s Hospital, Dublin, Ireland in October 2002. He is a professor in dermatology in Trinity College Dublin.

His research interests are in epithelial genetics, disease mechanisms in, and therapy of atopic dermatitis. His work on the genetics of atopic dermatitis with long term collaborative partner Irwin McLean has helped refocus attention on the role of the skin barrier in the pathogenesis of this disease and of allergic disease in general. He is funded by the National Children’s Research Centre and the Wellcome Trust amongst others and has attracted approximately €12M research funding. He has published more than 210 peer reviewed articles, cited more than 17, 500 times, with an H-index of 57. Alan has received several international awards including the Paul Gerson Unna Prize (German Dermatology Society), and the MB Sulzberger (AAD). Dolovich (AAAAI), Watson-Smith (RCP Edinburgh), RW Goltz (U Minnesota), Lectureships

Dmitri Wall is a University College Dublin medical graduate and a member of the Royal College of Physicians. After completing higher specialist training with the Royal College of Physicians of Ireland, he worked as a consultant dermatologist in St James’s Hospital, Dublin and currently he is a clinical fellow in hair and nail disease with Professor Rod Sinclair in Melbourne, Australia.

Since 2010 Dr Wall has developed a special interest in health informatics, completing an MSc in health informatics in Trinity College Dublin in 2015 with distinction, winning the Health Informatics Society of Ireland Award for contribution to Health Informatics by a student for his thesis.

Dr Wall contributes to a number of national and international committees and projects, such as the European Reference Network for Rare and Undiagnosed Skin Disease (ERN Skin), for which he is the eHealth lead, and the British Association of Dermatologists Health Informatics Sub Committee. Amongst other publications, he has authored sections of the PAtient REgistries iNiTiative (PARENT) EU “Methodological guidelines and recommendations for efficient and rationale governance of patient registries.

Neil O’Hare is a professor of health informatics in University College Dublin and group chief information officer for the Ireland East Hospital Group, which is the largest group of acute hospitals in Ireland. He was previously a chief physicist in St. James’s Hospital (Dublin) and his research interests are in health informatics, medical imaging and UV phototherapy.

M-Tahar Kechadi is professor of computer science in University College Dublin (UCD), Ireland. He was awarded PhD and master degree in computer science from University of Lille 1, France. He is currently a principal investigator in the INSIGHT Centre for Data Analytics. His research interests span the areas of big data analytics, distributed data analytics, heterogeneous distributed systems, cloud computing, and forensic computing and cybercrime investigations. The core and central focus of his research for the last decade is how to manage and analyse data quickly and efficiently at larger scale. He is in the editorial board of the Journal of Future Generation of Computer Systems and of IST Transactions of Applied Mathematics-Modelling and Simulation. He is a member of the International Knowledge Cloud Consortium (IKCC). He is full member at CERN. He is a member of ACM and IEEE.