جامعة الملك سعود
King Saud University

كلية علوم الحاسب والمعلومات

TOWARD ROBUST HEART FAILURE PREDICTION MODELS USING BIG DATA TECHNIQUES

Heba Rammal and Ahmed Emam, PhD

eTELEMED 2018

Rome, ITALY

AGENDA

- Introduction
- Survey study on HF current work.
- Proposed architecture.
- Proposed methodology.
- Conclusion and future work.

INTRODUCTION

- It has been said that ‘data is the new oil’. It needs to be refined like the oil before it generates value.
- Using **Big Data** analytics, organizations can extract information out of massive, complex, interconnected, and varied datasets (both structured and unstructured) leading to valuable insights. Using tools such as Mahout, WEKA, and R.
- A study by McKinsey points out that U.S. spend at least 600$ - 850$ billion on healthcare.
- One area we can leverage in healthcare using Big Data analytics is Heart Failure (HF).
- HF is the heart's inability to pump a sufficient amount of blood to meet the needs of the body tissues.

But what about Refining ?

IMPORTANCE OF THE STUDY

- Diagnosis is typically done based on doctor's intuition and experience rather than on rich data knowledge hidden in the database which may lead to late diagnosis of the disease, early prediction avoids
  - Unwanted biases, Errors and excessive medical costs, which improve quality of life and services provided to patients.
  - It can identify patients who are at risk ahead of time and therefore manage them with simple interventions before they become chronic patients.

RESEARCH OBJECTIVES

- To develop a survey study of current analytics techniques that support heart failure prediction.
- To mine the useful multi-structure patients data that were collected from a well-known hospital in Saudi Arabia: Kind Saud University Medical City (KSUMC) with the help of cardiologists and radiologist to design a predictive model that will give us the prediction of HF.

RELATED WORK

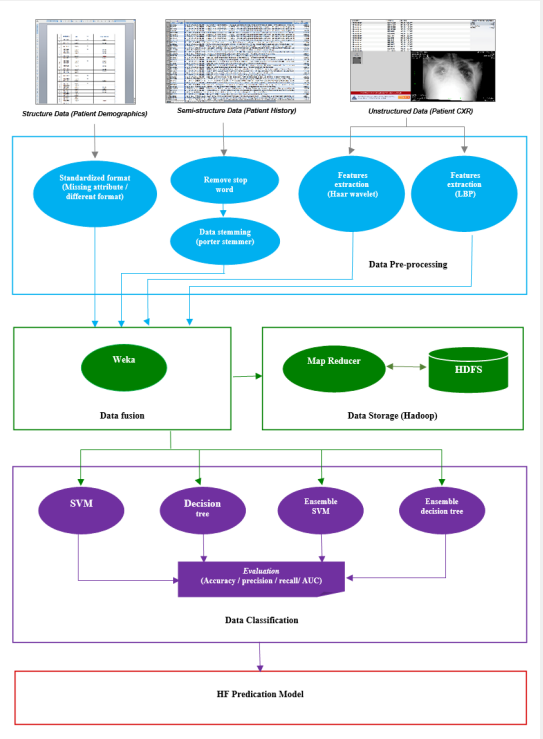
Author		Prediction Technique Used	Platform	Objective
Zolfaghar K, et al (2013)	Supervised learning	Logistic regression, Random forest	Mahout	BD solution to predict the 30- day RoR of HF
Meadam N., et al (2013)		Logistic regression, Naive Bayes, Support Vector Machines	R	Evaluation preprocessing techniques for Prediction of RoR for CHF Patients
Yang, G. et al (2010)		support vector machine (SVM)	n/a	A heart failure diagnosis model based on support vector machine
Panahiazar et al. (2015)		Decision trees, Random Forests, Adaboost, SVM and logistic regression	n/a	Using EHRs and Machine Learning for Heart Failure Survival Analysis
Donzé, Jacques et al (2013)		Cox proportional hazards	SAS	Avoidable 30-Day RoR of HF
K. Zolfaghar et al (2013)		Naive Bayes classifiers	R	Intelligent clinical RoR of HF calculator
Bian, Yuan et al (2015)		Binary logistic regression	n/a	Scoring system for the prevention of acute HF
Suzuki, Shinya et al (2012)		logistic regression	SPSS	Scoring system for evaluating the risk of HF
Auble, T. E. et al (2005)		Decision tree	SPSS	Predict low-risk patients with HF
Pocock, S. J. et al (2005)		Cox proportional hazards	n/a	Predictors of Mortality and Morbidity in patients with CHF
Miao, Fen et al (2014)		Cox proportional hazards	R	Prediction for HF incidence within 1-year
S.Dangare et al (2012)		Decision Trees, Naïve Bayes, and Neural Networks	Weka	HD prediction system using DM classification techniques
Rupali R. Patil (2014)		Naive Bayes classifiers	MATLAB	HD prediction system
Rupali R. Patil (2012)		Artificial Neural network	Weka	A DM approach for predication of HD
Wu, Jionglin et al (2010)		Logistic regression, SVM, and Boosting	SAS, R	HF prediction modeling using EHR
Zebardast, B. et al (2013)		Generalized Regression Neural Networks	MATLAB	Diagnosing HD
Vanisree K. & Singaraju J. (2011)		Multi layered Neural Network	MATLAB	Decision Support System for CHD Diagnosis
Guru N. et al. (2007)		Neural network	MATLAB	HD prediction system
R. Chitra and V, Seenivasagam (2013)		Cascaded Neural Network	n/a	HD Prediction System
Sellappan Palaniappan and Rafiah Awang (2008)		Decision trees, naïve bayed and neural network	.Net	HD prediction system using DM techniques
K. Srinivas et al (2010)		Naive Bayes classifiers	Weka	DM technique for prediction of Heart Attacks
Saqlain, M. et al (2016)		Naive Bayes	n/a	Identification of HF using unstructured data of Cardiac Patients
Strove, Sigurd et al (2004)	Structured predication (Bayesian network)		HUGIN	Decision Support Tools in Systolic HF Management
Gladence, L.M. et al (2014)			Weka	Method for detecting CHF
Liu, Rui et al (2014)			MicrosoftAzure (R & python)	Framework to recommend interventions for 30-Day RoR of HF
C. Ordonez (2006)	Association rules		n/a	HD Prediction
M. Akhil jabbar et al. (2012)	Associative classification	Gini index, Z-statics & genetics algorithm	n/a	Decision Support System for HD prediction
K. Chandra Shekar et al (2012)		association rule mining and classification	Java	Algorithm for prediction of HD

RELATED WORK

- The literature shows a gap in multi-structured predictors for HF prediction and data fusion which will be our main task.
- It is easy to observe that our effort is orthogonal to this related work but, unlike us, none of these works deal with the problem semi-structured or unstructured HF predictor variable.
- They did not generate Big Data analytics predication model, nor do they perform on large scale or distributed data.

PROPOSED ARCHITECTURE

- **Layer 1: Data collection** from KSUMC in the form of structure, unstructured, and semi-structured
- **Layer 2: Data pre-processing** to prepare and filter the data set to make it ready for the next step in building the model
- **Layer 3: Data fusion and storage** which is an important layer that used to integrate all preprocessed data and store it to be then fed to the next step
- **Layer 4: Data classification and evaluation** ware the two final step that includes training, testing then evaluating the model.



METHODOLOGY

- The following, are each step in details toward our proposed model which was implemented using two analytics tools (WEKA and MATLAB)
  - Data collection:
  - Data preprocessing
  - Data fusion and storage
  - Data classification
  - Data evaluation

METHODOLOGY
DATA COLLECTION

- In collaboration with King Saud Medical City (KSUMC) system located in Riyadh, Saudi Arabia all needed clinical and demographic were collected
- The dataset contained 100 real patient records extracted form KSUMC Electronic Health Recode (EHR) and Picture Archiving and Communication System (PACS) with approval from KSUMC administrative office.
- One of the major steps is to determine the subset of attributes (i.e., predictor variables) that has a significant impact in predicting patient with HF from the myriad of attributes present in the dataset.

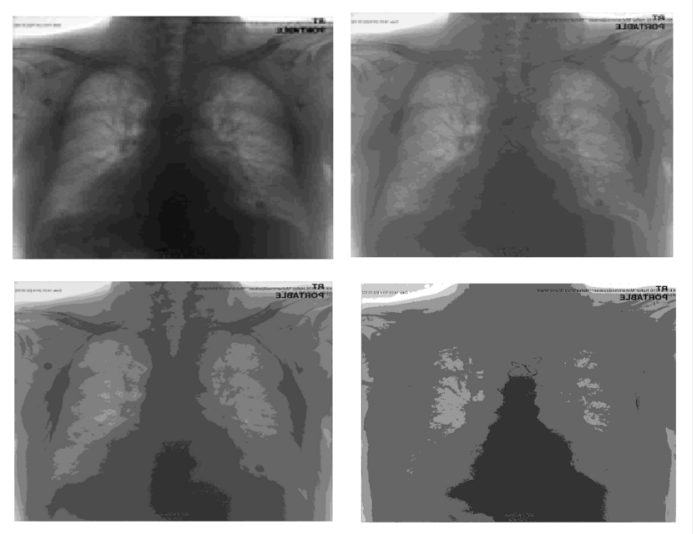
	Label	Feature	Format
Structured	Demographics	Age	Numeral
		Sex	Binary
		Place of birth	Nominal
Semi-Structured	Clinical indications / History	Hypertension, Anemia, Diabetes, Chronic Kidney Disease, Ischemic heart disease, SOB, Swilling hands, Cough, Previous CHF	String
Un-Structured	Front CXR Back CXR Side CXR	64 Features (Haar) 61 Features (LBP)	Numeral

METHODOLOGY
DATA PREPROCESSING

- *Structured Data* (Patient Demographics): We had to do screening before data analyzing; Those data with too many missing attributes were all wiped off. Also, all data formats were standardized.
- *Semi-structured Data* (history / clinical indication): Text analysis techniques were applied:
  - **Stop word removal**: helped in removing all common words, such as ‘a’ and ‘the’ from the text.
  - **Stemming**: porter algorithm was used as the to identify and remove the commoner morphological and inflexional endings from words

METHODOLOGY
DATA PREPROCESSING

- *Un-structured Data* (Chest X-Ray (CXR)):
  - Using MATLAB, Haar wavelet and local binary pattern (LBP) feature extraction functions were applied to over 150 CXR images.
    - Haar was used since it is the fastest technique that can be used to calculate the feature vector. 64 features were found using Wavelets features.
    - On the other hand, LBP has been found to be a powerful and simple feature yet very efficient texture operator. 61 features were found using LBP.
  - Principle component analysis (PCA) was applied to properly rank and compute the weights of the features to find the most promising attributes to predicate HF from 64 / 61 features found.

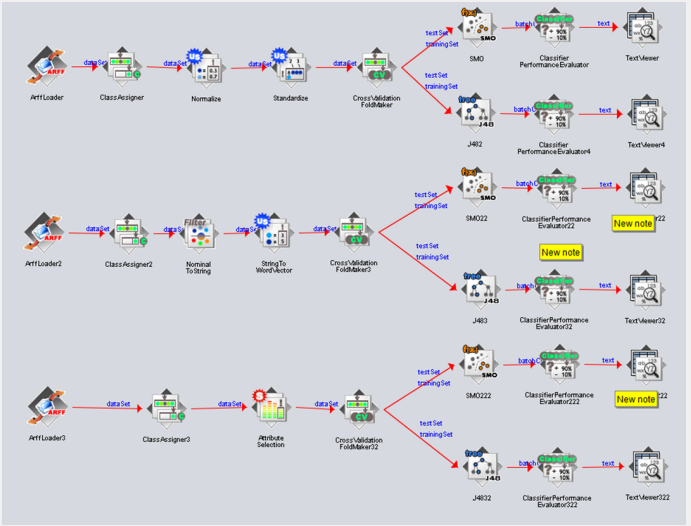


METHODOLOGY
DATA FUSION AND STORAGE

- After pre-processing the data and extracting all the needed attributes, the statistics feature from CXR scan images with other attributes will be integrated and stored using Big Data tools.
- We leverage the power of Hadoop; Hadoop is not a database, so it lacks functionality that is necessary for many analytics scenarios.
- Fortunately, there are many options available for extending Hadoop to support complex analytics, including real-time predictive models such as Weka (Waikato Environment for Knowledge Analysis), which we used in our study.
- We added distributedWekaHadoop to Weka: which works as a Hadoop wrapper for Weka.

METHODOLOGY
DATA CLASSIFICATION AND EVALUATION

- In this study, each set of the data (Structured, Semi-structured, and Unstructured) trained and tested using data mining algorithms in Weka.
- Knowledgeflow was used in Weka which presents, a workflow inspired interface.
- Data was trained using two state-of-the-art classification algorithms including, SVM and Decision Tree.
- In the end, accuracy, precision, recall and, Area under the Curve (AUC) were used as performance measures.



CONCLUSION AND FUTURE WORK

- Non-Communicable Diseases like Heart Failure is one of a major health hazard in the KSA. The literature shows a gap in multi-structured predictors for HF prediction and data fusion which is our main task
- In this research, data fusion played a vital role in combining multi-structure dataset. The goal of this research deals with the study of HF predication in healthcare industry using Big Data analytics technique.
- As future work, we will use a larger dataset for training.
- We will also incorporate more medical data into the model, better simulating how a cardiologist makes a decision.
- Finally, we will use different data mining techniques to extract the buried information from the patient semi-unstructured /unstructured reports.

REFERENCES

- McKinsey and Company, McKinsey Global Institute, Big Data: The next frontier for innovation, competition and productivity. Available at http://lateralpraxis.com/download/The_big_data_revolution_in_healthcare.pdf , Last accessed March 2017.
- National Heart, Lung, and Blood Institute, What is heart failure. Available at http://www.nhlbi.nih.gov/health/health-topics/topics/hf, Last accessed April 2017.