# Heart Failure Prediction Models using Big Data Techniques

Heba F. Rammal
Information Technology Department
King Saud University
Riyadh, Saudi Arabia

Ahmed Z. Emam
Information Technology Department
King Saud University, Riyadh, Saudi Arabia
Computer Science and Math Department,
Menoufia University, Egypt

***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, both nationally and internally. 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 research 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 an integrated model based on Big Data technologies using WEKA analytics tool. 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) using distributed-WEKA-spark package. 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. Our proposed model achieved an accuracy and Area under the Curve (AUC) of 93.75% and 94.3%, respectively.**

***Keywords*—Big data; hadoop; healthcare; heart failure; prediction model**

## I. INTRODUCTION

In the 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 datasets. Like any other industry, healthcare has a huge demand to extract a value from data. A study by McKinsey [5] points out that the U.S. spends at least 600$ - 850$ billion on healthcare. The report points to the healthcare sector as a 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 accident 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 the prediction model. But, there is a lot of valuable information buried in the semi-structured and unstructured data format because those data are very discrete, complex, and noisy [13]. In our study, we collected patient's reports from a well-known hospital in Saudi Arabia: King 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 K, 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 recourse 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. Due to their negative impacts on healthcare systems' budgets and patient loads, RoR for CHF gained the interest of researchers. Thus, the development of predictive modeling solutions for risk prediction is extremely challenging. Prediction of RoR was addressed by, Vedomske et al. [15], Shah et al. [16], Royet al. [17], Koulaouzidis et al. [18], Tugerman et 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] used a dataset of 5044 HF patients admitted to the Mayo Clinic from 1993 to 2013. They applied 5 training algorithms to the data that includes decision trees, Random Forests, Adaboost, SVM and logistic regression. 43 predictors were selected which express demographic data, vital measurements, lab results, medication, and co-morbidities. The class variable corresponded to survival period (1-year, 2-year, 5-year). 30% of the dataset were used for training and the rest 70% 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, M. 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 was 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 in 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 so satisfied 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 were listed in Table I, which was collected and summarized as recent analytics techniques and platform to predict heart failure. The table shows that supervised learning technique is the most dominant techniques in building HF prediction 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 prediction model, nor do they perform on large scale or distributed data.

TABLE I. STATE OF ART FOR HF PREDICTION STUDIES

| 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 prediction 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 prediction (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 |

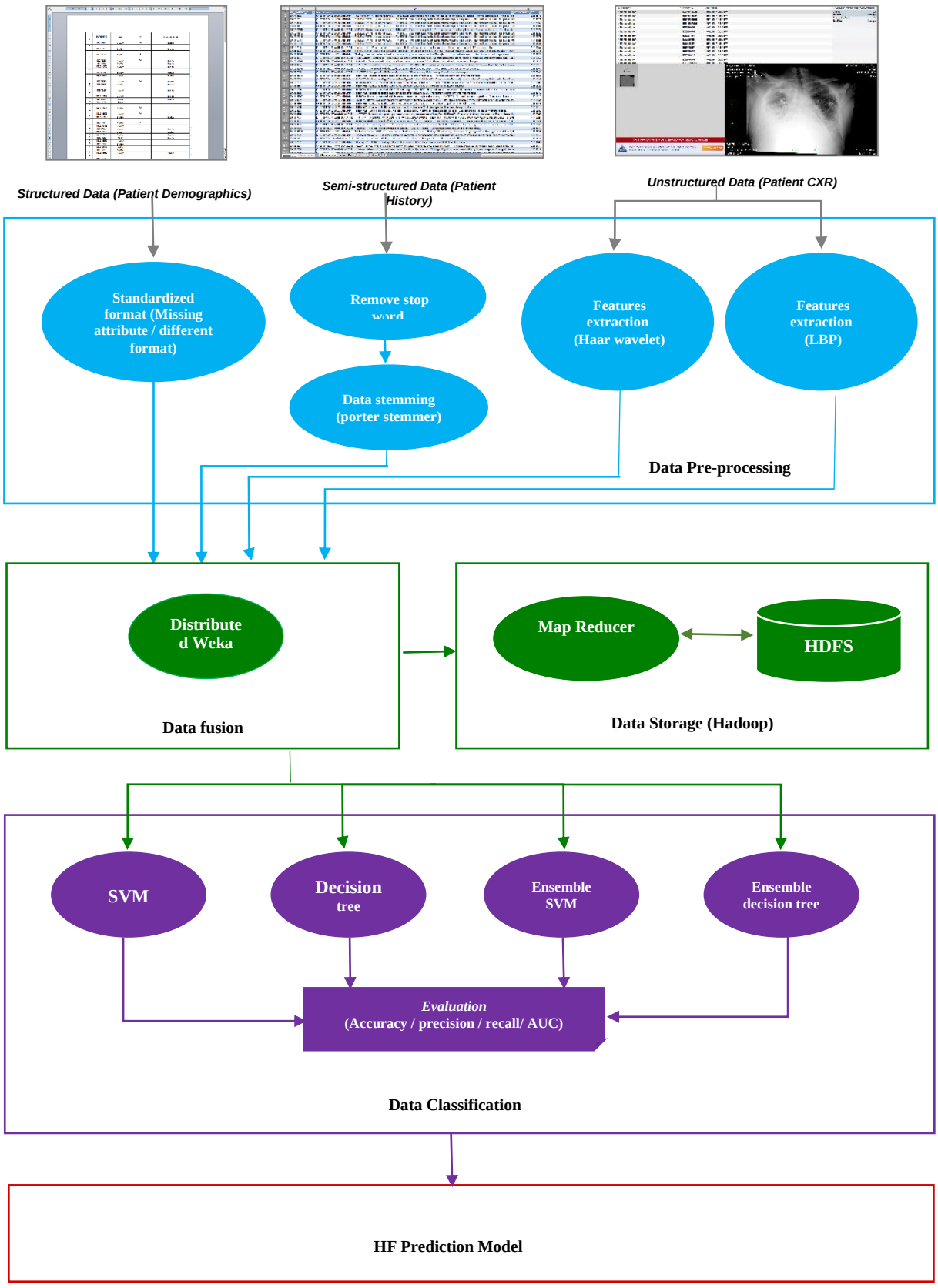


Fig. 1. HF prediction model.

## 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 (Fig. 1 shows the complete architecture of proposed model).

- 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 dataset 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 in HDFS to be then fed to the next step.
- Layer 4: Data classification and evaluation are the two final steps that include training, testing then evaluating the model.

## IV. Proposed Methodology

In the following, we will describe the adapt methodology and each step in toward our proposed model.

### A. Data Collection

In our study, we collaborated with King Saud University Medical City (KSUMC) system located in Riyadh, Saudi Arabia to extract manually all needed clinical and demographic that we needed to adapt to evaluate the performance of the proposed model in identifying HF risk, from January 2015 to December 2015.

The dataset contained 100 real patient records extracted form KSUMC Electronic Health Record (EHR) and Picture Archiving Communication System (PACS), with approval from KSUMC administrative office. Due to patients' privacy, some demographic information that includes name, national ID number or iqama number, phone, address were excluded, Basic characteristics of the samples' demographic information are shown in Table II. Obviously, our sample doesn't have a uniform distribution in terms of gender. Also, patients aged from 60 years old to 70 years old account for the most part of our data. 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 this study, parameters are selected from 3 datasets which are summarized in Table III.

The validation of the selected dataset achieved by consolidating some cardiologist and according to their evaluation all cases were labeled into two groups. 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). Our dataset contains 69 predictor variable, having 1 binary variables (gender) 3 text values (place of birth, history, and symptoms) and 65 numerical variables (including age and all CXR features) and a single response variable 'Result' having only two values HF and Non-HF.

### B. Data Preprocessing

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

TABLE II. Demographic basic Characteristics

| Characteristic | Group | | |
|---|---|---|---|
| | **HF group** | **Non-HF group** | **Total** |
| **Female** | 21 | 23 | 44 |
| **Male** | 25 | 31 | 56 |
| **Age (mean ± SD)** | 69 ± 12 | 61 ± 15 | 65 ± 14 |

TABLE III. Selected Attributes 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. Those data with too many missing attributes were all wiped off when we selected the dataset. Also, all data formats were standardized, see Table IV.
- Apply text analysis techniques on the semi-structured dataset to get the needed information. Three steps were applied to the text to process the data, tokenizer, stop word removal, and stemming. Before any real text processing is to be done, the text needs to be segmented into words, punctuation, phrases, symbols, and other meaningful elements called tokens. Next, stop word removal, illustrated in Fig. 2, helped in removing all common words, such as 'a' and 'the' from the text. Then, Porter algorithm was used as the stemmer to identify and remove the commoner morphological and inflexional endings from words, which is part of the snowball stemmers in WEKA [24].

TABLE IV. Unstandardized Structured Data

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

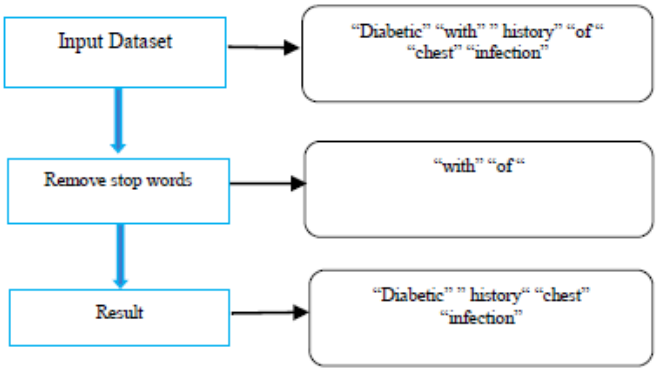


Fig. 2. Stop word removal.

- 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 4 times to divide the input image into 16 sub-images, illustrated in Fig. 3. 64 features that include Energy, Entropy, Homogenous were found, Fig. 4 illustrate the resulted images after first level of Haar. Each CXR image represents certain features in the image of heart values of Energy_Entropy_Homo and wavelet features. A total of 16 for Energy_Entropy_Homo in each level since we have 4 levels in Haar wavelet so 64 features are extracted in total. 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. We applied LBP to all CXR images by first, labeling all the pixels, absent the borders, using the LBP operator, then dividing the image into 60 segments. A feature vector is created by obtaining the histogram of each region, and finally concatenating all the histograms into one vector which result in finding 60 features. A typical LBP application to a CXR is shown in Fig. 5.

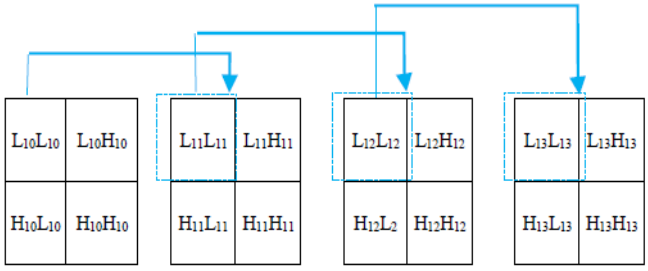


Fig. 3. Applying haar wavelet four times.

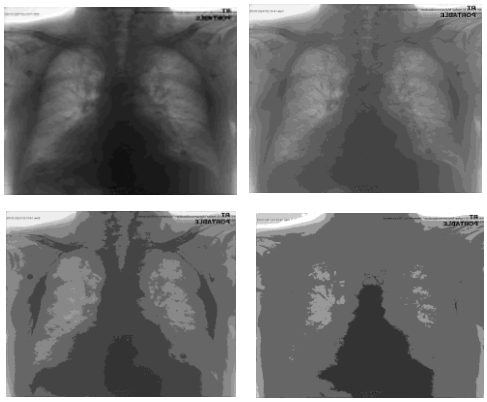

Fig. 4. The result from wavelet.

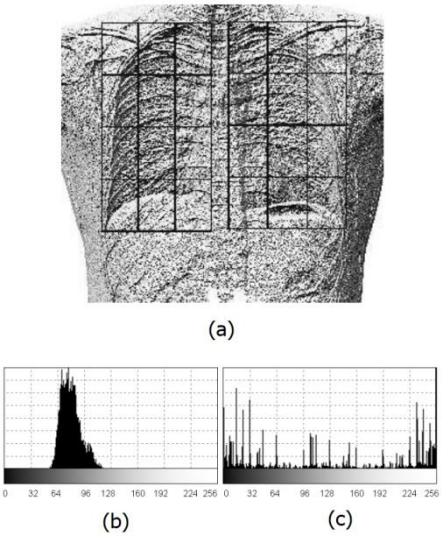


Fig. 5. Histogram of image obtained by applied LBP algorithm on CXR: (a) LBP applied image and (c) histogram before LBP (b) histogram of a.

- Principle component analysis (PCA) was applied to properly rank and compute the weights of the features to find the most promising attributes to predict HF from the features found. The selected attributes were used to train the classifiers to get a better accuracy. Also, to circumvent the imbalanced problem, we applied resampling method. This method alters the class distribution of the training data so that both the classes are well represented. It works by resampling the rare class records so that the resulting training set has an equal number of records for each.

### 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 data fusion techniques to generate the needed data that will be used for training and testing and finally produce the predictive model. Complementary data fusion classification technique was used as each dataset represents part of the scene and was used to build a reliable information. 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 distributed WekaSpark 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. Knowledge flow was used in Weka which presents, a workflow inspired interface, see Fig. 6. Data was trained using two state-of-the-art classification algorithms including, Random Forest (RF) and Logistic Regression (LR) as they both have been known to result in high accuracy in binary class prediction.

The ability to handle and analyze various types of data (structured, semi-structured or unstructured) is one of the most important characteristics of Big Data analytic techniques. We will perform the classifications in two phases to show that using the proposed integrated learning analytics technique is more efficient than a traditional single predictive model, especially if the data is multi-structured and has unique characteristics. In the end, model quality was assessed through common model quality measures such as accuracy, precision, recall and, Area under the Curve (AUC). Depending on the final goal of the HF prediction, the different evaluation measures are less or more appropriate. Recall is relevant as the detection of patients that belong to HF class is the main goal. The precision is considered less important as cost related to falsely predicting patients to belong to the class HF is low. The accuracy is the traditional evaluation measure that gives a global insight in the performance of the model. The AUC measure is typically interesting in our study because the problem is imbalanced. It is observed that the number of instances with HF label significantly outnumbers the number of instances with class label Non-HF.

## V. Results and Insights

It is clear from Tables V and VI that integrated dataset has the highest accuracy and AUC: ~92% and ~90%, respectively, then using each dataset by its own for HF patient's prediction. Also, using LBP features extraction methods achieved better performance results then Haar with 93% compared to 91% for Haar. We can also note that logistic regression did great in the integrated models compared to its poor performance in the single dataset models with over 90% recall, which can be resulted from the nature of the algorithm as it predicts better for problems with many attributes. Based on the experiment, we can provide evidence of the importance of the integration of unstructured, semi-structured, and structured data. This indicates that there are some indicators within textual patient report and images that can be extracted and used as important predictors of Heart failure. Also, the discovery of feature selection as a suite of methods that can increase model accuracy, decrease model training time and reduce overfitting.

Our proposed approach is also very important because it provides a knowledge discovery and intelligent model to the cardiologists and researchers such as: (1) the dataset contains 56% male patients, which mean male patients have more probability to get an HF diagnose than females. (2) 73% patient's age over 65 which indicate that aged people have more chance to get HF. (3) 70% of patients coming for HF complain having hypertension, diabetes, and SOB which means this disease has the main impact of HF.

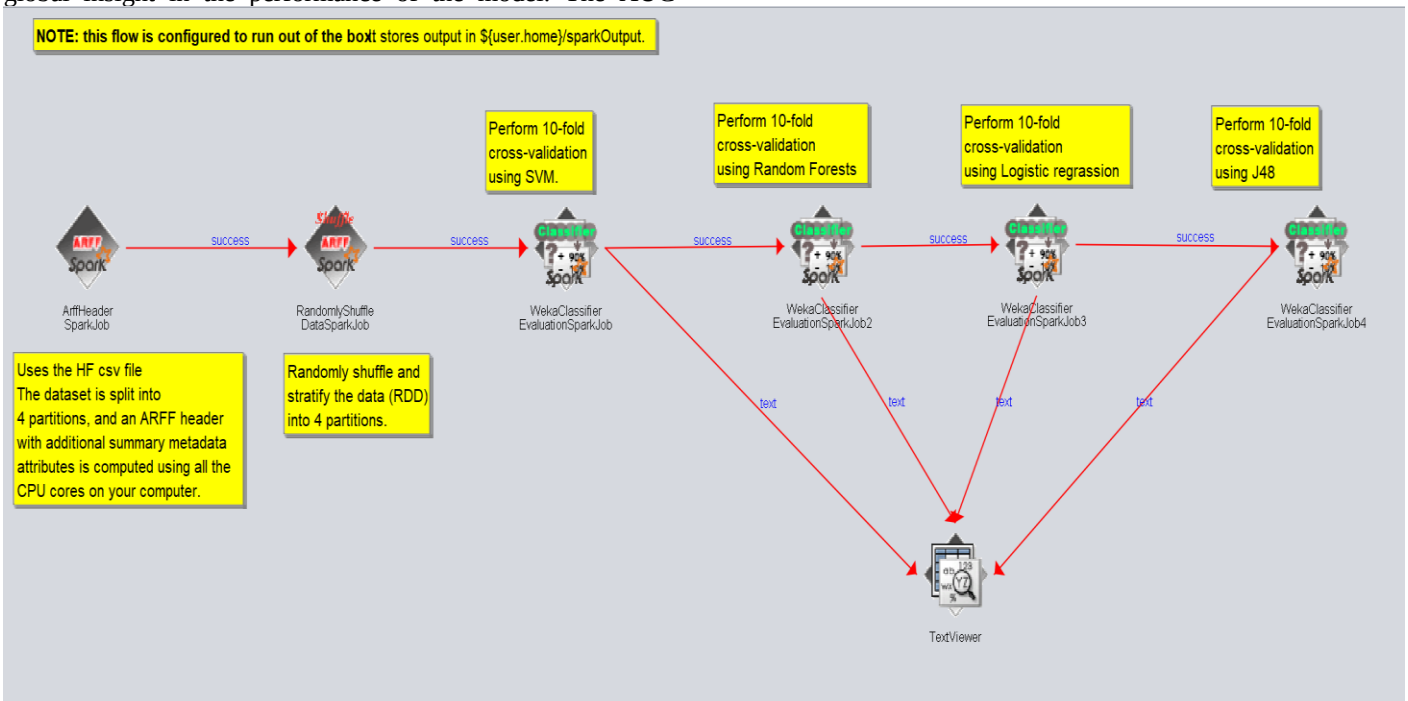


Fig. 6. The Proposed Knowledge flow using distributed Weka.

TABLE V. A PERFORMANCE MEASURE BASED ON RANDOM FOREST CLASSIFICATION ALGORITHM

| | Precision % | Recall % | AUC % | Accuracy % |
|---|---|---|---|---|
| **Structured** | 84.8 | 76.7 | 84.2 | 76.6 |
| **Semi-Structured** | 85.5 | 79.3 | 97.6 | 79.3 |
| **Un-Structured (Haar)** | 80.5 | 78.3 | 93.9 | 78.2 |
| **Un-Structured (LBP)** | 85.5 | 80 | 88.4 | 80 |
| **Integrated Dataset (Haar)** | 88.9 | 87.5 | 90 | 87.5 |
| **Integrated Dataset (LBP)** | 94.3 | 93.3 | 94.2 | 93.3 |

TABLE VI. A PERFORMANCE MEASURE BASED ON LOGISTIC REGRSSION CLASSIFICATION ALGORITHM

| | Precision % | Recall % | AUC % | Accuracy % |
|---|---|---|---|---|
| **Structured** | 79.2 | 60 | 78.3 | 60 |
| **Semi-Structured** | 40.1 | 41.3 | 70.5 | 41.3 |
| **Un-Structured (Haar)** | 80.5 | 78.3 | 93.9 | 78.2 |
| **Un-Structured (LBP)** | 76.1 | 56.7 | 66.5 | 56.6 |
| **Integrated Dataset (Haar)** | 91.7 | 91.7 | 80.3 | 91.6 |
| **Integrated Dataset (LBP)** | 93.3 | 93.3 | 94.3 | 93.3 |

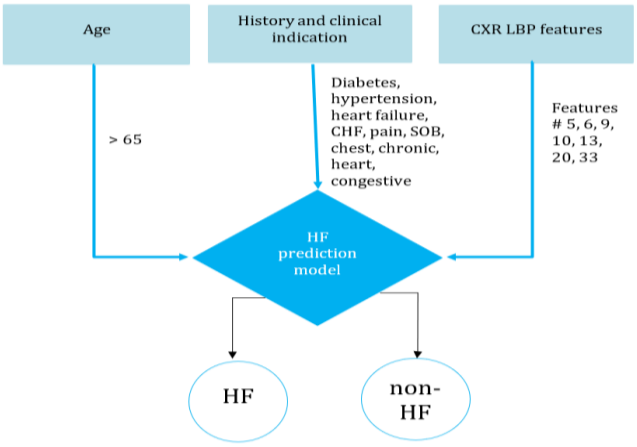


Fig. 7. HF integrated model.

The results of this study found 21 predictors, where the most powerful ones were older age, diabetes, hypertension and LBP features. It explains that if a patient came to the hospital having age > 65, suffering from Diabetes, hypertension, SOB and has some key features in the CXR, then there will be a high chance of having HF. Fig. 7 illustrated the proposed integrated model with the most promising attributes in all dataset. LBP features were used as LBP based model achieved better recall then Haar as mentioned in the previous sections. This also applied to the age range and words selected from the semi-structured data.

## VI. CONCLUSION AND FUTURE WORK

Big Data Analytics provides a systematic way for achieving better outcomes of healthcare service. Non-Communicable Diseases like Heart Failure is one of a major health problems internationally. By transforming various health records of HF patients to useful analyzed result, this analysis will make the patient understand the complications to 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. Combining several characteristics from each patient demographical information, patient clinical information, and patient's Chest X-Ray is a very hard task. In this research, data fusion played a vital role in combining multi-structure dataset. We extracted different important factors of heart failure from King Saud Medical City (KSUMC) system. The extracted data were in the form of structured (patients demographics), semi-structured (patient history and clinical indication), and un-structured (patient chest X-Ray) data. Then we applied some preprocessing techniques to enhance the parameters of each dataset. After that, data was stored in HDFS to be trained and tested using different modeling algorithms on two phases to compare the performance measures of the resulted models before and after integrating them in the first phase we train each dataset as a traditional single predictive. Then, we integrated the most promising attributes form all dataset in the second phase and build 2 models based on Haar and LBP feature extraction. The results showed that the performances of the classifiers were better using the fused data (~93 % accuracy). For further improving, other intelligent algorithms need to be prospectively analyzed as well and more subjects should be investigated to keep upgrading the classifier. We will also incorporate more medical data into the model, better simulating how a cardiologist makes a decision.

## 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: Hadoop, 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 21st 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, 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, 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, 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, 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, J. Ramírez, "A survey of stemming algorithms in information retrieval." Information Research, vol. 19 no. 1, March 2014.

[25] S. Arora, Y. Brar, 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.