Toward Robust Heart Failure Prediction Models Using Big Data Techniques

Heba F. Rammal
Information Technology Department
King Saud University
Riyadh, Saudi Arabia
Email: hrammal@ksu.edu.sa

Ahmed Z. Emam
Information Technology Department
King Saud University, Riyadh, Saudi Arabia
Computer Science and Math Department,
Menoufia University, Egypt
Email: aemam@ksu.edu.sa

Abstract— **Big Data technologies have a great potential in transforming healthcare, as they have revolutionized other industries. In addition to reducing the cost, they could save millions of lives and improve patient outcomes. Heart Failure (HF) is the leading death cause disease globally. The social and individual burden of this disease can be reduced by its early detection. However, the signs and symptoms of HF in the early stages are not clear, so it is relatively difficult to prevent or predict it. The main objective of this paper is to propose a model to predict patients with HF using a multi-structure dataset integrated from various resources. The underpinning of our proposed model relies on studying the current analytical techniques that support heart failure prediction, and then build a model based on Big Data technologies. To achieve this, we extracted different important factors of heart failure from King Saud Medical City (KSUMC) system, Saudi Arabia, which are available in structured, semi-structured and unstructured format. Unfortunately, a lot of information is buried in unstructured data format. We applied some pre-processing techniques to enhance the parameters and integrate different data sources in Hadoop Distributed File System (HDFS). Then, we applied data-mining algorithms to discover patterns in the dataset to predict heart risks and causes. Finally, the analyzed report is stored and distributed to get the insight needed from the prediction.**

Keywords- Big Data; Hadoop; Healthcare; Heart Failure; Prediction Model.

I. INTRODUCTION

In recent years, a new hype has been introduced into the information technology field called ‘Big Data’. Big Data offers an effective opportunity to manage and process massive amounts of data. A report by the International Data Corporation (IDC) [1] found that the volume of data the whole humanity produced in 2010 was around 1.2 Zettabytes, which can be illustrated physically by having 629.14 Million 2 Terabytes external hard drives that can fill more than 292 great pyramids. It has been said that ‘data is the new oil’, so 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. Analytics can be done on big data using a new class of technologies that includes Hadoop [2], R [3], and Weka [4]. These technologies form the core of an open source software framework that supports the processing of huge data sets. Like any other industry, healthcare has a huge demand to extract a value from data. A study by McKinsey [5] points out that U.S. spend at least 600$ - 850$ billion on healthcare. The report points to the healthcare sector as potential field where valuable insights are buried in structured, unstructured, or highly varied data sources that can now be leveraged through Big Data analytics. More specifically, the report predicts that if U.S. healthcare could use big data effectively, the hidden value from data in the sector could reach more than 300$ billion every year. Also, according to the ‘Big Data cure’ published last March by MeriTalk [6], 59% of federal executives working in healthcare agencies indicated that their core mission would depend on Big Data within 5 years.

One area we can leverage in healthcare using Big Data analytics is Heart Failure (HF); HF is the leading cause of death globally. It is the heart's inability to pump a sufficient amount of blood to meet the needs of the body tissues [7]. Despite major improvements in the treatment of most cardiac disorders, HF remains the number one cause of death in the world and the most critical challenges facing the healthcare system today [8]. A 2015 update from the American Heart Association (AHA) [9] estimated that 17.3 million people die due to HF per year, with a significant rise in the number to reach 23.6 million by 2030. They also reported that the annual healthcare spending would reach $320 billion, most of which is attributable to hospital care. According to World Health Organization (WHO) statistics [10], 42% of death in 2010 (42,000 deaths per 100,000) in the Kingdom of Saudi Arabia (KSA) were due to cardiovascular disease. Also, in KSA, cardiovascular diseases represent the third most common cause of hospital-based mortality, second to accidents and senility.

HF is a very heterogeneous and complex disease which is difficult to detect due to the variety of unusual signs and symptoms [11]. Some examples of HF risk factors are: breathing, dyspnea, fatigue, sleep difficulty, loss of appetite, coughing with phlegm or mucus foam, memory losses, hypertension, diabetes, hyperlipidemia, anemia, medication, smoking history and family history. Heart failure 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. Thus, the effort to utilize clinical data of patients collected in databases to facilitate the early diagnosis of HF patients is considered a challenging and valuable contribution to the healthcare sector. 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. Clinical data are available in the form of complex reports, patient's medical history, and electronics test results [12]. These medical reports are in the form of structured, semi-structured and unstructured data. There is no problem to use structured data for risk prediction model. But, there is a lot of valuable information buried in unstructured data format because this data is very discrete, complex, multidimensional and noisy [13]. In our study, we collected patient's reports from a well-known hospital in Saudi Arabia: Kind Saud University Medical City (KSUMC).

The objective of our research is to mine the useful information from these reports with the help of cardiologists and radiologist to design a predictive model that will give us the prediction of HF. The paper is organized as follows. Section II introduces the related work. Section III describes the proposed architectural model and each process involved. In Section IV, the proposed research methodology is explained. The conclusion and future work of this research are found in Section V.

II. Literature Review

Big Data predictive analytics represents a new approach to healthcare, so it does not yet have a large or significant footprint locally or internationally. To the best of our knowledge, no prior work has investigated the benefits of Big Data analytics techniques in heart failure prediction problem. A work by Zolfaghar et al. [14] proposed a real-time Big Data solution to predict the 30- day Risk of Readmission (RoR) for Congestive Heart Failure (CHF) incidents. The solution they proposed included both extraction and predictive modeling. Starting with the data extraction, they aggregate all needed clinical & social factors from different resources and then integrated it back using a simple clustering technique based on some common features of the dataset. The predictive model for the RoR is formulated as a supervised learning problem, especially binary classification. They used the power of Mahout as machine learning based Big Data solution for the data analytics. To prove quality and scalability of the obtained solutions they conduct a comprehensive set of experiments and compare the resulted performance against baseline non-distributed, non-parallel, non-integrated dataset results previously published. RoR for CHF gained the interest of researchers due to their negative impacts on healthcare systems' budgets and patient loads. Thus, the development of predictive modeling solutions for risk prediction is extremely challenging. Prediction of RoR was addressed by, Vedomske et al. [15], Shahet al. [16], Royet al. [17], Koulaouzidis et al. [18], Tugermanet al. [19], and Kang et al. [20]. Although our studied problem is fundamentally different, as they are all using structure data; nevertheless, our proposed model could benefit from the proposed large-scale data analysis solutions.

Panahiazar et al. [21] applied decision trees, Random Forests, Adaboost, SVM and logistic regression to a dataset extracted from the EHR of the Mayo Clinic. The dataset included 5044 HF patients admitted to the Mayo Clinic from 1993 to 2013. For each patient, 43 predictor variables, expressing demographic data, vital measurements, lab results, medication, and co-morbidities, were recorded. The class variable corresponded to mortality status, consequently, three versions of the dataset were created, each one corresponding to survival period (1-year, 2-year, 5-year). 1560 instances out of 5044 were used for training and the rest 3484 instances for testing. The authors observed that logistic regression and Random Forests were more accurate models compared to others, also, among the scenarios, the best prediction accuracy was 87.12%.

Saqlain et al. [22] worked on 500 HF patients from the Armed Forces Institute of Cardiology (AFIC), Pakistan, in the form of medical reports. They started by manually applying pre-processing steps to transform unstructured reports into the structured format to extract data features. Then they perform multinomial Naïve Bayes (NB) classification algorithm to build 1-year or more survival prediction model for HF diagnosed patients. The proposed model achieved an accuracy and Area under the Curve (AUC) of 86.7% and 92.4%, respectively. Even though the above model is based on some attributes extracted from the unstructured data, they used a manual approach to achieve this. On the other hand, our model deals with unstructured data by automatically recognizing attributes using Machine Learning (ML) approaches without the need for a radiologist opinion. A scoring model for HF diagnosis based on SVM were proposed by Yang, G. et al. [23]. They applied it to a total of 289 samples clinical data collected from Zhejiang Hospital. The sample was classified into three groups: healthy group, HF-prone group, and HF group. They compared their results to previous studies which showed a considerable improvement for HF diagnosis with a total accuracy of 74.44%. Especially in HF-prone group, accuracy reaches 87.5%, and this implies that the proposed model is feasible for early diagnosis of HF. However, accuracy in the HF group is not satisfactory due to the absence of symptoms and signs and also due to the high prevalence of conditions that may mimic the symptoms and signs of heart failure.

More studies are listed in Table 1, which were collected and summarized as recent analytics techniques and platform to predict heart failure. The table shows that supervised learning technique is the most domainant techniques in building HF predication model, also Weka and Matlab are the preferable platforms to build HF prediction model.

The literature presented above 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.

TABLE 1. STATE OF ART FOR HF PREDICTION STUDIES

Author	Prediction Technique Used		Platform	Objective
Zolfaghar et al. (2013)	**Supervised learning**	Logistic regression, Random forest	Mahout	BD solution to predict the 30- day RoR of HF
Meadam et al. (2013)		Logistic regression, Naive Bayes, Support Vector Machines	R	Evaluation preprocessing techniques for Prediction of RoR for CHF Patients
Yang 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 et al. (2013)		Cox proportional hazards	SAS	Avoidable 30-Day RoR of HF
Zolfaghar et al. (2013)		Naive Bayes classifiers	R	Intelligent clinical RoR of HF calculator
Bian et al. (2015)		Binary logistic regression	n/a	Scoring system for the prevention of acute HF
Suzuki et al. (2012)		logistic regression	SPSS	Scoring system for evaluating the risk of HF
Auble et al. (2005)		Decision tree	SPSS	Predict low-risk patients with HF
Pocock 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
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 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 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 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

Copyright (c) IARIA, 2018. ISBN: 978-1-61208-618-7

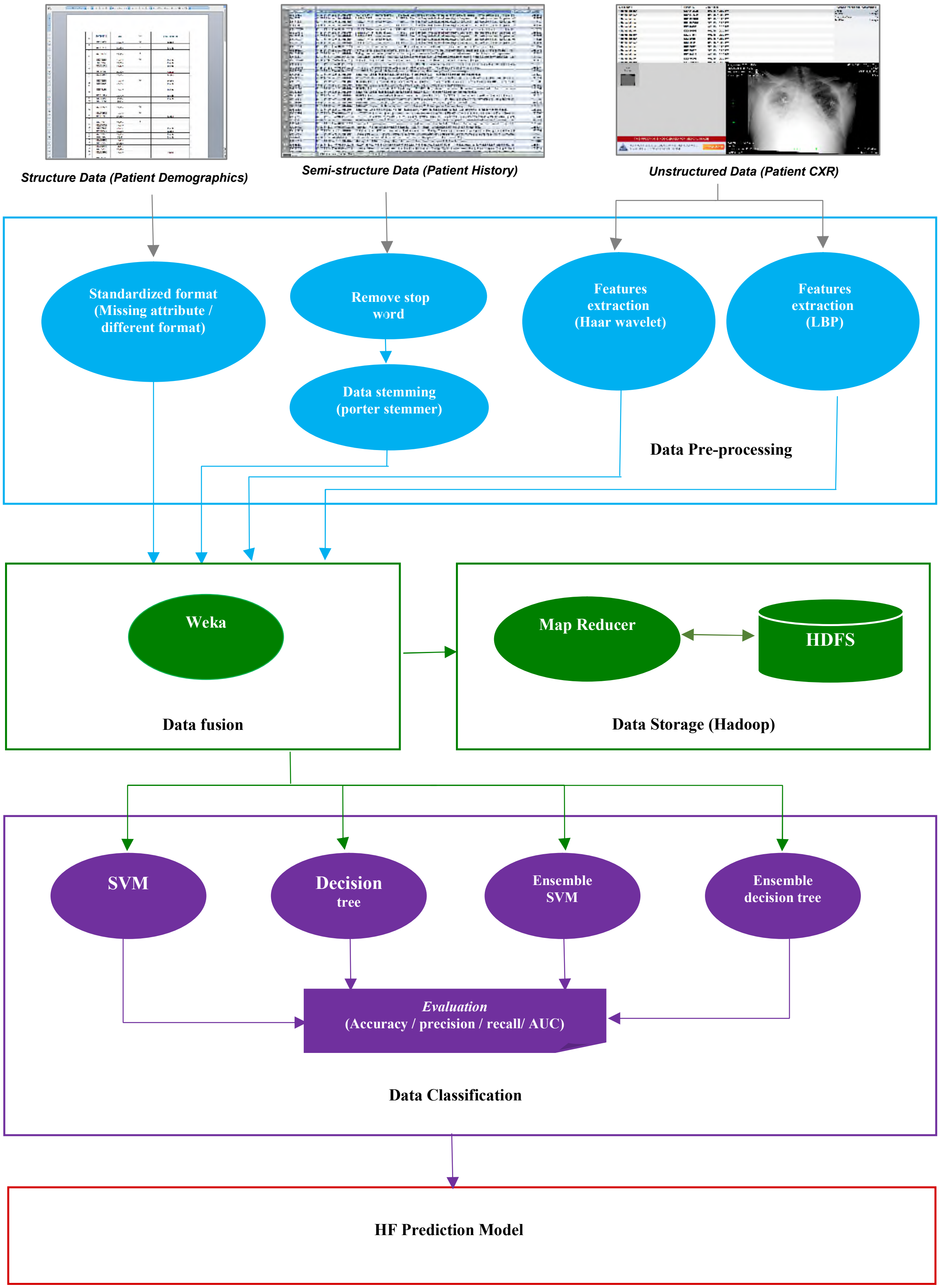
Structure Data (Patient Demographics)
Semi-structure Data (Patient History)
Unstructured Data (Patient CXR)
Standardized format (Missing attribute / different format)
Remove stop word
Data stemming (porter stemmer)
Features extraction (Haar wavelet)
Features extraction (LBP)
Data Pre-processing
Weka
Data fusion
Map Reducer
HDFS
Data Storage (Hadoop)
SVM
Decision tree
Ensemble SVM
Ensemble decision tree
Evaluation
(Accuracy / precision / recall/ AUC)
Data Classification
HF Prediction Model

Figure 1. HF Prediction Model.

III. Proposed Architecture

Predictive analysis can help healthcare providers accurately expect and respond to the patient needs. It provides the ability to make financial and clinical decisions based on predictions made by the system. Building the predictive analysis model includes various phases as mentioned in the literature. (Figure 1 shows the complete architecture of the proposed model).

IV. Proposed Methodology

In the following, we will describe the adaptive methodology and each step towards our proposed model.

A. Data Collection

In collaboration with King Saud Medical City (KSUMC) system located in Riyadh, Saudi Arabia all needed clinical and demographic data were adopted to evaluate the performance of the proposed hybrid method in identifying HF risk in patients. The dataset contained 100 real patient records extracted form KSUMC electronic health recod (EHR) with approval from KSUMC administrative office. The selected sample covers most distribution domain from different ages and different sociodemographic. At the same time, validation of the selected dataset was achieved by consolidating some cardiologist and data scientist. The selected dataset has many noises such as missing values and misidentified attributes. The output values were categorized into two labels denoted as Non-HF (meaning HF is absent) and HF (meaning HF is present). One of the major steps is the distillation of data, which responsible of determining 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. In our study, the definitions and categories of the HF attributes are summarized in Table 2.

B. Data Preprocessing

In this phase structured, semi-structured, and unstructured data are accumulated, cleansed, prepared, and made ready for further processing.

TABLE 2. SELECTED ATTRIBUTED FROM 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)	Numeral
		61 Features (LBP)	

- Raw structured information has some missing values and written in different formats during information entry or management, we had to do screening before data analyzing. Those data with too many missing attributes were all wiped off when we selected the sample set. Also, all data formats were standardized, see Table 3.
- Apply text analysis techniques on the semi-structured dataset to get the needed information. Two steps were applied to the text to process the data, stop word removal and stemming. Stop word removal - illustrated in Figure 2 - helped in removing all common words, such as ‘a’ and ‘the’ from the text. Next, Porter algorithm was used as the stemmer to identify and remove the commoner morphological and inflexional endings from words [24].
- Extracting all needed features from the unstructured dataset, which includes 3 types of Chest X-Ray (CXR) images (front CXR, back CXR, and side CXR) using MATLAB. Haar wavelet and local binary pattern (LBP) 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 [25]. This was performed based on applying the Haar wavelet four times to divide the input image into 16 sub-images, illustrated in Figure 3. 64 features that include Energy, Entropy, Homogenous as 3D XYZ were found using Wavelets features, see Figure 4.

  On the other hand, LBP has been found to be a powerful and simple feature yet very efficient texture operator which labels the pixels of an image by calculating each pixels’ neighborhoods’ thresholding then considers the result as a binary number. 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. The selected attributes were used to train the classifiers to get a better accuracy.

TABLE 3. UNSTANDARIZED STRUCTRED DATA

	Age	Sex	P_B	Diagnosis
1	045Y	Female	Riyadh	HF
2	62	F	?	HF
…..	….	……	…..	…..
100	098	male	riy	Non-HF

Explanatory Data

Label

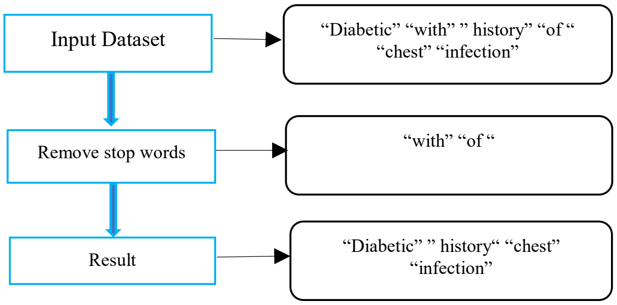

Figure 2. Stop word removal.

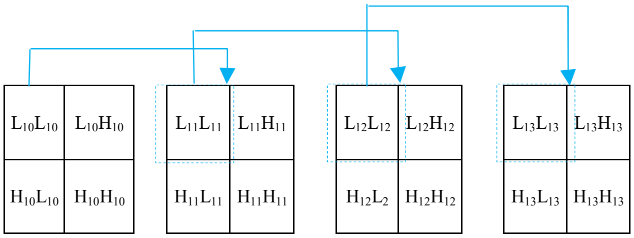

Figure 3. Applying haar wavelet four times.

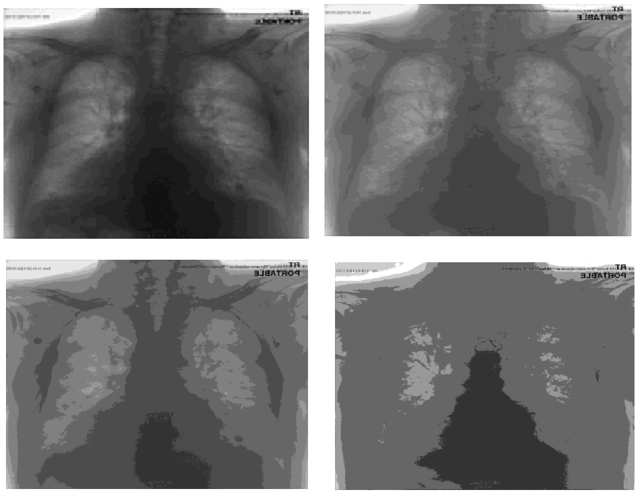

Figure 4. Result from wavelet.

C. *Data Storage and Fusion*

After pre-processing the data and extracting all the needed attributes, the statistics feature from CXR scan images with other attributes will be integrated using Big Data tools to generate the needed data that will be used for training and testing & finally produce the predictive model. We leverage the power of Hadoop as a framework for distributed data processing and storage. 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.

D. *Data Classification*

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, see Figure 5. 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. We get all these measures by using confusion matrix because it contains all True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) assessments.

V. Conclusion And Future Work

Big Data analytics provides a systematic way for achieving better outcomes like availability and affordability of healthcare service to all members of the population. Non-Communicable Diseases like Heart Failure is one of the major health hazard in the KSA. By transforming various health records of HF patients into useful analyzed result, this analysis will make the patient understand the complications that can occur.

The literature shows a gap in multi-structured predictors for HF prediction and data fusion, which is 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.

Combining structured dataset (sociodemographic), semi-structured dataset (cardiology diagnoses), and unstructured dataset (Patient X-ray images) is a very hard task. In this research, data fusion played a vital role in combining multi-structured datasets. The goal of this research deals with the study of HF predication in healthcare industry using Big Data analytics technique. The design of predictive analysis model of HF may give enhanced data and analytics for better results in healthcare.

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

[1] J. Gantz and D. Reinsel, "The digital universe decade – are you ready?" External publication of IDC (Analyse the Future) information and data, pp. 1- 16, 2010.

[2] Hadoop Apache. Available at http://hadoop.apache.org/, Last accessed March 2017.

[3] R Project. Available at https://www.r-project.org/about.html, Last accessed March 2017.

[4] P. Navas, Y. Parra, and J. Molano, “Big Data Tools: Haddop, MongoDB and Weka”. International Conference on Data Mining and Big Data, pp 449-456, 2017.

[5] 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.

[6] Meritalk, The Big Data cure. Available at http://www.meritalk.com/bigdatacure, Last accessed March 2017.

[7] 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.

[8] World Health Organization (WHO) (2015). Cardiovascular diseases (CVDs). Available at http://www.who.int/mediacentre/factsheets/fs317/en/ , Last accessed March 2017.

[9] American Heart Association. Heart Disease and Stroke Statistics – At-a-Glance. Available at http://www.heart.org/idc/groups/ahamah-public/@wcm/@sop/@smd/documents/downloadable/ucm_470704.pdf , Last accessed March 2017.

[10] Mistry Of Health (MOH), “Cardiovascular Diseases Cause 42% of Non-Communicable Diseases Deaths in the Kingdom”. Available at https://www.moh.gov.sa/en/Ministry/MediaCenter/News/Pages/News-2013-10-30-002.aspx, Last accessed March 2018.

[11] Ishwarappa and J. Anuradha, “A Brief Introduction On Big Data 5Vs Characteristics And Hadoop Technology”. Procedia Computer Science 48, pp. 319-324, 2015.

[12] J. A. Hartigan and M. A. Wong, “Algorithm AS 136: A k-means clustering algorithm,”, Series C (Applied Statistics), vol. 28, pp. 100-108, 1979.

[13] E. AbuKhousa and P. Campbell, “Predictive data mining to support clinical decisions: An overview of heart disease prediction systems,” Proc. IEEE, Innovations Information Technology (IIT), pp. 267-272, March 2012.

[14] K. Zolfaghar et al., “Big Data Solutions for Predicting Risk-of-Readmission for Congestive Heart Failure Patients”, IEEE Inter. Conf. on Big Data, 2013.

[15] M. A. Vedomske, D. E. Brown and J. H. Harrison, “Random forests on ubiquitous data for heart failure 30-day readmissions prediction”, Proceedings of the 12th international conference on machine learning and applications, vol. 2, pp. 415-421, 2013.

[16] S. J. Shah et al. “Phenomapping for novel classification of heart failure with preserved ejection fraction”. Circulation. Vol 131, pp. 269–279, 2015.

[17] S. B. Roy, A. Teredesai, Zolfaghar K., Liu R., and Hazel D., “Dynamic hierarchical classification for patient risk-of-readmission”. Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1691–1700, 2015.

[18] G. Koulaouzidis, D. K. Iakovidis, and A. L. Clark, “Telemonitoring predicts in advance heart failure admissions”. Int J Cardiol, vol. 216, pp. 78–84, 2016.

[19] L. Turgeman and J. H. May, “A mixed-ensemble model for hospital readmission.”, Artif Intell Med, vol. 72, pp. 72–82, 2016.

[20] Y. Kang, M.D. McHugh, J. Chittams and K.H. Bowles, “Utilizing home healthcare electronic health records for telehomecare patients with heart failure. A decision tree approach to detect associations with rehospitalizations”. Comput Inform Nurs, vol. 34 no. 4, pp.175–182, 2016.

[21] M. Panahiazar, V. Taslimitehrani, N. Pereira and J. Pathak, “Using EHRs and machine learning for heart failure survival analysis.” Stud Health Technol Inform, vol. 216, pp. 40–44, 2015.

[22] M. Saqlain, W. Hussain, N. Saqib and A. Muazzam Khan, “Identification of Heart Failure by Using Unstructured Data of Cardiac Patients.”, 45th International Conference on Parallel Processing Workshops, 2016.

[23] G. Yang et al., “A heart failure diagnosis model based on support vector machine”. 3rd International Conference on Biomedical Engineering and Informatics. 2015;

[24] C. Moral, A. Antonio, R. Imbert and J. Ramírez, “A survey of stemming algorithms in information retrieval.” Information research, vol. 19 no. 1, March 2014.

[25] S. Arora, Y. Brar and S. Kumar, “HAAR wavelet transform for solution of image retrieval.” International Journal of Advanced Computer and Mathematical Sciences ISSN, vol. 5, no. 2, pp 27-3, 2014

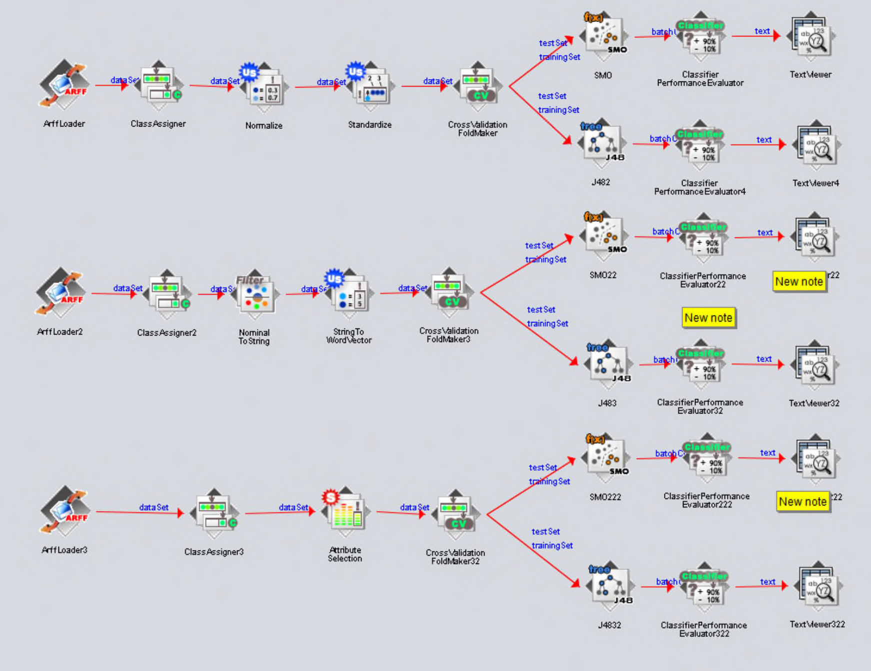

Figure 5. The Proposed Knowledgeflow using Weka.