PJAEE, 17 (8) (2020)

PalArch's Journal of Archaeology of Egypt / Egyptology

# DATA MINING IN THE TIME OF COVID-19

Ahmad Okash[1], Firuz Kamalov[2], Samer Hamidi[3], Claire Roberts[4], Safa Abdulnasir[5]

[1,2,4, 5] Canadian University Dubai, [3] Hamdan Bin Mohamad Smart University

Corresponding Author[1] a.okasha@cud.ac.ae

**DATA MINING IN THE TIME OF COVID-19, Ahmad Okash1, Firuz Kamalov2, Samer Hamidi3, Claire Roberts4, Safa Abdulnasir5 . -Palarch's Journal Of Archaeology Of Egypt/Egyptology 17(8), 222-246. ISSN 1567-214x**

**Keywords: Micro Business Unit, Competitive Strategy, Performance**

## ABSTRACT:

Healthcare organizations, like other organizations, are facing a major global challenge. In a recent Mckinsey survey (From "wartime" to "peacetime": Five stages for healthcare institutions in the battle against COVID-19, 2020), many consumers indicated that the COVID 19 pandemic has the most significant challenge on their economic and social lives in the last 100 years. Being patient centric rather than reactive is one of the ways to succeed in this uncertain environment. Being patient centric means to identify the needs of patients and design specific programs to address their needs whether they are financial, personal, or clinical. COVID-19 accelerated utilizing data and online applications. Many healthcare organizations have access to consumer related data. Data mining capabilities provide health care organizations with the ability to extract hidden predictive information from large databases.

The paper surveyed hospital Chief Information Officers (CIO), health information managers, and healthcare managers to find out measure awareness of data mining applications in healthcare and to determine the use and reason for data mining applications in healthcare.

The results indicate that many healthcare organizations are aware of descriptive and simple data mining tools. For more sophisticated data mining tools, most healthcare organization managers in the Middle East as expected are not aware of them. When it comes to using data mining as an application for disease diagnoses, marketing, and education simulation, many healthcare managers indicate that they are already using data mining in these areas.

Healthcare organizations in the UAE like other healthcare organization worldwide are already using data mining applications to address clinical and business needs.

## INTRODUCTION

Healthcare organizations, like other organizations, are facing a major global challenge. In a recent Mckinsey survey (From "wartime" to "peacetime": Five stages for healthcare institutions in the battle against COVID-19, 2020), many consumers indicated that the COVID 19 pandemic has the most significant challenge on their economic and social lives in the last 100 years. This shock to the whole global economy has added on the other existing challenges such as maintaining the rising costs and reducing medical errors.

The survival of health care institutions depends heavily on how these organizations respond to the COVID-19 pandemic and market and patient needs. Many countries have enacted social distancing programs to limit social interactions. Many individuals were using online applications during these times. Healthcare organizations have access to huge amount of consumer related data. This type of needed data enhanced the existent administrative and clinical health information.

The healthcare industry is experiencing a turbulent global transformation. The ability of governments to keep funding healthcare at the current escalating cost is questionable. The current global healthcare expenditure is estimated at 9.6 trillion USD (Dieleman, et al., 2017). Healthcare expenditures take a significant share of the gross domestic product (GDP) of many countries. For instance, in the US, the health care spending now represents 17% percent of GDP compared to 8.3% globally and 5.2% in North Africa and Middle East region (Dieleman, et al., 2017). On the other hand, the medical practice which represents a large component of the healthcare industry is still based on an old practice model. The healthcare industry incentivizes specialization and super specialization. The need to control healthcare costs is paramount and it is driving the change of the healthcare industry. For instance, many countries are moving from fee for service payment models that are based on the specialization model, which represent a modest progress toward cost containments.

There are many ways the healthcare industry can be improved. However, the financing framework will have the biggest impact on improving the healthcare industry. The payers of healthcare are under pressure to finance the escalating costs. Payers are starting to force health care organizations to rethink their existing businesses from purely medical care transactions to managing the health of served populations. To improve the outcome of care for their members, payer organizations are turning to demand management. Since a large portion of the health care bill is attributed to chronic diseases, improving the outcomes of care for patients with chronic diseases will lead to reducing the overall costs of care. At the buyers' side, cost cutting measures are creating so-called managed care backlash. Consumers are demanding a broader choice of providers, quality for their health care expenditures and accountability from their health care providers.

The survival of health care organizations, specifically health insurance and provider organizations depend heavily on how these organizations respond to market needs. The market requires health care organizations to lower their operating costs and meet customer needs. Many health care organizations are responding to these needs by merging, acquiring similar organizations or creating alliances. The reason for this is to create large organizations that can invest in developing care management programs for chronic diseases, sharing information with providers and reaching patients and individuals in their homes using the internet. However, mergers and acquisitions may not be the panacea and may only ameliorate operating efficiency in the short run. Therefore, health organizations need to be patient-centric to be successful in the long run.

Being patient centric means to identify the needs of patients, and design specific programs to address their needs (e.g. identify the population at risk and target them with a care management program). To become patient centric, health care organizations need to use available information to derive knowledge about their patients, providers and competitors. Over the past few years, hospitals and payers have been collecting massive numbers of transactions that are buried in their legacy operating systems. Historically data collected were either from insurance claims or hospital administrative data. Many large health care organizations'-built data warehouses. With time, the sizes of these data warehouses were expanding exponentially and the complexity of extracting valuable information has led to the proliferation of online analytical processing (*OLAP*) tools. Nowadays, for the first time, the healthcare sector can collect life-long datasets. Such data include medical and insurance records, wearable sensors, genetic data and social media. This data can be utilized to personalize care, contain epidemics and prevent chronic illnesses through predictive modelling techniques.

The healthcare industry is still fragmented and super-specialized. Health care organizations need to put the patient needs foremost and use data mining to address their needs. Doing this will have a significant impact on consolidating the health industry components into an effective connected system. A recent *Gartner Group Advanced Technology Research* noted data mining and artificial intelligence at the top of the five key technology areas that will have a major impact across a wide range of industries within the next three to five years (Gartner, 2017).

Data mining capabilities provide health care organizations with the ability to extract hidden predictive information from large databases. For example, insurance companies can make use of their claims data to detect early signs of adverse outcomes. Organizations can then engage patients to prevent these adverse outcomes, which will lead to better quality of care and lower overall costs. Analysis of social network interactions can also lead to a prediction of an individual need to consult with a psychiatric. Data mining applications in healthcare has been an active field of research in the recent years (Jothi & Husain, 2015), (Raghupathi, 2016), (Joudaki, et al., 2014; M Shariff et al., 2020). Srinivas, Rani, & Govrdhan (2010) applied various data mining techniques such as *Decision Trees* and *Neural*

*Networks* to large volumes of healthcare data. In particular they used data mining to produce accurate prediction methods for heart attack patients. Aljumah, Ahmad, & Siddiqui (2013) used support vector machines (SVM) to study the effectiveness of diabetic treatments. The authors studied the patients in Saudi Arabia and their diabetic treatments with respect to different age groups. The study showed that younger patients should receive delayed treatment for diabetes to avoid side effects whereas older patients should receive immediate treatment. Zolbanin, Delen, & Zadeh (2015) studied the significance of concurrent chronic diseases during treatment. The authors analyzed datasets for breast and female genital cancer as well as prostate and urinal cancer. Using a variety of data mining methods, authors built a number of predictive models. They concluded that having more information on comorbid conditions improves accuracy of the predictive models. Zheng, Zhang, Yoon, Lam, Khasawneh, & Poranki (2015) studied the risk of hospital readmission using several data mining techniques such as neural networks (NN), random forest (RF) algorithm, and SVM. The proposed classifiers are used to model patients' characteristics such as age, insurance payers, medication risks etc. Experimental results show that the proposed SVM model with particle swarm parameter tuning yields robust accuracy results with very high sensitivity level.

We argue that the healthcare industry should put the patient at the center of the healthcare structure and meet their needs. This paper addresses the issue of how to leverage the mountain of available data to become a patient-centric organization. This article discusses the areas where health care organizations can apply data mining and explores the different techniques and business applications for data mining as well as providing guidelines for implementing a successful data mining project.

## LITERATURE REVIEW AND HYPOTHESIS

### *Applications of Data Mining*

Healthcare organizations can employ data mining as an instrument to extract hidden predictive information from large databases to become patient centric organizations.

### *Disease diagnoses*

According to Cindy McConnell, a spokeswoman at *NIH's National Center for Advancing Translational Sciences*, the number of known human diseases is in the thousands of which only 500 diseases have approved treatments (Kessler, 2016; Muhammad et al., 2019). The task of identifying a diagnosis of a disease is very difficult and sometimes overwhelming for physicians. Improving the accuracy of diagnosis will lead to significant improvement in the quality of healthcare and

reduce the cost. Many physicians overestimate their level of confidence in their accuracy of diagnosis (Nederhand, Tabbers, Splinter, & Rikers, 2017; Munir et al., 2019; Noorollahi et al., 2019). Meyer, Payne, Meeks, & Rao (2013) examined physicians' accuracy and have found that physicians' accuracy level was 31% across four medical cases. Hanrahan, et al, (2018) found that physicians' prognosis accuracy for terminally ill patients was 20%.

Data mining has been used recently to diagnose a wide array of diseases. Data mining techniques has been used to diagnose brain abnormality by analyzing MRI images with an accuracy percentage of 71% (Noreen et al., 2019; Ramani & Sivaselvi, 2017). Siuly, Li, & Wen (2011) have used clustering techniques to classify electroencephalogram (EEG) for elliptic, motor imagery and mental imagery with classification accuracy of 94%, 84%, 62%, respectively. Machine learning has been used to design computer aided diagnosis (CADx) for detecting the presence of diabetes, heart and Parkinson disease with an accuracy of 72%, 78% and 84%, respectively (Normalini et al., 2019; Ozcifta & Gultenb, 2011).
Therefore, data mining can be leveraged by physicians to improve diagnoses accuracy and thus provide more effective and more efficient services to patients.

One of the main issues with healthcare data is the imbalance distribution of the target values. There exist a number of machine learning techniques that allow to balance the data in order to improve the performance of the AI algorithms. Kernel density estimation of minority class data has been used successfully to balance data including in healthcare application (Kamalov, 2020a). More recently, gamma distribution has also been applied to generate new minority sample to balance data (Kamalov & Denisov, 2020; Shabbir et al., 2019). In addition, outlier detection techniques based on combination of principal component decomposition and kernel density estimation have been used effectively in imbalanced data context (Kamalov & Leung, 2020)

### *Marketing and Sales Effectiveness*

Getting the right and timely information from potential or existing patients is critical to the survival of healthcare organizations. Many successful data mining projects have been implemented in marketing and sales. Organizations outside health care are using data mining effectively in marketing in addition to huge savings. A private healthcare company has developed a consumer health utilization index (Koh & Tan, 2011). Several healthcare companies such as payers of care are starting to use data mining in areas of market segmentation, direct marketing, and customer profiling. For example, in a proactive health maintenance organization (HMO), where the organization is focusing on wellness and preventing diseases, data mining can be used to identify the factors that are associated with smoking cessation. Once these factors have been identified, direct marketing can target individuals with similar characteristics to those who stopped smoking. This is not

to say that those who failed to stop smoking should be left alone. More research should be carried out to identify ways to convince these persons to stop smoking.

### *Outcomes and Disease Management*

Saving lives and finding better ways to treat patients and improve their outcomes is at the core of healthcare organization mandates. Early detection of life-threatening conditions is crucial to saving lives. Massive amounts of available clinical and claims data can be leveraged to find early signs of life threating conditions. Traditional statistics have been extensively used in outcome management and clinical analysis. In a subsequent section, data mining will be shown to be a more efficient way of performing clinical analysis. Data mining automates many of the steps taken to perform traditional statistical analysis. In addition, data mining overcomes many of the limitations of traditional statistics such as that of linearity for regression analysis. Data mining can be used to answer questions such as: what is the likelihood that a congestive heart failure (CHF) patient will develop an adverse outcome after open heart surgery? What are the predicted asthma inpatient admissions for a given hospital? How many pregnant women are most likely to undergo cesarean section? What is the likelihood that patients taking drugs A and B will also be taking drug C?

### *Fraud and Waste Detection*

The World Health Organization (WHO) estimates that about 20–40% of resources spent on health are wasted. The most common causes of inefficiency include inappropriate and ineffective use of medicines, medical errors, suboptimal quality of care, waste, corruption and fraud (Hamidi, 2016). For example, in the US health system, billions of dollars are wasted each year because of fraudulent practices. King (2014) maintained that an improper payment by CMS is estimated around $50 billion. Some of these services might be fraudulent. Detecting fraud is not an easy process. Abuse and waste involve the question of medical necessity. Were specific tests really needed? Was this high-cost treatment the best cost? A practice pattern may prove fraud. Data mining has been used to detect fraudulent payments in claims data (Copeland, Edberg, Panorska, & Wendel, 2012; Shabbir et al., 2020). Bauder, Khoshgoftaar, & Seliya, (2016) maintained that data mining can be used to detect up coding. Up coding occurs when a healthcare provider obtained additional reimbursement by coding a certain provided service as more expensive as what was delivered. Minimizing waste and efficiencies in healthcare is essentially a financial resource value care process from a patient's point of view.

### *Simulation and Medical training*

Data mining has been used recently to develop virtual laboratory for medical and biology students to develop their practices of immunological technics (Slimani, Elouaai, Elaachak, Yedri, & Bouhorma, 2018). Data mining techniques has also been used to develop adaptive clinical decision support systems (Anima & Kumar, 2018). These clinical decision support systems can be used effectively for medical training. For instance, Eliot, Williams, & Woolf (1996) have developed an intelligent tutoring system software to teach teaching cardiac resuscitation techniques. The system is personalized for each student based on their interaction with the system.

## DATA MINING TECHNIQUES

Let us look further at data mining techniques that can help health care organizations transform into customer-focused organizations. It is noteworthy to restate that one of main advantages of data mining is that it automates the analysis steps. Thus, it saves time and resources. The following sections discuss several techniques which are well-developed than others and have been applied in many areas in health care and in industries such as banking, insurance, and defense. These techniques are presented in the context of business applications that can be used by health care organizations to become more customer focused.

### *Descriptive statistics*

The healthcare industry is fast paced. Countless research are conducted every year for diagnosis and preventive purposes, whether it is for vaccines or new medications. All these studies generate a big amount of data which would be impossible to understand without the use of statistics.

Descriptive statistics serve as the first step to understand the data. Various descriptive analyses can serve to understand data. Descriptive statistics (mean, median, and mode) and charts (bar charts, pie charts, etc.) are used to describe the given set of data sample in an understandable manner.

Descriptive statistics is used to summarize and analyses a large amount of data into something that is simpler to present. This type of data helps us recognize patterns which can be the basis of a future research (Kenton , 2017).

For instance, to conduct a study to evaluate the prevalence of diabetes in a population, a sample would be taken that represents the population. After choosing the sample and collecting the necessary data, it would be presented through charts and graphs to summarize the findings generate an analysis about the population which would later help in creating an initiative to decrease the prevalence of diabetes, if needed.

### *Basic Predictive Statistics*

Preventive care is an essential part of healthcare system. By taking preventive steps and being aware of potential risks a patient care avoids serious diseases. As part of preventive care, it is important to understand and identify the relevant risks. For instance, a patient with a history of heart problems needs to know what factors increase the risk of a heart disease. Likewise, a patient needs to be aware of the factors that reduce the risk of a disease. However, it is not always easy to determine the relevant factors. This is the problem of feature selection. In the context of data mining, the goal of feature selection is to identify the independent variables (features) that are most relevant to the dependent variable (target class). The task of feature selection is complicated by the interactions among the features. It is possible that a pair of features may not be relevant to the target class individually but be highly correlated with the target class when taken together. In theory, given a set of features one must perform an exhaustive search through the space of all possible subspaces of the feature set to determine the optimal subset. An exhaustive search is very time consuming. Therefore, in practice various heuristic methods are used to approximate the optimal subset. Effective feature selection also helps in classification tasks. For instance, in the case of micro-array data analysis the number of features can reach thousands. To build a classification model based on such a high dimensional dataset would be computationally difficult. So, researchers often perform feature selection as a preprocessing step before building a classification model. Another advantage of having fewer features in a model is the interpretation of its behavior.

Another part of data mining that deals with features is feature engineering. In feature engineering, new features are derived from existing features. One such technique is called Principal Component Analysis (PCA). Given a dataset a researcher can use PCA to create an orthogonal set of features that span the sample space. The advantage of orthogonal features is that they do not interact with one another. Therefore, it is easy to identify the most relevant features in the new set.

Traditional statistics can be used to address all the identified data mining business applications. However, traditional statistics is time consuming and resource intensive. It requires highly skilled statisticians and analysts to design and implement the analysis step by step. Conversely, data mining automates many of the steps in traditional statistics and model building. For example, if an analyst wants to predict risk factors for CHF patients, it will take one-week to generate, validate, and test a regression model, which is a traditional statistical technique. Using decision trees, a data mining technique, the same model building steps can be performed in less than a day. If the analyst wants to predict risk by age group, it will take three weeks to develop the regression models, but it takes less than a day to develop the models using decision trees. Table 1 shows overlaps between data mining techniques and what they can do. In some cases, the data mining analysts can choose freely between two techniques to perform an analysis. For instance, to perform prediction, analysts can use either neural networks or decision trees. However, in other cases, the business problem and data characteristics determine which data mining technique should be used. For example, when the output

variable is a categorical variable such as age group, neural networks cannot be used to model this variable. Decision trees can model a categorical variable.

### *Linear Regression*

Regression analysis is a statistical method that attempts to find relationships within a data set. Concretely, the goal of linear regression is to apply a linear model to estimate the relationship between the dependent and independent variables. A linear model is the simplest functional method to describe the relationship between variables. Linear regression is less likely to over fit the data compared to other more complex non-linear models. It is a well-known and widely used tool in statistical analysis. Linear regression is a standard tool implemented in all statistical packages and software. The results of linear regression are easily understood and interpretable. In the context of medical applications, linear regression can be used in a number of different ways. For instance, we can use linear regression to study the effects of various factors on patient blood pressure. Linear regression can be also modified to accommodate non-linear models making it indeed a powerful statistical analysis tool.

### *Decision Trees*

A multi-hospital system has built a women's health care center. To increase visits, the marketing department wants to roll out a direct marketing campaign. To ensure the success and cost effectiveness of the direct marketing program, decision tree models can be used to predict the women who are most likely to visit the center. This information helps the organization in focusing its marketing messages. Another example of the use of decision trees is to find factors associated with blood pressure. Blood pressure scores are then converted into low, medium, and high. This is the dependent variable. Examples of independent variables are those related to exercise, diet regimen, stress, attitude, heredity and medical history. Decision trees can find the variables associated with low, medium and high blood pressure. Once these factors are identified, specific health education materials can be offered to patients with a high risk of developing high blood pressure and to their providers. Decision trees are well known in data mining since they produce powerful models.

Decision trees are tree shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Each node in a tree represents a question and the branches of the node correspond to the answers to the question. For instance, a node can correspond to the question "Is the patient male or female?" and two branches emanating from the node would be a "Yes" or "No" answer to the question. Another way to think about nodes and branches is that each node corresponds to an explanatory (independent) variable and the branches

correspond to the values of the variable. To determine the next variable on which to split a subtree we calculate the Information Gain with respect to dependent variables. Information Gain is a measure of mutual information between the independent and dependent variable. It is calculated as the difference between the entropy of the dependent variable and the conditional entropy of the dependent variable on the independent variable. The variable that has the highest Information Gain is used for further division of the tree.

Until the previous decade decision trees were one the most popular classification algorithms. However, they have been supplanted by the more modern methods such as SVM and Neural Networks. Nevertheless, decision trees remain in use by many practitioners due to their ease to understand structure. Decision trees are user friendly and straightforward to interpret. This allows a user to analyze the tree and make necessary adjustments. The decision trees were originally designed to work with categorical data. However, one can adjust the decision tree algorithm to work with continuous data using discretization. Given continuous data one can discretize it by various methods and then use it as if it were categorical data. Decision tree methods include Classification and Regression Trees (CART), Chi Square Automatic Interaction Detection (CHAID), and C-5.

One of the main approaches to improve the effectiveness of decision trees is by using informative features. To this end a number of machine learning approaches have been proposed. The least squares approach uses the minimum L2 norm to select the best features (Thabtah et al., 2020). Variance decomposition allows to account for interaction between the features to select the best subset (Kamalov, 2018).

### *Advanced Neural Networks: Artificial Intelligence*

Now suppose a hospital wants to predict utilization or to estimate the likelihood that patients will develop adverse outcomes. Neural networks would be the optimal analytical technique here. Neural networks, non-linear predictive models that learn through training and resemble biological neural networks in structure, produce powerful models. Neural networks can predict adverse outcomes for patients with congestive heart failure (CHF). Using claims data and other available clinical data, data mining analysts train neural networks to predict the likelihood that CHF patients will develop adverse outcomes. A high score indicates a greater likelihood of developing adverse outcomes. This knowledge can be used to devise a special care management program to reach people with a high likelihood of developing adverse outcomes. Neural networks have been used in industries such as banking, insurance, high-tech and health care to predict risks and estimate churn rates. Neural networks enjoy wide applicability, since they can predict non-linear behaviors. The one drawback however, is the difficulty of explaining their output models. Even though they are very powerful tools for prediction, neural networks mathematical techniques are considered a “black box,” because while the model

can identify a CHF patient with a high probability of developing adverse outcomes, it cannot explain why or how.

Neural networks consist of input, output, and "hidden" layers. The input layer of a network consists of "n" nodes where "n" is the number of input (independent) variables. The output layer consists of "m" nodes corresponding to the number of different values of the target class. The number of hidden layers can vary depending on the context, as does the number of nodes in each hidden layer. For instance, if a neural network is designed to classify whether a patient has a particular disease based on ten symptoms, then the input layer would have ten nodes. Each node in the input layer would correspond to a symptom. The output layer would consist of two nodes corresponding to "yes" and "no" values of the dependent variable. We can also add a couple of hidden layers to this network. As mentioned above the number of hidden layers and the number of nodes in each hidden layer depends on the problem.

Every node between two adjacent layers of a network is connected via a "branch." The branches are assigned different weights based on which the nodes affect one another. Thus, the values of the nodes in the preceding layer together with their weights determine the values of nodes in the next layer via a linear combination. Initially, the weights are assigned at random. Afterwards, as the network analyzes examples from a training set it adjusts the weights to maximize the accuracy of the network. In this sense a neural network is nothing more than a nonlinear mathematical function, albeit a very complicated one, with the weights in the network as its variables. Since the number of branches grows exponentially neural networks require a large computing capacity. Therefore, they became popular only recently with the advent of supercomputers. As the number of hidden layers in a network increases so does the accuracy of the model. However, practitioners must be wary of using too many hidden layers in a network. First, each new layer increases the computational task of the network. Secondly, the risk of over fitting also increases as more variables are added to the model as part of a new layer. Although neural networks were originally developed for classification tasks such as digit recognition, they can also be applied in regression (Kamalov, 2020b).

## RESEARCH METHODOLOGY
## METHOD AND DATA

Based on the mainstream literature on data mining, the paper identifies main data mining techniques and their relevant application in healthcare sittings. A questionnaire was developed and sent to healthcare information technology

managers within hospitals and healthcare institutions. Survey data were collected and analysis and tabulated as shown in the results.

### *Survey description*

The survey consists of two parts. The first part measures the awareness of data mining applications among technology mangers. In particular, the survey participants were asked the following question: "*How familiar are you with the following data mining application?*" The data mining techniques included in the survey are as follows:

1. Descriptive statistics: summary statistics (mean, median, standard deviation) and charts (pie charts, bar charts).
2. Basic predictive statistics: hypothesis testing, confidence intervals
3. Linear regression (OLS)
4. Decision trees
5. Advanced classification methods (SVM, XGBoost)
6. Neural Networks

The participants were asked to identify their level of expertise on the scale of

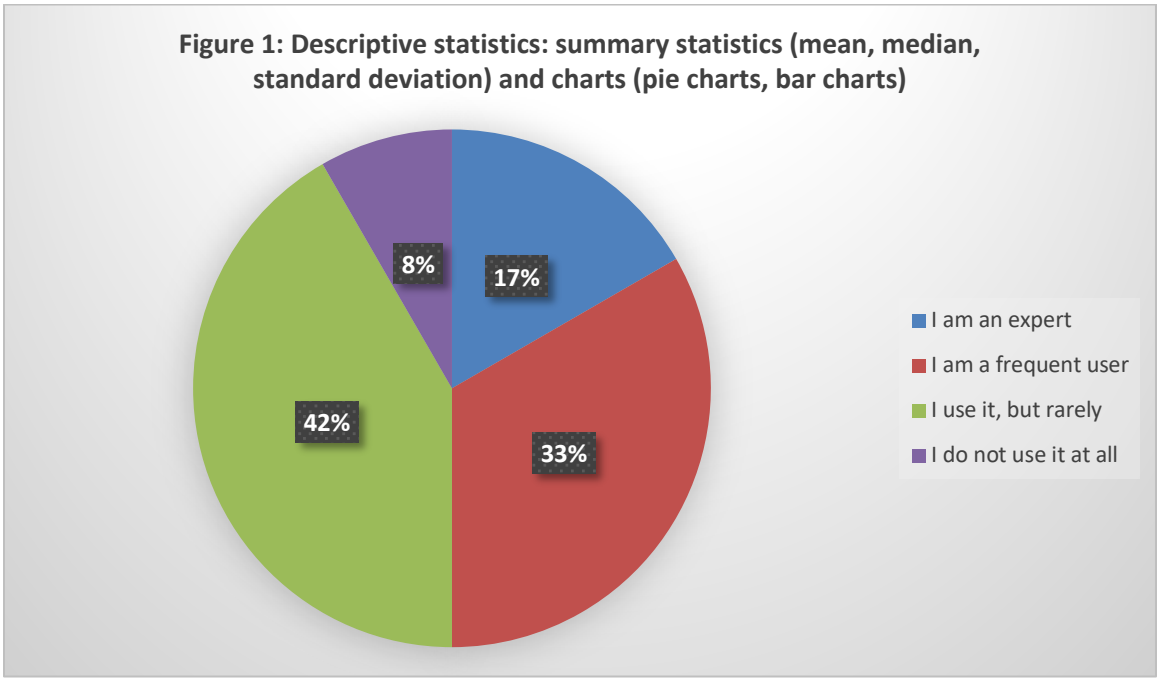
I am an expert
I am a frequent user
I use them but rarely
I do not use them at all

The second part is related to the applications of data mining techniques in healthcare setting. The survey participants were asked the following question: "*How often are you using data mining for any of the following any of the following business application?"* The data mining business applications included in the survey are

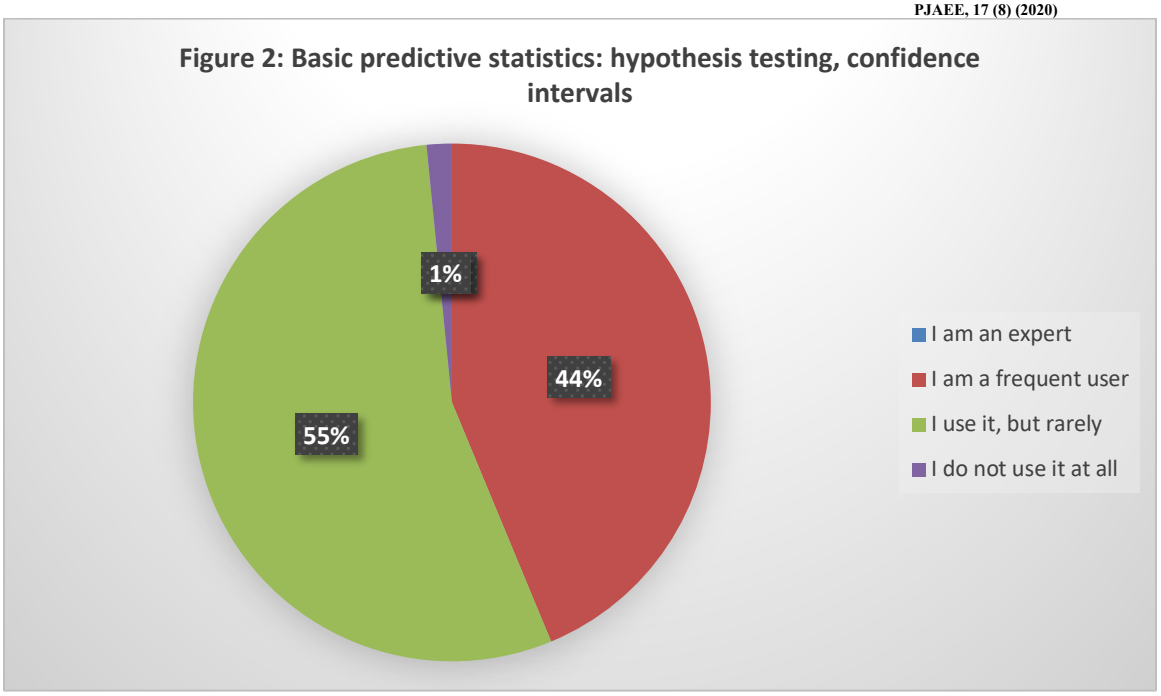
1. Disease Diagnoses
2. Marketing and Sales Effectiveness
3. Outcomes and Disease Management
4. Fraud and Waste Detection

## RESULTS & DISCUSSION

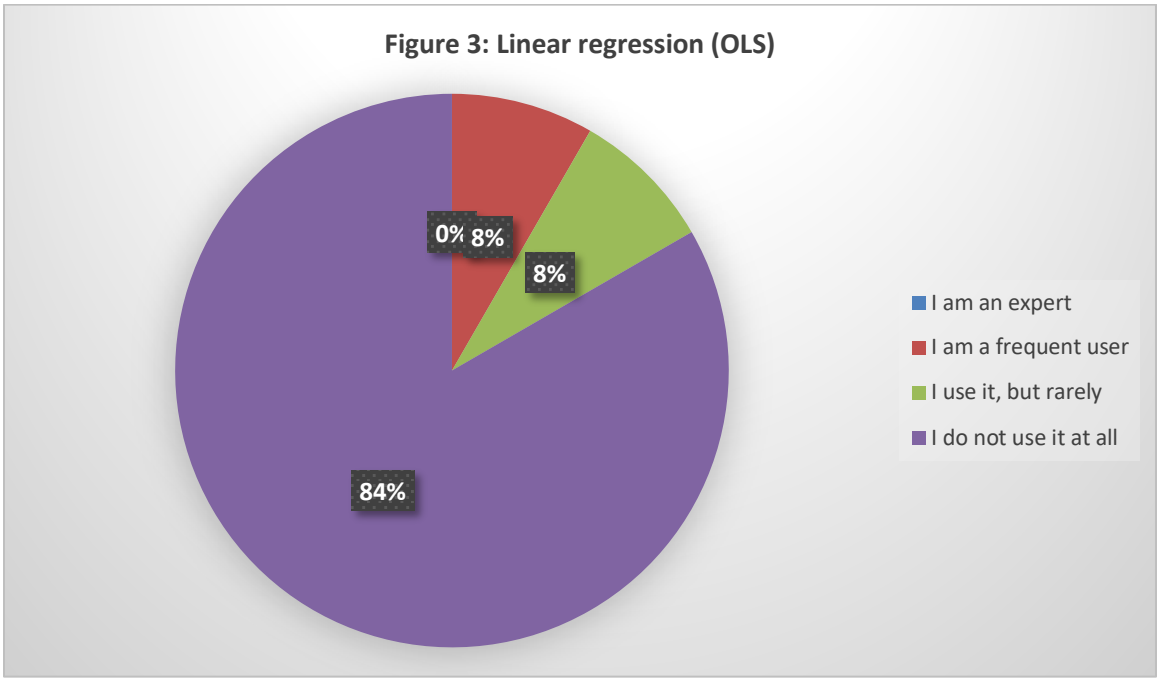
The results of the survey are presented in the figures below. As shown in Figure 1, descriptive statistics is a widely used tool exploratory data mining tool among the technology managers with at least half of the survey participants identifying themselves as either experts or frequent users. Only 8% of the participants indicated that they never use the descriptive statistics in their work. This result is not surprising as descriptive statistics is the most basic tool of data analysis. It is a standard starting point in any kind of data analysis.



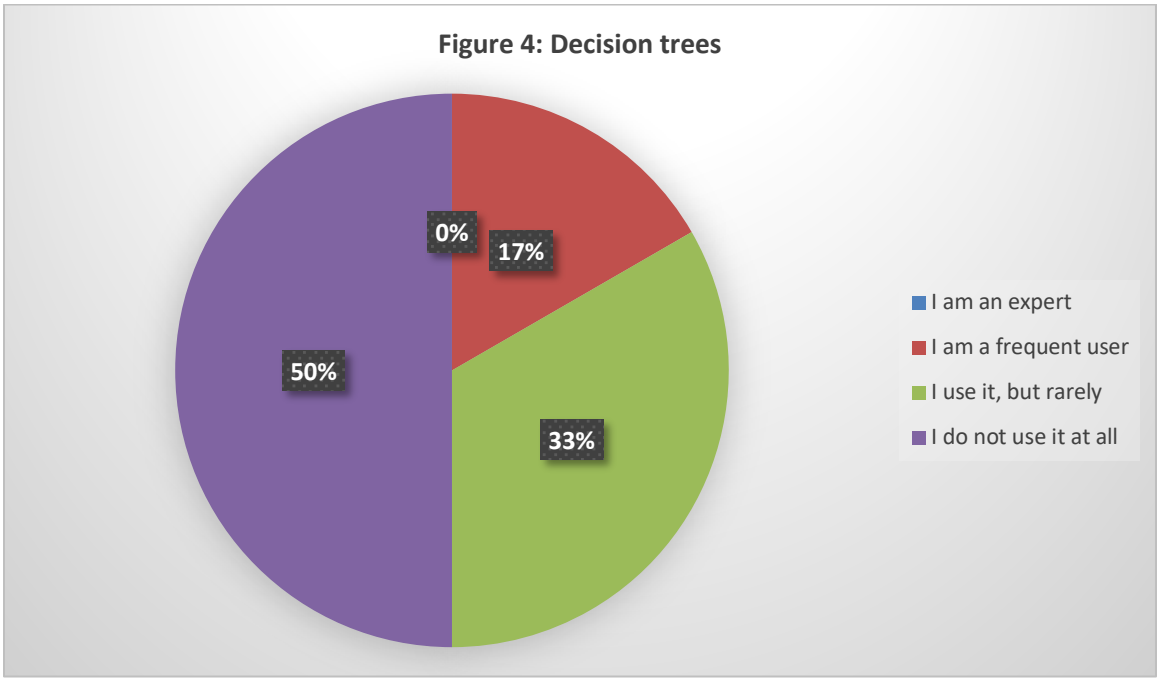

As shown in Figure 2, basic predictive statistics such as hypothesis testing and confidence intervals are another popular tool among technology managers. Concretely, 99% of the survey participants indicated some level of familiarity with basic predictive techniques. This result is again unsurprising since tools such as confidence intervals are widely used in many areas of industry and academia.



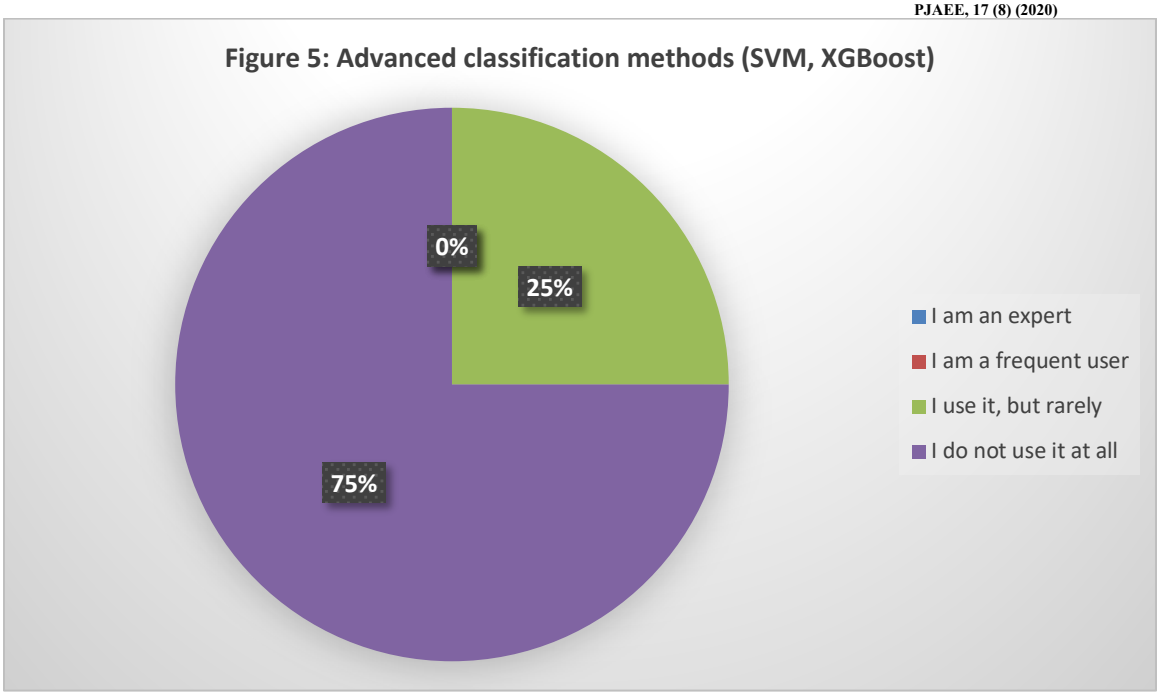

As shown in Figure 3, linear regression has a limited application among the technology managers with only 16% of the participants having some level of familiarity with the method. This result is somewhat surprising since linear regression is a fundamental forecasting tool in statistics. It is implemented in every available software package and can be of great use for planning and managing. Many healthcare organizations have been using linear regression for many years **(Pakt, 2017)**. The results of the study indicate that the UAE healthcare industry is still behind in terms of using linear regression.

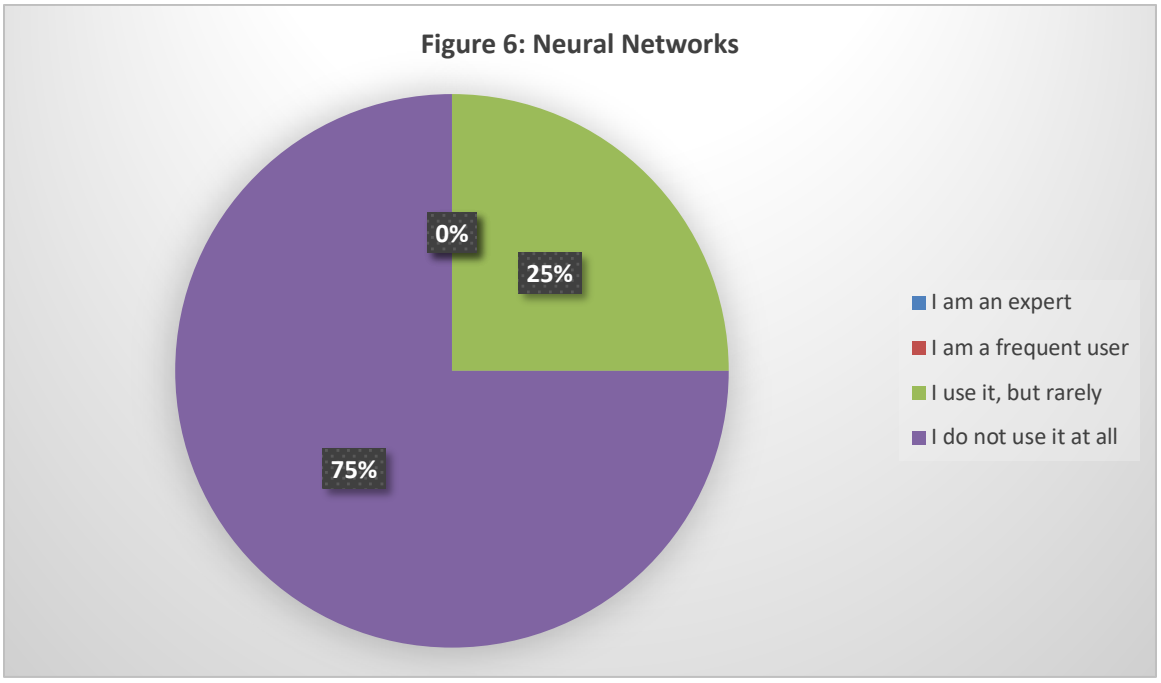


As shown in Figure 4, decision trees are more popular among technology managers then linear regression with half of the participants indicating at least partial familiarity with the method. Decision trees are an intuitive tool that can be effective in decision making. It is structured in an accessible manner that is easy to understand which undoubtedly contributes to its popularity among the users. It is worth noting that none of the survey participants are experts in neither linear regression nor decision trees.



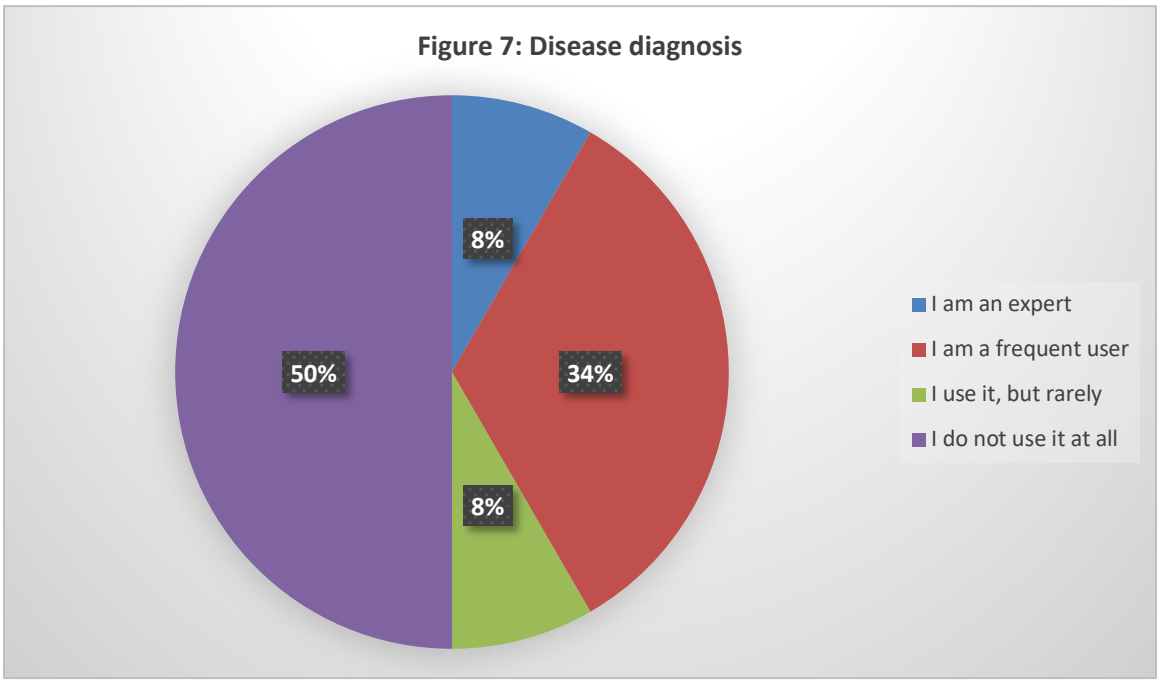

As shown in Figures 5 & 6, technology managers' familiarity with the more advanced analytical tools such as SVM, XGBoost, and Neural Networks is limited. Only 25% of the survey participants indicated a spare use of the advanced methods. This result indicates that the cutting edge research in machine learning is yet to be widely adopted by the UAE healthcare industry. One of the main reasons for the limited adoption of advanced analytics is the absence of reliable and easy to use software that would allow less technically oriented managers to use the tools. Therefore, there remains a great deal of work between the healthcare industry experts and technology companies to deliver the right software allowing for implementation of the new machine learning methods.



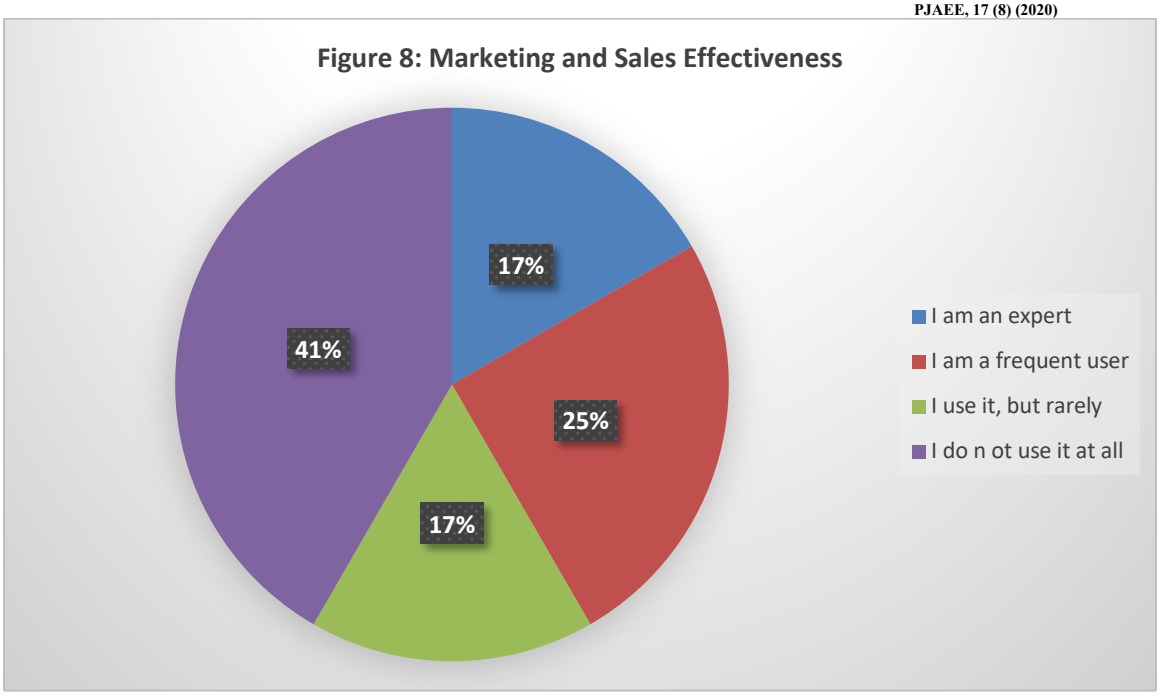




The second portion of the survey attempted to identify the extent to which the technology managers applied data mining and machine learning techniques in their business activities. As shown in Figure 7, half of the survey participants use data mining in disease diagnosis with 42% using it frequently. The frequent use of data mining in disease diagnosis is an encouraging sign as health care providers are taking advantage of the modern analytics tools to make better informed decisions. It also indicates that the modern machine learning methods are advanced enough to be applied in real life situations. Although such a considerable rate of adoption is a positive sign there is a considerable room for improvement. Disease diagnosis is an

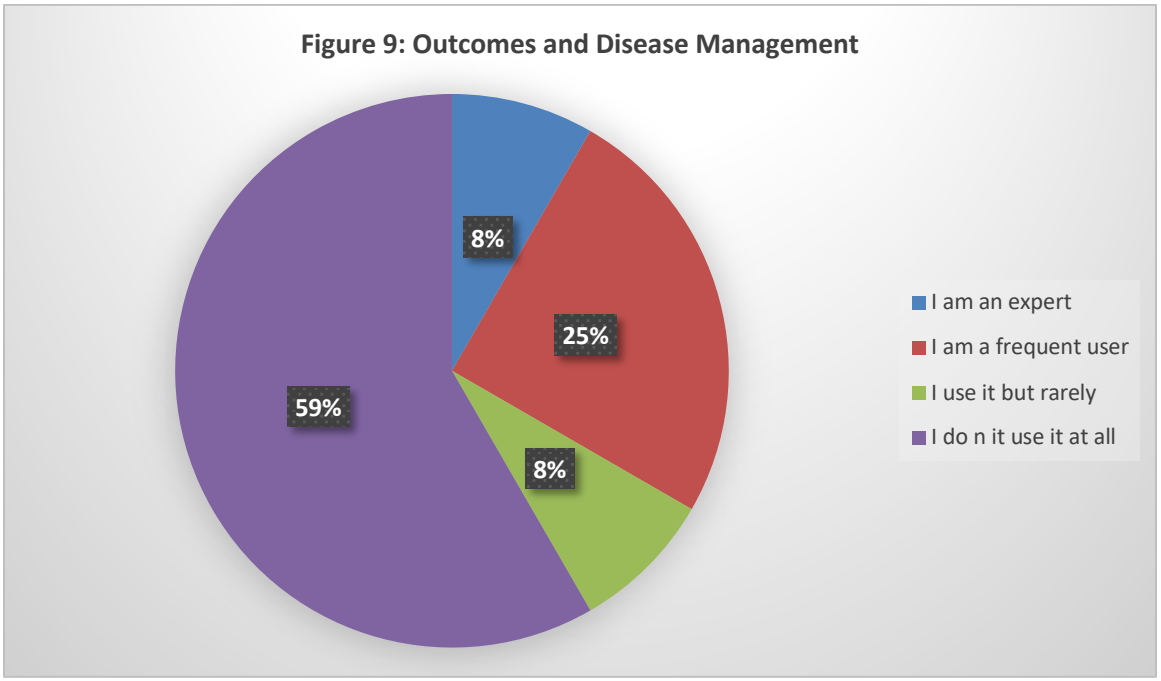
extremely active research area in academia and holds a lot of promise. The current advances such as Google's breast cancer detection system already outperforms human physicians. We expect machine learning to play a major role in the coming years in predictive health analytics.



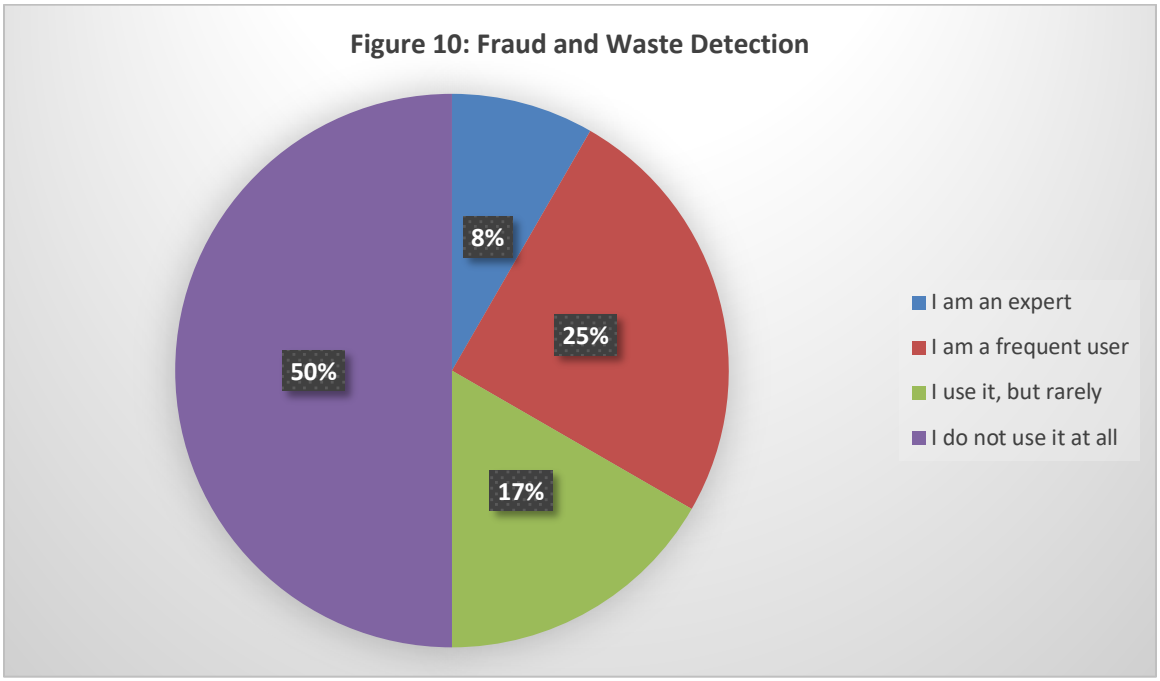

As shown in Figure 8, approximately 60% of the technology managers employ data mining techniques to improve marketing and sales effectiveness. In particular, 42% of the respondents indicate being either a frequent or expert users. This is a positive sign of the industry adoption of the latest data mining techniques in marketing. Market segmentation and analytics is an active research area. Many technology companies including Amazon, Netflix, and others have already widely adopted data mining and machine learning tools to improve their marketing efforts. We expect the health care industry to continue adoption of data mining tools in marketing.



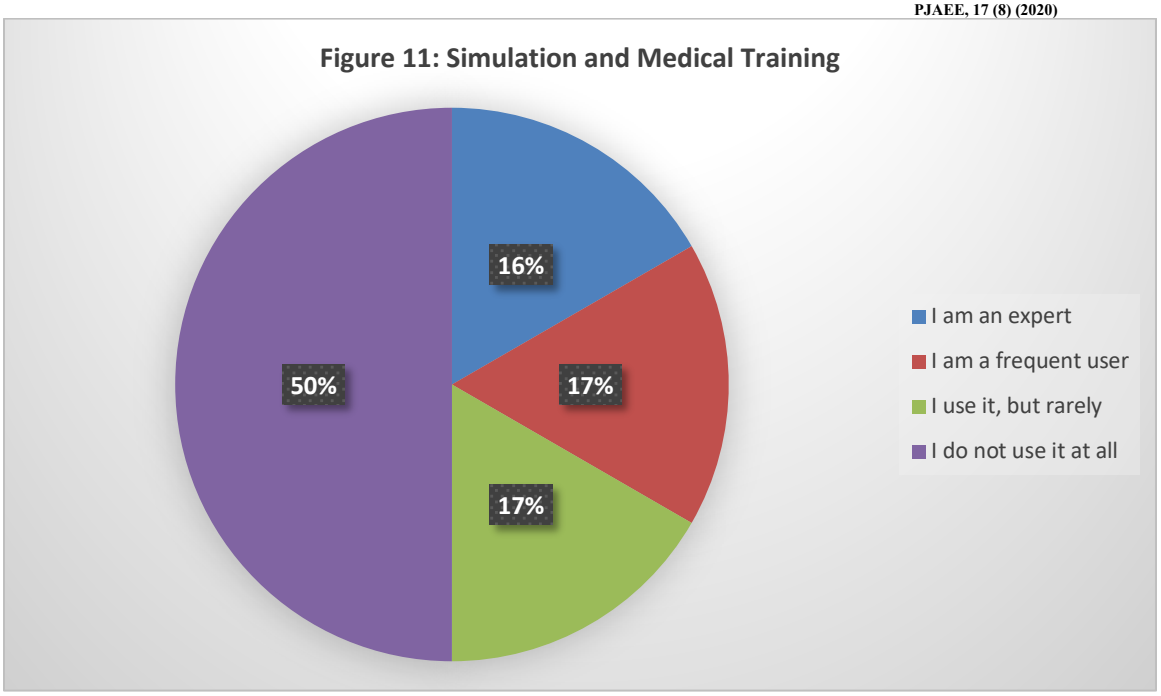

As shown in Table 9, approximately 40% of technology managers employ data mining in outcomes and disease management. Although it is a smaller rate of adoption than the previous categories such results are unsurprising. Disease management is a complicated issue that requires a great deal of human involvement and decision making. The current data mining tools have a limited application in outcomes and disease management. We expect outcomes and disease management to remain a human driven domain augmented by the AI technology.



As shown in Figure 10, half of the respondents employ data mining and machine learning in fraud and waste detection with 32% identifying themselves as frequent or expert users. Fraud detection is one of the major applications of the modern machine learning. Insurance and credit card fraud are actively research use cases in academia. Many financial and insurance companies employ some form of AI enabled technology to flag potentially fraudulent claims and transactions. Machine learning research has made great strides in this area resulting in its adoption by the healthcare industry. We expect data mining and machine learning to be widely adopted in health care industry for the purpose of fraud detection. The continued improvements and capabilities of machine learning systems will result in greater use of these tools in the industry.




As shown in Figure 11, half of the technology managers employ data mining tools in simulation and medical training with a third of the respondents using it frequently. The recent advances in deep learning such Bayesian Neural Networks and Generative Adversarial Networks have shown the creative side of machine learning. These and other technologies are poised to revolutionize the field of medical simulation and training. Therefore, we expect a continued growth of applications of data mining in this field.




It's noteworthy to indicate that the results of the study may seem contradictory. From the one hand, the results indicate low awareness and familiarity of data mining techniques. From the other hand, results indicate that data mining is being used by the healthcare industry to for disease diagnoses, marketing, and fraud and abuse. One explanation is that healthcare information technology and medical equipment are adopting data mining. The UAE healthcare industry is using the latest healthcare technology and medical equipment. For instance, Most of the government hospitals in Abu Dhabi are using Cerner hospital information system, while government hospitals in Dubai are using Epic hospital information system.

## CONCLUSION & RECOMMENDATIONS

### CONCLUSION

Building a patient focused organization is not easy. Data mining can make the chore more tolerable. Data mining can be used effectively to enhance an organizations' understanding about their customer's needs, attitudes and biases. The organizations that predict and understand their customers and re-align their resources to meet those needs are organizations that will succeed and carry their services into the next decade. This study explores the awareness of data mining tools inside healthcare organization and seeks to discover where data mining is being utilized to address business needs. The results indicate that many healthcare organizations are aware of descriptive and simple data mining tools. For more sophisticated data mining

tools, most healthcare organization managers in the Middle East as expected are not aware of them. When it comes to using data mining as an application for disease diagnoses, marketing, and education simulation, many healthcare managers indicate that they are already using data mining in these areas. The explanation for this is that health information technology providers and medical device providers in the Middle East are utilization data mining such as predictive statistics, machine learning, and artificial intelligence to enhance their products and services.

Healthcare organizations in the Middle East focus more on building awareness of data mining tools and leverage their capability to be more customers centric. Their ability to build local predictive models for heart diseases would be valuable.

Additional studies should investigate data mining applications in more details. Understanding which diseases data mining is being used to diagnosed is very important. In addition, it is interesting and useful to understand the linkage between data mining and personal and mobile health applications.

## RECOMMENDATIONS

This study is a foundation study for understanding the prevalence of data mining with the healthcare sector in the Middle East. Future study needs to understand the factors that influence adopting data mining applications in the Middle East.

## References

Aljumah, A. A., Ahmad, M. G., & Siddiqui, M. K. (2013). Application of data mining: Diabetes health care in young and old patients. *Journal of King Saud University-Computer and Information Sciences*, 127-136.

Anima, P., & Kumar, B. (2018). Adaptive Personalized Clinical Decision Support System Using Effective Data Mining Algorithms. *Journal of Network Communications and Emerging Technologies*, 13-18.

Bauder, R., Khoshgoftaar, T. M., & Seliya, N. (2016). A survey on the state of healthcare upcoding fraud. *Health Services Outcomes Research Method*, 31–55.

Berry, M., & Linoff , G. (2004). *Data Mining Techniques For Marketing, Sales, and Customer Relationship Management.* Indianapolis: Wiley Publishing, Inc.

Copeland, L., Edberg, D., Panorska, A., & Wendel, J. (2012). Applying Business Intelligence Concepts to Medicaid Claim Fraud Detection. *Journal of Information Systems Applied Research*, 51-61.

Dieleman, J., Campbell, M., Chapin, A., Eldrenkamp, E., Fan, V. Y., Haakenstad, A., & Reynolds, A. (2017). Evolution and patterns of global health financing 1995–2014: development assistance for health, and government, prepaid private, and out-of-pocket health spending in 184 countries. *The Lancet*, 1981-2004.

Eliot, C. R., Williams, K. A., & Woolf, B. P. (1996). An Intelligent Learning Environment for Advanced Cardiac Life Support. *AMIA Annual Fall Symposium Proceedings* (pp. 7-11). Washington: AMIA.

Gartner. (2017, October 4). *Gartner Identifies the Top 10 Strategic Technology Trends for 2018*. Retrieved March 21, 2018, from www.gartner.com: https://www.gartner.com/newsroom/id/3812063

Hamidi, S. (2016). Measuring technical efficiency of governmental hospitals in Palestine using stochastic frontier analysis. *Cost Effectiveness and Resource Allocation*, 1-12.

Hanrahan, B., Ghearing, G., Urban, A., Plummer, C., Bagic, A., & Antony, A. (2018, January). Diagnostic accuracy of paroxysmal spells: Clinical history versus observation. *Epilepsy & Behavior, 78*, 73–77.

Jothi, N., & Husain, W. (2015). Data mining in healthcare–a review. *Procedia Computer Science*, 306-313.

Joudaki, H., Rashidian, A., Minaei-Bidgoli, B., Mahmoodi, M., Geraili, B., Nasiri, M., & Arab, M. (2014). Using data mining to detect health care fraud and abuse: a review of literature. *Global Journal of Health Sciences* , 194-202.

Kamalov, F. (2018, December). Sensitivity analysis for feature selection. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 1466-1470). IEEE.

Kamalov, F. (2020a). Kernel density estimation based sampling for imbalanced class distribution. *Information Sciences*, *512*, 1192-1201.

Kamalov, F. (2020b). Forecasting significant stock price changes using neural networks. *Neural Computing and Applications*, 1-13.

Kamalov, F., & Leung, H. H. (2020). Outlier detection in high dimensional data. *Journal of Information & Knowledge Management*, *19*(01), 2040013.

Kamalov, F., & Denisov, D. (2020). Gamma distribution-based sampling for imbalanced data. *Knowledge-Based Systems*, *207*, 106368.

Kenton , W. (2017, June 27). *What is Descriptive Statistics?* (Investopedia) Retrieved May 2019, from https://www.investopedia.com/terms/d/descriptive_statistics.asp

Kessler, G. (2016, November 17). *Are there really 10,000 diseases and just 500 'cures'?* Retrieved March 21, 2018, from www.washingtonpost.com: https://www.washingtonpost.com/news/fact-checker/wp/2016/11/17/are-there-really-10000-diseases-and-500-cures/?utm_term=.011d38299df2

King, K. M. (2014). *Medicare fraud: progress made, but more action needed to address Medicare fraud, waste, and abuse.* Washington: The U.S. Government Accountability Office.

Koh, H. C., & Tan, G. (2011). Data Mining Applications in Healthcare. *Journal of Healthcare Information Management*, 64-72.

Meyer, A. N., Payne, V. L., Meeks, D. W., & Rao, R. (2013). Physicians' Diagnostic Accuracy, Confidence,. *JAMA International Medicine, 173*(21), 1952-1959.

Mickinsey. (2020, Aprill 24). *From "wartime" to "peacetime": Five stages for healthcare institutions in the battle against COVID-19*. Retrieved from Mckinsey: https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/from-wartime-to-peacetime-five-stages-for-healthcare-institutions-in-the-battle-against-covid-19

M Shariff, M. N., Ahmad, N. R., & Shabbir, M. S. (2020). Moderating effect of access to finance of the gem and jewelry industry. Utopía y Praxis Latinoamericana, 25, 264-279. http://doi.org/10.5281/zenodo.3809397

Muhammad, S., Shabbir, M. S., Arshad, M. A., & Mahmood, A. (2019). 4th Industrial Revolution and TVET: The Relevance of Entrepreneurship Education for Development. Opcion, 11-21. https://doi.org/10.1201/9780429281501-1

Muhammad, S., Shabbir, M. S., & Kassim, N. M. (2019). Entrepreneur as an Individual: Review of Recent Literature on Entrepreneurial Skills. Opcion, 35, 582-599.

Munir, S., Yasin, M. A., Shabbir, M. S., Ali, S. R., Tariq, B., Chani, M. I., Orangzab, M., & Abbas, M. (2019). Mediating role of organizational citizenship behavior on authentic leadership and employee job performance: A study of higher educational institutes in Pakistan. Revista Dilemas Contemporáneos: Educación, Política y Valores. http://www.dilemascontemporaneoseducacionpoliticayvalores.com/

Noorollahi, Y., Shabbir, M. S., Siddiqi, A. F., Ilyashenko, L. K., & Ahmadi, E. (2019). Review of two decade geothermal energy development in Iran, benefits, challenges, and future policy. Geothermics, 77, 257-266. https://doi.org/10.1016/j.geothermics.2018.10.004

Noreen, T., Abbas, m., Shabbir, M. S., & Al-Ghazali, B. M. (2019). Ascendancy Of Financial Education To Escalate Financial Capability Of Young Adults: Case Of Punjab, Pakistan. International Transaction Journal of Engineering, Management, & Applied Sciences & Technologies. https://doi.org/10.14456/ITJEMAST.2019.200

Normalini, M., Ramayah, T., & Shabbir, M. S. (2019). Investigating the Impact of Security Factors In E-business and Internet Banking Usage Intention among Malaysians. Industrial Engineering & Management Systems, 18(3), 501-510. https://doi.org/10.7232/iems.2019.18.3.501

Nederhand, M. L., Tabbers, H. K., Splinter, T., & Rikers, R. M. (2017, December ). The Effect of Performance Standards and Medical Experience. *Health Professions Education*, 1-8.

Ozcifta, A., & Gultenb, A. (2011). Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms. *computer methods and programs in biomedicine*, 443–451.

Pakt. (2017, December 20). *Healthcare Analytics: Logistic Regression to Reduce Patient Readmissions.* Retrieved February 27, 2020, from https://hub.packtpub.com/healthcare-analytics-logistic-regression-to-reduce-patient-readmissions/

Quality, A. f. (2020, January 28). *Regression Analysis*. (U.S. Department of Health & Human Services) Retrieved January 28, 2020, from https://digital.ahrq.gov/health-it-tools-and-resources/evaluation-resources/workflow-assessment-health-it-toolkit/all-workflow-tools/regression-analysis

Raghupathi, W. (2016). Data Mining in Healthcare. In S. Kudyba, *Healthcare informatics* (pp. 353-372). New York: CRC Press.

Ramani, R., & Sivaselvi, K. (2017). Classification of Pathological Magnetic Resonance Images of Brain using Data Mining Techniques. *Second International Conference on Recent Trends and Challenges in Computational Models (ICRTCCM).* Tindivanam: IEEE.

Ramakrishnan, J., Shabbir, M. S., Kassim, N. M., Nguyen, P. T., & Mavaluru, D. (2020). A comprehensive and systematic review of the network virtualization techniques in the IoT. International Journal of Communication Systems, 33(7). https://doi.org/10.1002/dac.4331

Shabbir, M. S., Abbas, M., Aman, Q., Ali, R., & Orangzeb, K. (2019). Poverty Reduction Strategies. Exploring the link between Poverty and Corruption from less developed countries. Revista Dilemas Contemporáneos: Educación, Política y Valores. http://www.dilemascontemporaneoseducacionpoliticayvalores.com/

Shabbir, M. S., Abbas, M., & Tahir, M. S. (2020). HPWS and knowledge sharing behavior: The role of psychological empowerment and organizational identification in public sector banks. Journal of Public Affairs. https://doi.org/10.1002/pa.2512

Siuly, Li, Y., & Wen, P. (2011). Clustering technique-based least square support. *c omputer methods and programs in biomedicine*, 358–372.

Slimani, A., Elouaai, F., Elaachak, L., Yedri, O., & Bouhorma, M. (2018). Learning Analytics Through Serious Games:Data Mining Algorithms for Performance Measurement and Improvement Purposes. *International Journal of Emerging Technologies in Learning, 13*(1).

Srinivas, K., Rani, B., & Govrdhan, A. (2010). Applications of data mining techniques in healthcare and prediction of heart attacks. *International Journal on Computer Science and Engineering*, 250-255.

Thabtah, F., Kamalov, F., Hammoud, S., & Shahamiri, S. R. (2020). Least Loss: A Simplified Filter Method for Feature Selection. *Information Sciences*.

Zheng, B., Zhang, J., Yoon, J., Lam, S. S., Khasawneh, M., & Poranki, S. (2015). Predictive modeling of hospital readmissions using metaheuristics and data mining. *Expert Systems with Applications*, 7110-7120.

Zolbanin, H. M., Delen, D., & Zadeh, A. H. (2015). Predicting overall survivability in comorbidity of cancers: A data mining approach. *Decision Support Systems*, 150-161.