2015 IEEE Symposium Series on Computational Intelligence

A Novel Ontology and Machine Learning Inspired Hybrid Cardiovascular Decision Support Framework

Kamran Farooq
Department of Computing Science
and Mathematics
University of Stirling
Stirling, UK
Email: kfa@cs.stir.ac.uk

Amir Hussain
*Department of Computing Science
and Mathematics
University of Stirling
Stirling, UK
†Visiting Professor
School of Computer Science and Technology
Anhui University
Hefei, Anhui, China
Email: ahu@cs.stir.ac.uk

Bin Luo
School of Computer Science and Technology
Anhui University
Anhui, China
Email: luobinahu@gmail.com

Warner Slack
Harvard Medical School
Beth Israel Clinical Informatics Division
Boston, US
Email: wslack@bidmc.harvard.edu

Abstract—**Healthcare information management systems (HIMS) have a substantial amount of limitations such as rigidity and nonconformity to complex clinical processes like Electronic Healthcare records and effective utilisation of clinical practice guidelines to help provide effective clinical decision support. The conventional healthcare systems suffer from a general lack of intelligence, they are successful in offering basic patient management capabilities, but they do not offer consistent and holistic decision support capabilities for clinicians working under tight deadlines in a fast paced environment. The conventional healthcare information management systems are designed using branching logic based rigid architectures, which are hard to maintain and upgrade without considerable labour intensive effort. The proposed ontology and machine learning driven hybrid clinical decision support framework comprises of two key components (1) ontology driven clinical risk assessment and recommendation system and (2) machine learning driven prognostic system. The key aim of our research is to utilise information collected through the knowledge based ontology driven clinical risk assessment and recommendation system and non-knowledge based/evidence based machine learning driven prognostic system to deliver a holistic clinical decision support framework in the cardiovascular domain. The ontology driven clinical risk assessment and recommendation system could be used as a triage system for cardiovascular patients as a preventative solution, this could help clinicians prioritise patient referrals after reviewing a snapshot of patient's medical history (collected through an ontology driven intelligent context aware information collection using standardised clinical questionnaires) containing patient demographics information, cardiac risk scores, cardiac chest pain score, medication and recommended lab tests details. The machine learning driven prognostic system is developed using a chest pain clinical case study identified by the consultant cardiologist, Professor Stephen Leslie from Raigmore Hospital in Inverness. The key aim of this clinical case study UK is to provide a clinical decision support mechanism for Raigmore Hospital's Rapid Access Chest Pain Clinic (RACPC) patients by combining evidence, extrapolated through legacy patient data (based on machine learning driven techniques) to facilitate evidence based cardiovascular preventative care. The machine learning driven prognostic system provides cardiac chest pain prognosis through a cardiac chest pain specific prognostic model which is validated through consultant cardiologist from Raigmore Hospital. The cardiac chest pain prognostic model could help clinicians diagnose cardiac chest pain patients efficiently and could also help clinicians reduce load on overly prescribed angiography treatment in a cost effective manner. Additional two clinical case studies in the heart disease and breast cancer domains are considered for the development and clinical validation of the machine learning driven prognostic system. The proposed novel ontology and machine learning driven hybrid clinical decision support framework will also be validated in other application areas.**

I. Introduction

Clinical decision support operations are an integral part of modern healthcare management systems. The information encoded in these intelligent clinical systems is used for inference purposes to improve clinical effectiveness and quality of healthcare. Automated cardiovascular decision support systems are being deployed in hospitals and primary care organizations in order to meet the ever growing clinical needs of prognosis in the areas of cardiovascular disease and Coronary heart disease. Computerized decision support strategies have already been implemented successfully in several areas of cardiovascular care [1]. These applications are being used as part of the extension of clinical informatics infrastructure in the UK and US. These systems are being used in both primary

978-1-4799-7560-0/15 $31.00 © 2015 IEEE
DOI 10.1109/SSCI.2015.122

and secondary care settings for providing efficient healthcare delivery to its patients. In order to capitalise on the benefits provided by cardiovascular decision support systems, a strong foundation in evidence-based care and well-established clinical practice guidelines (CPGs) have to be considered to ensure clinical governance in the next generation clinical systems.
Computational Intelligence and healthcare informatics are transforming healthcare to a proactive p4 medicine that is predictive, preventive , personalised and participatory. Computational intelligence - holistic, and integrative approach has given rise to machine learning driven predictive modelling. Machine learning refers to a type of artificial intelligence algorithm designed to identify patterns in input data, such as patient characteristics, in order to perform complex classification tasks. Machine Learning based clinical decision support systems can avoid the bottleneck of knowledge acquisition because knowledge is directly learned through the clinical data. In addition, ML-based clinical decision support systems are able to give recommendations that are generated by non-linear forms of knowledge, and are easily maintainable by simply adding new cases [2].

In [3], considered a clinical use case of predicting cases of POAF (post atrial fibrillation) following CABG (coronary artery bypass graft) surgery. Predictive features such as age, body mass index (BMI), and systolic blood pressure (SBP), were selected to predict whether patients could develop AF (Atrial Fibrillation) during the recovery period following CABG. Authors utilised k-NN algorithm in their experimental setups. The k-NN algorithm was provided with a number of labelled training samples, which in this case consisted of a set of three features for a series of patients who have undergone CABG in the past, as well as their clinical outcome in terms of AF occurrence, or lack thereof, during the recovery period.

A novel cardiovascular decision support framework was presented in [4], with a view to provide a triage mechanism for primary and secondary care clinicians in the UK and US hospitals. The aim of their novel clinical decision support framework was to help improve the diagnostic and performance capabilities of Rapid Access Chest Pain Clinic (RACPC), by reducing delay and inaccuracies in the cardiovascular risk assessment of patients with chest pain by helping clinicians effectively distinguish acute angina patients from those with other causes of chest pain. The key components of the proposed framework were presented in [4], [5] and [6]. Their proposed framework is also capable of learning from legacy patient data containing missing information and its effective utilisation in the over all clinical decision making was demonstrated in [7]. Their work was further extended through the exploitation of RACPC (chest pain) patient dataset in [8] and [9], authors have demonstrated the clinical effectiveness of the hybrid clinical decision support mechanism through utilisation of ontology and machine learning driving techniques. The proposed framework was also validated using real chest pain patient data provided by Raigmore Hospital in the UK.

Hybrid clinical decision support systems in various clinical domains are playing an important role in assisting medical professionals in making decisions, based on current patient data and best practices encoded in a rule base, in scenarios where there may be missing data. In [10] a novel context-aware hybrid reasoning framework through the exploitation of fuzzy rule-based reasoning has been proposed to achieve pervasive healthcare in smart home environment. Authors presented a personalised, flexible and extensible hybrid reasoning framework for context aware real-time assistant (CARA)in a smart home environment which provides context-sensitive sensor data as well anomaly detection mechanisms that supports Activity of Daily Living (ADL) analysis and alert generation. They deployed a pervasive healthcare system in a lab setting comprises of wearable wireless sensors, smart home sensors, remote monitoring system and a data reviewing system. In [11] and [12], an ontology inspired approach was utilised to develop a clinical decision support framework for lung cancer patients. They exploited ontological inference using dynamic logic reasoner to create patient-specific treatment arguments by automatically grouping patients based on set of guidelines (British Thoracic Society Guidelines into Lung Cancer Assistant system) written in the ontology. A novel feature of their proposed lung cancer assistant property was its ability to provide a rule-based and probabilistic decision support within a single platform. The guideline-based CDS is based on clinical guideline rules, while the probabilistic CDS is based on a Bayesian network trained on the English Lung Cancer Audit Database (LUCADA). Matt-Mouley Boumrane from the University of Glasgow implemented an ontology driven approach for the development of clinical decision support system in the pre-operative risk assessment domain. In [13], they combined a preventative care software system in the pre-operative risk assessment domain with a decision support ontology developed with a logic based knowledge representation formalism. In [14], [15], [16], Cambria et al utilised ontology and semantically inspired sentiment mining techniques to develop patient centric applications as part of providing cost effective preventative care mechanism for patients. In [17], an ontology driven approach has been utilised for the diagnosis of mild cognitive impairment (MCI), specialised clinical knowledge is coded into an ontology for the construction of a rule set utilised by machine learning algorithms. The reasoning engine is also exploited to automatically distinguish MCI patients from normal ones. The rule set was trained by MRI data of 187 patients, support vector machine (SVM), Bayesian Network (BN) and back propagation (BP) neural networks were used for the construction of reasoning rules. Their evaluation results suggested that their approach would be useful to assist clinicians in effectively diagnose patients with mild cognitive impairment. Their framework demonstrated that domain ontology combined with machine learning techniques are useful in diagnosing complex chronic illnesses. The paper is organised as follows. Section 2 focuses on the Novel Ontology and Machine Learning Driven Hybrid Cardiovascular Decision Support Framework. Section 3, discusses the Methodology, followed by Results in section 4. Finally overall findings are concluded in section 5.

II. The Proposed Framework

We propose an ontology and machine learning driven hybrid clinical decision support framework in Figure 1 to provide a clinical decision support mechanism for primary and secondary care clinicians in the UK and US hospitals. The proposed clinical decision support framework could also be utilised by patients in order to build their medical records which could be used as part of triage mechanism. The ontology and machine learning driven hybrid clinical decision support framework comprises of two key components to provide a cardiovascular preventative care solution. The key components of the proposed framework are as follows:

1) Ontology Driven Clinical Risk Assessment and Recommendation System.
2) Machine Learning Driven Prognostic System.

Ontology driven techniques help system developers build more scalable, cost effective, reusable and modularised clinical decision support components which can be integrated in an intelligent manner to deliver framework functionality. The proposed ontology and machine learning driven hybrid clinical decision support framework could be exploited in the disease management of other chronic illnesses by updating clinical rules encoded in the clinical rules engine and domain specific ontologies without altering the interface, database and the framework design. The key components of the framework are reusable (through mapping of disease specific questionnaire ontology, recommendation ontology (based on clinical rules for recommendation of lab tests and medication) and NICE guidelines).

Our proposed ontology and machine learning driven hybrid clinical decision support framework builds on Matt-Mouley Bouamrane's clinical decision support framework by incorporating machine learning driven prognostic system and a refined ontology driven clinical risk assessment and recommendation system in the cardiovascular domain. The ontology driven clinical risk assessment and recommendation system provides an ontology driven intelligent context aware information collection for conducting patient interviews in order to gather patient medical records which are utilised by NICE/Expert driven clinical rules engine for the cardiac risk scores calculation for various cardiovascular diseases. Patient medical records are transformed into patient semantic profile (to alleviate interoperability issues) using answers provided in the patient interviews. The patient semantic profile combines with recommendation ontology is utilised for the recommendation of lab tests and prescription of medication for cardiovascular patients. The machine learning driven prognostic system and ontology driven risk assessment and recommendation system are integrated as a complete system in order to provide a cardiovascular preventative care solution for patients as well as primary and secondary care clinicians using dedicated graphical user interfaces for clinicians and patients.

The proposed machine learning driven prognostic system is developed (based on legacy patient data for RACPC patients) with a view to provide RACPC specific cardiac chest pain prognosis as part of cardiovascular preventative care. The machine learning driven prognostic system is also validated in different application areas and disease specific prognostic models in the chest pain, heart disease and breast cancer have also been developed and deployed online for further clinical trials and clinical validation. The proposed ontology and machine learning driven hybrid clinical decision support framework provides a learning mechanism, which is built using machine learning techniques. The learning facility is provided through exchange of patient data (collected in patient interviews conducted through ontology driven intelligent context aware information collection) amongst the machine learning driven prognostic system and the ontology driven clinical risk assessment and recommendation system, specifically from the cardiac chest pain and heart disease risk scores calculation perspective. The machine learning driven cardiac chest pain prognostic model's risk score calculation (cardiac chest pain risk score) along with other cardiac risk scores are provided through the integration of both components with a view to provide a holistic view of multiple cardiac risk scores calculation for each patient.

The proposed clinical decision support framework could be used as a cardiovascular preventative care solution for automatically conducting patient pre-visit interviews. It will not replace a human doctor, but would be used before a hospital visit to prepare the patient, deliver educational materials, cardiac risk assessment scores, cardiac chest pain and heart disease scores and pre-order appropriate tests, making better use of doctor-patient consultation time. It could also be used as a triage system to help clinicians prioritise patient appointments after reviewing a snapshot of patient's medical history (collected through an ontology driven intelligent context aware information collection using standardised clinical questionnaires) containing patient demographics information, cardiac risk scores, cardiac chest pain and heart disease risk scores, recommended lab tests and medication details. Additional two clinical case studies in the heart disease and breast cancer domains are considered for the development and clinical validation of the machine learning driven prognostic system.

III. Methodology

A. Ontology Driven Clinical Risk Assessment and Recommendation System for Cardiovascular Preventative Care (OD-CRARS)

The proposed ontology driven clinical risk assessment and recommendation system is developed using a hybrid approach based on ontology driven techniques and clinical rules engine. Ontology driven approach is exploited in the development of Intelligent Context aware Information Collection Component and recommendation of lab tests and medication is carried out through the Recommendation Ontology based on clinical rules written by consultant cardiologist, Professor Calum MacRae Slack from Harvard Medical School. A dedicated clinical rules engine (through the utilisation of NICE guidelines and clinical rules written by clinical domain expert) is developed to carry

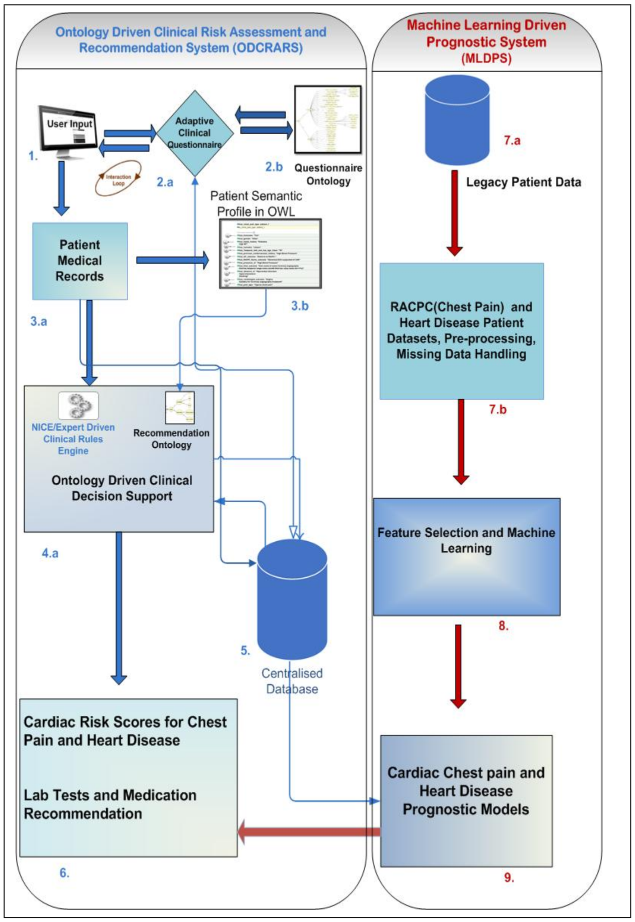

Fig. 1: A Novel Ontology and Machine learning-driven hybrid Clinical Decision Support Framework for Cardiovascular Preventative Care.

out various operations, the key operations of the clinical rules engine are ad follows :

1) To carry out cardiac risk assessment (for calculating global, absolute and relative risk scores) for various cardiovascular diseases like coronary heart disease (CHD), Myocardial Infarction (MI) etc.
2) To implement access control for system users (patients and clinicians).
3) Dynamic creation of the systematic examination questionnaires to conduct patient clinical reviews.

1) Ontology driven intelligent context aware information collection component: Healthcare information systems are widely used all over the world to alleviate diverse healthcare demands and supply gap [18]. Information collection through questionnaires based systems are fundamental to the core functioning of healthcare information management systems. With the advent of recent success of electronic healthcare records globally, information collection through intelligent means has now become one of the most important components of modern healthcare systems. In modern patient interviewing systems, one of the main challenges for the system developers is to develop usable, context sensitive interfaces so that patients could be involved in the information collection process in order to collate their medical records. Healthcare resources in most parts of the world are stretched to the limit which is why healthcare providers' main focus is to develop preventative care solutions based on patient medical records. Patient triage systems are more in demand than ever before which is why they are an essential component of healthcare information management systems to ensure safe record keeping of patient medical records along with clinical risk assessment information, details of recommended lab tests and medication as part of preventative care measure. Patient triage systems help clinicians optimise the referral process and enable them to utilise their consultation time more efficiently by focussing on providing more direct care for their patients.

2) Patient Medical Records: Patient medical records as shown in 3a in Figure 1 are generated using patient answers collated through the Ontology Driven Intelligent Context Aware Information Collection component. This information containing patient demographics and clinical review details is saved in the centralised database for its utilisation by clinical rules engine for clinical risk assessment purposes.

Patient medical records are generated using patient answers and it provides a snapshot of patient's medical history. These medical records are used by clinical rules engine for the cardiac clinical risk assessment of patients to check cardiac risk scores of various cardiovascular diseases which includes, Coronary Heart Disease, Myocardial Infarction etc. Cardiac global, absolute and relative risk scores are calculated using set of clinical rules executed by Java rules engine called jess.

3) Patient Semantic Profile: The information representation at the patient records level lacks flexibility in its structure and due to their static nature, patient medical records do not carry any intrinsic meaning. The information collection based on an ontology driven approach provides an opportunity to simultaneously generate patient semantic profile through a clinical ontology in order to preserve the semantics. The important benefit of utilising this approach is that patient medical records being a single repository of information could be used to provide a number of services within the proposed framework.

4) Ontology Driven Clinical Decision Support: The ontology driven clinical risk assessment and recommendation as in 4a in Figure 1 system provides clinical decision support mechanism based on a Recommendation ontology(for lab tests recommendation and medication prescription)and clinical rules engine. The proposed ontology driven clinical risk assessment and recommendation system shown in Figure 1 aims to provide an online cardiovascular preventative care solution with a view to enhance the doctor-patient consultation process effectively by facilitating patients to complete a standardised clinical review of their current and past medical histories prior to hospital visits. These reviews are conducted through the ontology driven intelligent context aware information collection component. The recommendation system exploits information held in the patient medical records and patient semantic profile to carry out clinical decision support operations using clinical rules engine and Recommendation ontology for the recommendation of lab tests and prescription of medication.

The ontology driven clinical risk assessment and recommendation system collects structured information (driven through context- sensitive standardised clinical questionnaires) through the web front end using a systematic medical extermination technique known as the patient clinical review and then provides a suggested list of laboratory tests and medication using domain specific recommendation ontology.

In addition to these cardiac risk assessment, the ontology driven cardiovascular risk assessment and recommendation system and the machine learning driven prognostic system are integrated to provide cardiac chest pain and heart disease scores through the recommendation system's front end in order to provide a holistic cardiovascular preventative care solution by providing clinicians an array of cardiovascular risk assessment scores for various cardiovascular diseases, recommendation of lab tests and medication prescription. The ontology driven clinical risk assessment and recommendation system as part of the proposed framework is deployed online for further clinical trials and validation work.

B. Machine Learning Driven Prognostic System (MLDPS)

An iterative development process based on machine learning and feature selection has been utilised in the development of machine learning driven prognostic models. This machine learning driven prognostic model construction process is general enough to handle a variety of healthcare datasets which will enable researchers to develop effective non knowledge-based/evidence based clinical decision support systems. The key stages of the machine learning driven prognostic model construction process are given as follows:

1) Data Acquisition
2) Data Pre-Processing
3) Feature Selection
4) Prognostic Model Construction
5) Prognostic Model Validation and Evaluation
6) Online Clinical Prognostic Model

Decision Tree and Logistic Regression models have been exploited using Forward Selection and Backward selection techniques to help build optimum models using the best feature set. As it can be seen clearly that the clinical risk factors (highlighted in bold) like ANG (Angio Result) and INA (Initial Assessment) are showing as significant among all four experiential setups which suggest that the initial assessment of chest pain patients along with their Angio results are the most important clinical risk factors in the risk assessment of RACPC patients.

As per our comparative analysis shown in Table I of different machine learning techniques, based on various experimental setups, patient's "Angio Result" and "Initial Result" outcomes (as shown above) could be deciding factors for patient's referral through to the next stage of cardiac assessment. RACPC patients get referred through different clinical pathways as per findings in each phase, there are exit points in each stage for patients with non cardiac symptoms, patients with cardiac related chest pain get referred through the clinical pathway called "Presentation Suggests Angina" for further clinical tests like ETT, Perfusion Scan and ETT, followed by angiography for patients who are unable to do ETT or with abnormal ECG (suspicious of CAD).

TABLE I
WEIGHTED CLASSIFICATION ACCURACIES WITH COMMON CLINICAL VARIABLES (HIGHLIGHTED IN BOLD) IN EACH ITERATION.

Iteration	FS-DT		BS-DT		FS-LR		BS-LR	
1	**ANG**	64.78	MPS	76.02	**INA**	66.05	ETT	74.34
2	**INA**	71.72	NOC	76.51	**AGE**	67.81	CHL	74.27
3	**CT**	77.34	CHL	76.83	**ANG**	71.94	DAB	74.42
4	**ETT**	78.43	SMR	77.11	**SEX**	72.6789	NOC	74.45
5	DAB	78.43	ETT	77.15	**MPS**	73.3831	MPS	73.89
6	SEX	78.46	DAB	76.87	**YOS**	74.0550	SMR	73.30
7	HPT	77.59	YOS	73.64	NOC	73.91	HPT	73.81
8	CHL	76.96	AGE	75.00	HPT	73.99	YOS	73.67
9	MPS	74.24	PWY	77.30	PWY	74.30	CT	72.71
10	NOC	73.96	SEX	76.62	ETT	74.30	PWY	72.67
11	PWY	76.37	HPT	77.34	CT	74.30	SEX	71.94
12	SMR	75.33	CT	71.72	SMR	74.42	**INA**	68.17
13	AGE	75.11	**INA**	64.7867	DAB	74.13	**ANG**	62.06
14	YOS	75.11	**ANG**		CHL	74.16		

C. Prognostic Model Development

1) Logistic Regression: After dataset preparation, a number of clinical variables are extracted through the legacy patient data for the prognostic model construction phase. The vector of selected candidate independent clinical variables is called X and B which is a vector of coefficients. Depending on the desired output, in most cases , linear and logistic regression are able to provide prognostic models with a reasonable level of accuracy.

Prognoses problems in healthcare can be distinguished by the form of the output space Y. If the predictive class is numeric or continuous (i.e. Y = R, the real line), then the prognostic problem is a regression problem (e.g. predicting a physical measurement such as height) [19]. If the predictive class is discrete (i.e.Y = 0, 1, . . . , K 1) then we have a classification problem.

In all of our clinical case studies , classification problems fall into this category (i.e.$y^{(m)}\epsilon$ 0,1), in this case the model $\hat{y} = f(B, X)$ is the probability of an input data value belonging to a certain class. A threshold is generally applied to the probability calculated from the model in order to predict the class to what the data point is expected to belong to. The threshold is often used to quickly evaluate the accuracy of the model. Besides being needed in the practical usage of the model, the threshold is also commonly used to quickly evaluate the accuracy of the model (i.e. once a threshold has been selected, the accuracy of the model is worked out using the receiver operating characteristic (ROC) curves in terms of providing sensitivity, specificity values for True Positive

(TP) , True negative (TN) , False Positive (FP) and False Negatives (FN). TP and FP values are utilised in calculating the precision of the prognostic model ,at the same time recall could be calculated by utilising TP divided by sum of TP + FN. Details of prognostic models evaluation will be provided in the forthcoming section on model evaluation in IV-A.

A logistic regression model is denoted as

$$f(B,X) = \frac{1}{1+e^{-(b_0+b_1x_1+b_2x_2....+b_nx_n)}} = \frac{1}{1+e^{-BTX}} \quad (1)$$

Table II gives an account of the feature selection techniques which are utilised in the RACPC clinical case study using three experimental setups based on LR, DT and SVM. It is to be noted that in the case of DT+BS, DT+SFFS and SVM+SFFS experimental setups, minimal amount of features are selected to classify the patient data. In all of these experimental setups, clinical variables such as 14 (Angio Results), 10 (Initial Assessment) and 12 (CT Result) were found common among some of the DT and SVM based experimental setups. This means that using the initial assessment, CT Scan and Angio results, clinicians will be able to diagnose cardiac chest pain patients with a classification accuracy of 78.63 % which has been attained using DT+SFFS experimental setup. At the same time, more transparent approaches like LR combined with BS wrapping method requires 10 clinical variables to classify patient data with 74.68 % classification accuracy. Due to imbalanced and limited RACPC datasets, high classification accuracies (with low standard errors) could not have been achieved. In spite of the data sparsity and missing data issues, we were able to achieve good results through the utilisation of state of the art machine learning and feature selection techniques. This clinical case study was carried out under the supervision of RACPC clinical domain expert, machine learning results were analysed and way forward towards the development of online prognostic models (based on transparent LR approach) was agreed among the project stakeholders. Details of online RACPC prognostic models will be provided in the forthcoming sections.

TABLE II
EXPERIMENTAL SETUPS BASED ON MACHINE LEARNING CLASSIFIERS AND FEATURE SELECTION TECHNIQUES.

	Setups	**Selected Features**	**Accuracy**
1	**LR+FS**	10,4,14,6,13,3,2,8,5,11,12,1	74.68%
2	**LR+BS**	1,3,4,5,6,8,10,12,13,14	74.68%
3	**LR+ED**	All	74.36%
4	**LR+SFFS**	10,4,14,6,13,3	74.20%
5	**LR+P-Value**	14,4,10,6,8,13,7,9,5,1,12,1,3,11	74.36%
6	**LR+mRMR**	14,4,10,5,6,8,13,7,12,9,11,1,2,3	74.36%
7	**DT+FS**	14,10,12,11,7,6	77.84%
8	**DT+BS**	10,12,14	77.68%
9	**DT+ED**	All	75.47%
10	**DT+SFFS**	14,10,12,11	78.6% 3
11	**DT+P-Value**	14,4,10,6,8,13,7,9,5,1,12,2,3,11	74.52%
12	**DT+mRMR**	14,4,10,5,6,8,13,7,12,9,11,1,2,3	75.00%
13	**SVM+FS**	14,10,12,6,11,5,4,13,9,3,8	78.16%
14	**SVM+BS**	3,4,5,6,8,9,10,12,13,14	78.32%
15	**SVM+ED**	All	77.05%
16	**SVM+SFFS**	14,10,12	77.37%
17	**SVM+P-Value**	14,4,10,6,8,13,7,9,5,1,12,2,3,11	77.05%
18	**SVM+mRMR**	14,4,10,5,6,8,13,7,12,9,11,1,2,3	77.05%

IV. RESULTS

A. *Prognostic Model Validation and Evaluation*

1) Prognostic Model Validation: The key aim of a classification task is to map each element of a dataset to its corresponding class amongst a number of possible ones. Logistic regression algorithm (as well as other supervised machine learning techniques) infer a model from labelled training data. The generated model is then evaluated on a separate testing set, which provides an estimate of the accuracy of the model. A correct estimation of the accuracy of a classifier (in this context, also referred to as model validation) is crucial both to predict its future predictive power and to choose among a number of possible classifier.

In the case of classification, if the number of data samples for training and testing are limited , k-fold cross validation can be utilised to predict the error rate of a learning technique. In the k-fold cross validation, a full dataset is divided randomly into k disjoint subsets of approximately equal size, in each of which the class is represented in approximately the sample properties as in the full dataset [20]. The process of k-fold cross validation works in the manner as follows:

1) Training and testing will be repeated k times on the k data subsets using k-1 partitions as the training set and the renaming partition as the testing set.
2) The classification error of this iteration is calculated by testing the classification model on the holdout set. Finally the k number of errors are added up to generate an overall error estimate. The most commonly used value of k = 10 which is the right number of folds to get the best estimate of error and some theoretical evidence also backs this value of k=10 [20].
   The leave one out cross validation (LOOCV) simply n-fold cross validation, where n is the number of samples in the full dataset. In LOOCV, each sample on its turn is discarded/left out whilst classifier is trained on the remaining n-1 data samples. Classification error for each iteration is determined on the class prediction for the holdout sample's success or failure. LOOCV utilise greater amount of data samples for training in each iteration and involves no random shuffling of samples.

2) Prognostic Model Evaluation: There are several approaches for the evaluation of classification performance. The most commonly used evaluation measure is the confusion matrix. A confusion matrix is also referred as a contingency table or an error matrix. This matrix visualizes the classifier's output in terms of representing the patterns in the classified class, while each row contains the patterns in the actual class. The overall evaluation of classifier performance is usually delivered by two characteristics: the weighted accuracy and unweighted accuracy. These two characteristics are identical

TABLE III
CONFUSION MATRIX FOR A TWO-CLASS CLASSIFICATION PROBLEM.

		Predicted Class	
		A	**B**
Actual Class	**A**	**TP** **True Positive**	**FN** **False Negative**
	B	**FN** **False Positive**	**TN** **True Negative**

only when all testing classes have the same number of data patterns.

The unweighted accuracy can be calculated as

$$A_{wa} = \frac{100 N_{cor}}{N_p} \tag{2}$$

Where N_{cor} is the number of correctly classified data patterns of all classes and N_p is the total number of data patterns.

The weighted classification accuracy is denoted by

$$A_{uw} = \frac{100}{C} \sum_{c=1}^{C} N_{cor}^{c} \tag{3}$$

Where N_{cor}^{c} is the number of correctly classified data patterns of class c and C is the number of classes.

The binary classification scenarios are most commonly used in healthcare prognostic modelling, the subjects are classified into two classes: positive and negative [8].

The binary classification scenarios are most commonly used in healthcare prognostic modelling, the subjects are classified into two classes: positive and negative [8].

The confusion matrix for binary classification is provided in Table III.

From the confusion matrix in Table III. The true positive (TP) and true negative (TN) are the correct classifications in samples of each class. A false positive (FP) is when a class B sample is incorrectly predicted as class A sample; a false negative (FN) is when a class A sample is predicted as a class B sample. Each element of a confusion matrix shows the number of test samples for which the actual class is the row and the predicted class is the column. The error rate can be calculated as $\frac{FP+FN}{TP+TN+FP+FN}$ .The error rate is a measure of the overall performance of a classifier; however a lower error rate does not necessarily mean better performance, for example in the case of imbalanced datasets, 10 samples in class A and 90 samples in class B. If TP = 5 and TN = 85, then FP = 5, FN=5, the error rate in this case is only 10%. However in the case of class A, only 50 % of the samples are correctly classified, which is why it is vital to report weighted and unweighted classification accuracies to evaluate the prognostic model's performance in terms of classification accuracies. There are a number of other evaluation metrics which can be utilised to correctly evaluate the classification results without any bias.

1) Sensitivity or Recall measures the proportions of samples in class A which are correctly classified as A. It is calculated as True Positive Rate $(TPRate) = \frac{TP}{(TP+FN)}$
2) Specificity measures the proportion of samples in class B which are correctly classified as class B. It is calculated as True Negative Rate $(TNrate) = \frac{TN}{(FP+TN)}$
3) False Positive Rate $(FPRate) = \frac{FP}{(FP+TN)} = 1 - Specificity$
4) False negative rate (FN Rate) = $\frac{FN}{(TP+FN)} = 1 - Sensitivity$
5) Positive Predictive Value(PPV) = $\frac{TP}{(TP+FP)}$, also known as precision, which measures the proportion of the claimed class A samples are indeed class A samples.

In classification tasks higher TP rate, normally co-exists with a higher FP rate and same is the case with the TN and FN rate. The receiver operating characteristic (ROC) curve is used to characterise the trade off between TP rate and FP rate. The ROC curve shown plots TP rate on the Y axis against FP rate on the X axis.With an ROC curve of a classifier , the evaluation metric is the area under the ROC curve. The larger the area under the curve (the more closely the curve follows the left-hand border and the top border of the ROC space), hence more accurate the test.The ROC curve for a perfect classifier has an area of 1.

B. Online Clinical Prognostic Model

After the detailed clinical validation and evaluation of clinical prognostic models, the next stage in the machine learning driven prognostic model construction process is to get these novel prognostic models part of the clinical workflows for primary and secondary care clinicians in the UK and US. This objective is reached through the implementation of cardiac chest pain and heart disease prognostic models as online clinical prototypes. The integration of the machine learning driven cardiac chest pain prognostic model and ontology driven clinical risk assessment and recommendation system is shown in Figure 2. Details of their implementation are beyond the scope of this paper. These online clinical risk assessment prototypes are used for the clinical validation and evaluation purposes by consultant cardiologist, Professor Stephen Leslie from Raigmore Hospital and Professor Warner Slack from Harvard Medical School as well as primary care clinician (GP) from Edinburgh who utilised heart disease prognostic models for clinical trials using real patient data. These online prognostic models could be used to collect new data for further research work and could to be used with an online training algorithm to improve performance of existing models and to optimise machine learning inputs. These online prognostic models have been developed using PHP scripts to acquire patient data and HTML front end was developed to provide the risk score.

V. CONCLUSIONS

The proposed framework will also pave the way for the development of cost effective and patient centric preventative care solutions for chronic diseases with high mortality rates,

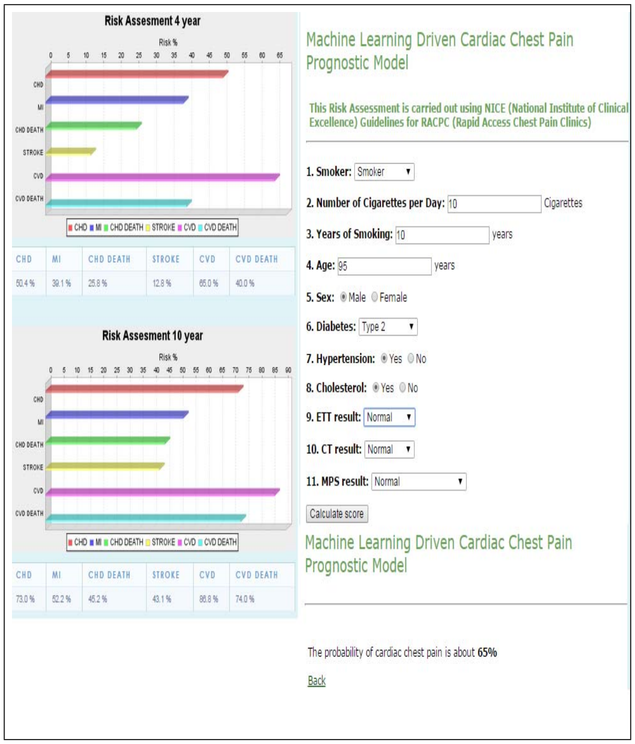

Fig. 2: System integration of the knowledge based Ontology driven Clinical Risk Assessment and Recommendation System with the evidence based Machine Learning driven Cardiac Chest Pain Prognostic model.

such as breast cancer, diabetes etc. These chronic diseases could be largely preventable through close partnership among healthcare providers, commercial partners and researchers working in the healthcare informatics domain towards developing innovative doctor-patient based interactive collaborative care solutions. The proposed framework will facilitate development of the next generation commercial clinical decision support systems with learning capabilities based on machine learning (for information exchange among key components for risk calculation for cardiac chest pain and heart disease conditions) , which could be utilised by primary and secondary care clinicians in the UK and US as a cardiovascular preventative care solution. The proposed novel ontology and machine learning driven hybrid clinical decision support framework exploits both (ontology and machine learning driven) approaches and combines both clinical expert's knowledge encoded in the form of clinical rules in the knowledge-based recommendation system and evidence based machine learning driven prognostic system (evidence extrapolated through legacy patient data) in an intelligent manner to deliver an effective clinical decision support framework for cardiovascular preventative care. Further evaluation , system implementation details and detailed results will be provided in future publications.

ACKNOWLEDGEMENT

This research project is funded by the EPSRC (Grant Ref.no. EP/H501584/1) and Sitekit Solutions Ltd. We would like to thank Professor Stephen Leslie who is a consultant cardiologist from Raigmore Hospital in Scotland for providing the required domain expertises as well as facilitating us to utilise the RACPC patient data for this case study.

REFERENCES

[1] G. J. Kuperman, A. Bobb, T. H. Payne, A. J. Avery, T. K. Gandhi, G. Burns, D. C. Classen, and D. W. Bates, "Medication-related clinical decision support in computerized provider order entry systems: a review," *Journal of the American Medical Informatics Association*, vol. 14, no. 1, pp. 29–40, 2007.

[2] C.-L. Chi, "Medical decision support systems based on machine learning," *Theses and Dissertations*, p. 283, 2009.

[3] C. Zhang and Y. Ma, *Ensemble Machine Learning*. Springer, 2012.

[4] K. Farooq, A. Hussain, S. Leslie, C. Eckl, and W. Slack, "Ontology-driven cardiovascular decision support system," in *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2011 5th International Conference on*. IEEE, 2011, pp. 283–286.

[5] K. Farooq, A. Hussain, S. Leslie, C. Eckl, C. MacRae, and W. Slack, "An ontology driven and bayesian network based cardiovascular decision support framework."

[6] ——, "Semantically inspired electronic healthcare records," *Advances in Brain Inspired Cognitive Systems*, pp. 42–51, 2012.

[7] K. Farooq, P. Yang, A. Hussain, K. Huang, C. MacRae, C. Eckl, and W. Slack, "Efficient clinical decision making by learning from missing clinical data," in *Computational Intelligence in Healthcare and e-health (CICARE), 2013 IEEE Symposium on*. IEEE, 2013, pp. 27–33.

[8] K. Farooq, A. Hussain, H. Atassi, S. Leslie, C. Eckl, C. MacRae, and W. Slack, "A novel clinical expert system for chest pain risk assessment," in *Advances in Brain Inspired Cognitive Systems*. Springer, 2013, pp. 296–307.

[9] K. Farooq, J. Karasek, H. Atassi, A. Hussain, P. Yang, C. MacRae, M. Mahmud, B. Luo, and W. Slack, "A novel cardiovascular decision support framework for effective clinical risk assessment," in *Computational Intelligence in Healthcare and e-health (CICARE), 2014 IEEE Symposium on*. IEEE, 2014, pp. 117–124.

[10] B. Yuan and J. Herbert, "Context-aware hybrid reasoning framework for pervasive healthcare," *Personal and ubiquitous computing*, vol. 18, no. 4, pp. 865–881, 2014.

[11] M. B. Sesen, R. Banares-Alcántara, J. Fox, T. Kadir, and J. Brady, "Lung cancer assistant: An ontology-driven, online decision support prototype for lung cancer treatment selection," in *OWL: Experiences and Directions Workshop (OWLED)*, 2012.

[12] M. B. Sesen, E. Jiménez-Ruiz, R. Banares-Alcántara, and M. Brady, "Evaluating owl 2 reasoners in the context of clinical decision support in lung cancer treatment selection." in *ORE*, 2013, pp. 121–127.

[13] M. Bouamrane, A. Rector, and M. Hurrell, "Gathering precise patient medical history with an ontology-driven adaptive questionnaire," in *Computer-Based Medical Systems, 2008. CBMS'08. 21st IEEE International Symposium on*. IEEE, 2008, pp. 539–541.

[14] E. Cambria, A. Hussain, T. Durrani, C. Havasi, C. Eckl, and J. Munro, "Sentic computing for patient centered applications," in *Signal Processing (ICSP), 2010 IEEE 10th International Conference on*. IEEE, 2010, pp. 1279–1282.

[15] E. Cambria, I. Hupont, A. Hussain, E. Cerezo, and S. Baldassarri, "Sentic avatar: Multimodal affective conversational agent with common sense," in *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces. Theoretical and Practical Issues*. Springer, 2011, pp. 81–95.

[16] E. Cambria, T. Benson, C. Eckl, and A. Hussain, "¡ i¿ sentic proms¡/i¿: Application of sentic computing to the development of a novel unified framework for measuring health-care quality," *Expert Systems with Applications*, 2012.

[17] X. Zhang, B. Hu, X. Ma, P. Moore, and J. Chen, "Ontology driven decision support for the diagnosis of mild cognitive impairment," *Computer methods and programs in biomedicine*, vol. 113, no. 3, pp. 781–791, 2014.

[18] D. A. Ludwick and J. Doucette, "Adopting electronic medical records in primary care: lessons learned from health information systems implementation experience in seven countries," *International journal of medical informatics*, vol. 78, no. 1, pp. 22–31, 2009.
[19] P. S. Bradley, U. M. Fayyad, and O. L. Mangasarian, "Mathematical programming for data mining: formulations and challenges," *INFORMS Journal on Computing*, vol. 11, no. 3, pp. 217–238, 1999.
[20] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.