

Introducing iMhotep Health System, A new Era of Data Mining in Health Care

Abstract:

There is a wealth of hidden gems within the healthcare data that are not yet discovered. Accordingly, it is essential to dig deeper into the medical big data by empowering national data interoperability in emerging healthcare markets. The initial challenges are focused on building-up the health database infrastructure through a comprehensive electronic health system. Also, the exponential growth of digital technology is playing a crucial role in healthcare innovation. Therefore, data mining holds great potential to leverage the full potential of the big data in the health insurance sector.

The paper introduces a new electronic big data management platform for health insurance, iMhotep. Furthermore, the study introduces various data mining techniques within iMhotep framework to extract valuable information and discover patterns that could transform the health insurance industry and public health in emerging markets.

Key Words:

Healthcare, Health Data, Medical Diagnosis, Data Mining, Deep Neural Network, Fraud Detection.

I. Introduction:

The health sector has witnessed a great evolution following the development of new computer technologies, and that pushed this area to produce more valuable medical data that will help the healthcare industry to be on the brink of massive change, which will give birth to multiple fields of research.

Health insurance is attracting more and more attention in low- and middle-income countries as a means for improving health care utilization and protecting households against impoverishment from out-of-pocket expenditures.

Most countries with emerging market are striving to achieve a satisfying health coverage using the available technological development and data mining revolution, but they face a number of challenges

Lack of technology is one of the main challenges, where technology helps in transforming how health care is delivered, giving more people in remote areas and around the world access to better care. Likewise, easier access to data helps both doctors and insurers make better-informed decisions about how to continue to improve the system.

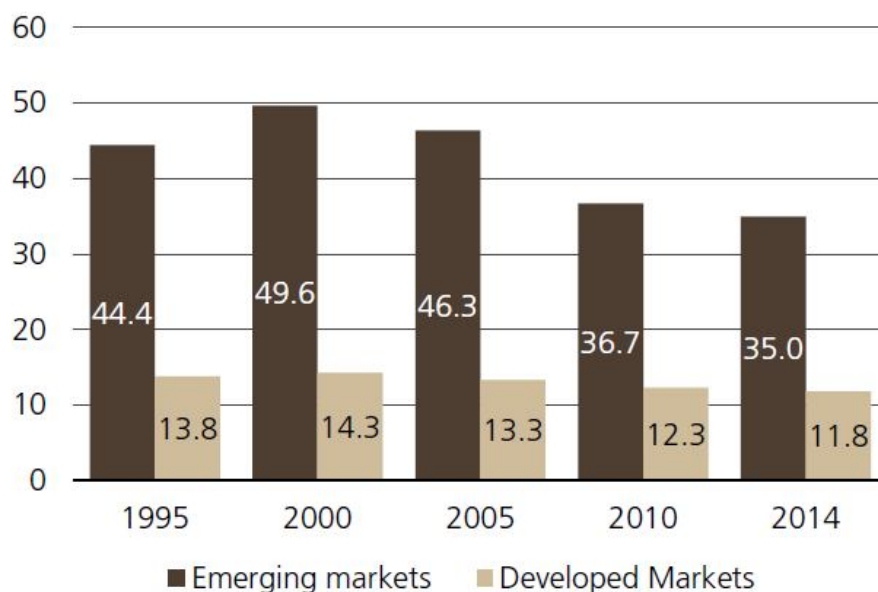
Many efforts are done to cope with the explosion of medical data on one hand, and to obtain useful knowledge from it on the other hand. These prompted researchers to apply all the technical innovations like *big data analytics*, *predictive analytics*, *machine learning* and *learning algorithms* in order to extract useful knowledge and help in making decisions.

With the promises of predictive analytics in big data, and the use of machine learning algorithms, predicting the future is no longer a difficult task, especially for medicine because predicting diseases and anticipating the cure became

possible, which will in turns make predicting insurance cost easier and eliminate the possibility of fraud as it will make detecting fraud easier process.

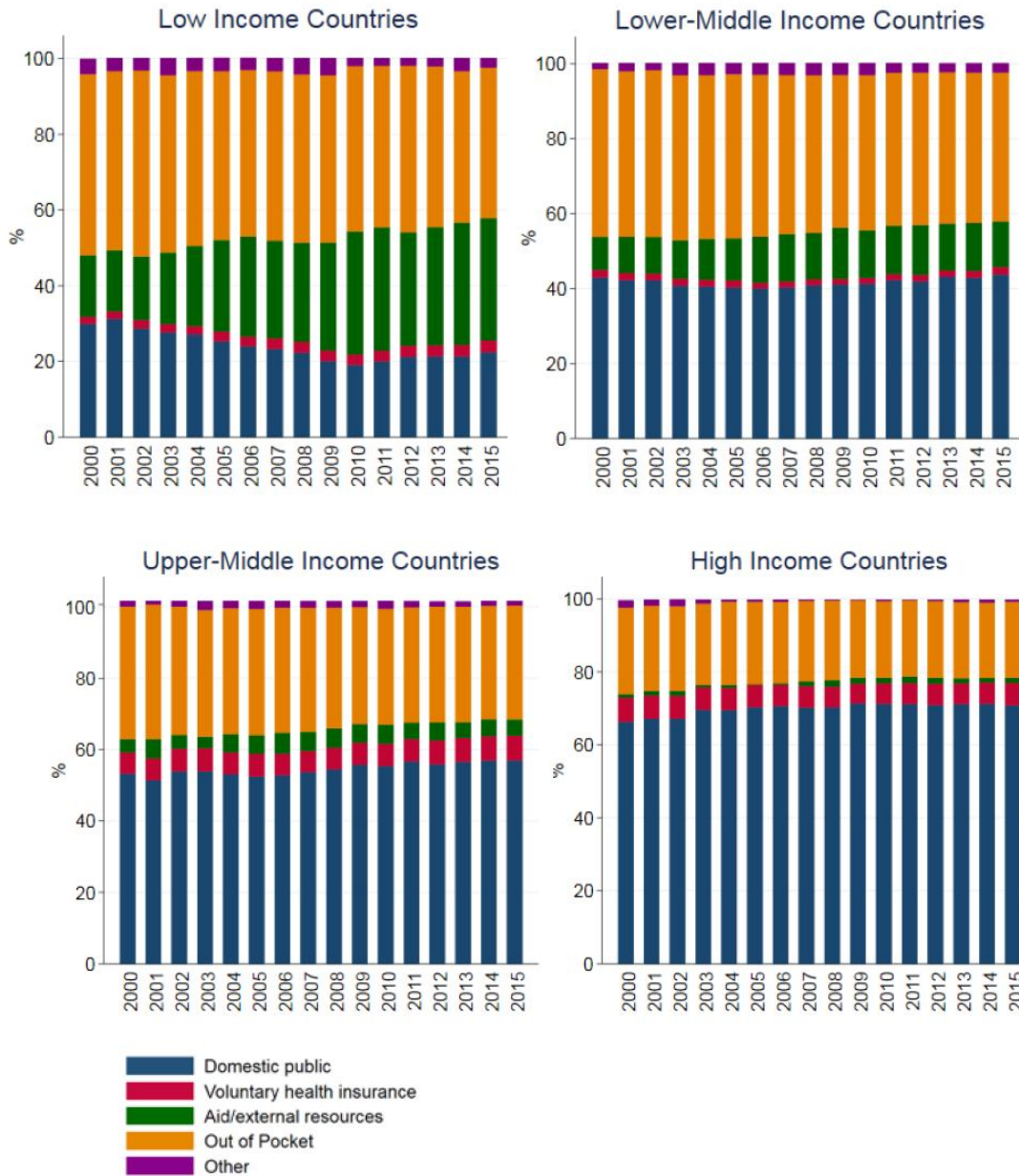
II. Research Problem:

The emerging market healthcare industry has depended on relatively high out-of-pocket expenses alongside rising public spending, where GDP per capita face a huge growth back to health care spending and the government are facing rising costs and inconsistent outcomes. There are billions of people living in countries where out-of-pocket spending is at least 50% of current health spending. Nevertheless, the insurance contributions are relatively minor in low-income and medium-income countries. This is illustrated in the following figure that shows the **out-of-pocket expenditure as a percentage of total healthcare**:



Source: World Health Organization (WHO), World Economic Outlook, UBS, as of 2014

Additionally, to have a deeper look, the following are another figures that represent the **trend in health expenditure sources, by country income group, 2000 - 2015**.



Source: World Health Organization (WHO), Global Report

Increasing the penetration rate of health creates many challenges for insurers and service providers to deliver a better health and better care at a lower cost, where challenges can be summarised in improving the quality of service, reducing error, enhancing the accessibility, increasing efficiency, and lowering costs. This also triggers another drawback; healthcare data tends to reside in multiple places. From different source systems, like HR software, to different departments, like radiology or pharmacy. The data also comes from different organizations and stakeholders. Healthcare data also occurs in different formats

(e.g., text, numeric, paper, digital, pictures, videos, multimedia, etc.). Aggregating this data into a single, central system, such as an enterprise data warehouse (EDW), makes this data accessible and gives a plenty room for predictive analytics.

III. Research Objective:

iMhotep health system, a central electronic health data management platform, aims to expand the realm of possible through using a technological evidence that break existing performance tradeoffs in the attainment of an outcome; providing “more for less” – more value, better outcomes, greater convenience, access and simplicity; all for less cost, complexity, and time required by the patient and the provider, in a way that expands what is currently possible.

iMhotep health system ensures that individuals and communities receive the health services they need without suffering financial hardships. **iMhotep** health system is about ensuring a progressive expansion of coverage of health services and financial protection as more resources become available. Also, **iMhotep** health system minimizes the fraud in claim management process through introducing data mining techniques in detecting fraud.

IV. Research Importance:

The World Health Organization (WHO) considers health insurance a promising means for achieving a satisfying health coverage and our research focus on how to transform this to reality through introducing the new **iMhotep** Health System through public-private partnerships contain all relevant parties to the health care processing.

V. iMhotep Electronic Data Management Platform Framework:

iMhotep, named after the father of medicine in the ancient Egyptian civilization, system is dedicated to providing a comprehensive framework and related standards for the exchange, integration, sharing, retrieval and warehousing of electronic health information. Another important part of the **iMhotep** is data predictive analysis and visualization tool, which supports clinical practice and health insurance, service delivery, patient behaviour tracking, advanced disease diagnostic prediction, and fraud detection.

The first and most important step in constructing **iMhotep** is data construction which is the process of gathering all the data from several sources. This helps the healthcare organizations and insurers treat their patients in a holistic manner, provide personalized treatments and enhance health outcomes.

The data collection phase can be acquired through numerous avenues, such as surveys, electronic financial transactions for health insurance claims, computer-based patient records (CPRs), and disease registries.

1- iMhotep Database building and data collection phase:

Firstly, the main key attributes of data base are limited on the following categories:

- **Demographic data** consists of facts such as age (or date of birth), gender, race and ethnic origin, marital status, address of residence, names of and other information about immediate family members, and emergency information. Information about employment status (and employer), schooling and education, and some indicator of socioeconomic class might also appear.

- **Administrative data** include facts about health insurance such as eligibility and membership, required copayments and deductibles for a given benefit

package. With respect to services provided (e.g., diagnostic tests or outpatient procedures), such data also typically include charges and perhaps amounts paid.

- **Pharmaceutical data** includes information on prescribed or dispensed medicines (including name, dose, duration, route of administration and therapeutic indication), immunisations, diagnosis data and referrals for laboratory investigations, imaging and other procedures.

- **Patient Clinical Records** contain adequate information to identify the patient; support the diagnosis; justify the care and treatment; document the course and results of treatment; promote continuity of care; and any specific content of patient clinical records has been determined by the organization.

- **Health risks and health status** Health risk information reflects behavior and lifestyle (e.g., whether an individual uses tobacco products or engages regularly in strenuous exercise) and facts about family history and genetic factors (e.g., whether an individual has first-degree family members with a specific type of cancer or a propensity for musculoskeletal disease).

2- iMhotep Database Benefits to Stakeholders:

Market drivers are increasing the demand for measuring physician cost and quality performance through integrated database based on the big data collected from the most relevant stakeholders, and stakeholders depend on the expertise of the healthcare analytics to collaborate the data collected from various sources to monitor the efficiency of their processes.

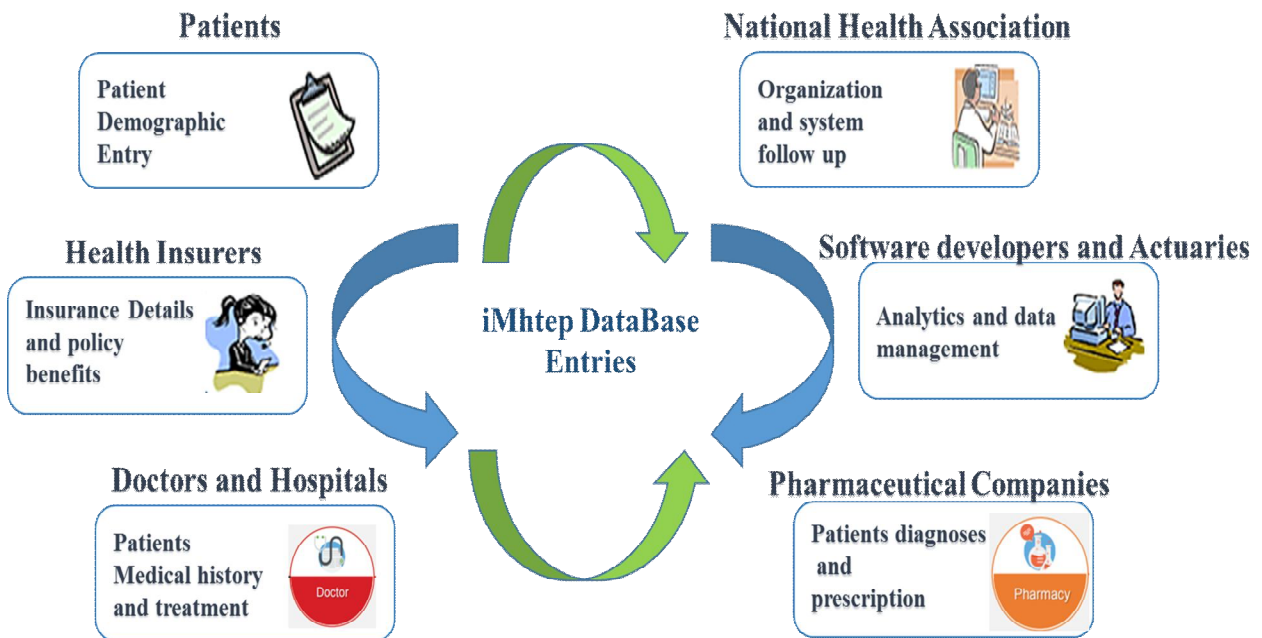
This would help them understand how the patients have responded to their course cure program and what their condition is currently. All these will help in enhancing future expectation and introduce a new era of artificial intelligence and data mining, the following table shows how each relevant parties benefit from iMhotep database:

| | |
|--|---|
| <p style="text-align: center;">Doctors</p> | <ul style="list-style-type: none"> • Use the medical/health record as a means of communication during an episode of care and treatment of a patient, and as an aide memoire for continuing care of that particular patient. • Use health care data to evaluate the services provided; and look up prescription medications. |
| <p style="text-align: center;">Hospitals</p> | <ul style="list-style-type: none"> • Evaluate data and develop critical pathways and patient care plans for admitted patients. |
| <p style="text-align: center;">Pharmaceutical comapnies</p> | <ul style="list-style-type: none"> • Check drug interactions • Gain influence through research and prescription tracking. • Use information they gather and understand what it takes to sell the drugs they represent. • Enrich the research area through introducing new drugs for some diseases. |
| <p style="text-align: center;">Health Insurers</p> | <ul style="list-style-type: none"> • Require information to reimburse the patient and/or health care facility for services rendered whether for an inpatient or ambulatory patient; • Data mining techniques used to help delivering a method for detecting fraudulent events from other parties. |
| <p style="text-align: center;">Courts</p> | <ul style="list-style-type: none"> • Facilitate access to patient’s records and official documents in order to protect the legal interests of the patient, doctor and other health care professionals. |
| <p style="text-align: center;">Patients</p> | <ul style="list-style-type: none"> • Accelerate access to patient health information enabling clinicians to keep track of patients’ conditions over time. • Give patients treatment that is more effective and |

| | |
|--------------------|--|
| | better aligned to patients' needs. |
| Researchers | <ul style="list-style-type: none"> Analyse and interpret data to determine causes, prevention methods and treatment for diseases, injuries and disabilities. Develop data mining methods for advanced analytics. |

A demand for creating a National Health Association (NHA) is required in order to organize the flow of data and keep data privacy as it will not be one of the main beneficiary of the System, it will help in reviewing vital statistics and the incidence and prevalence of disease in a city, state or country.

The following Figure shows how flow of data organized between different stakeholders:



The NHA will consolidate the data sharing and manage the governances of the data privacy through classifying accessibility to database into four categories: no access, indirect access through the database owner or a third party, direct access restricted to specific datasets and direct access to the full dataset; this

will safeguard patients' data, and patients should be provided with clear information on how the data will be stored, used and protected.

3- Data Quality Control for iMhotep Health System:

A popular method of performance improvement used in the United States of America is the "Plan, Do, Check and Act" (PDCA) method developed by Walter Shewhart (Abdelhak et al., 1996). The steps are:

- The plan phase consists of data collection and analysis to propose a solution to a specific problem;
- The do or implementation phase tests the proposed solution;
- The check phase monitors the effectiveness of the solution over a period of time; and
- The act phase formalizes the changes that have proven effective in the "do" and "check" stages.

These steps will be essential to monitor and review data quality in order to enhance data consistency and effectiveness for the purpose of researches and analytics.

4- Data Mining Techniques

Data mining have a great potential to enable healthcare systems to use data more efficiently and effectively hence it improves care and reduces cost. The following are the classifications of data mining tools that can be applied on health care database:

1. *Tracking patterns:*

One of the most basic techniques in data mining is learning to recognize patterns in your data sets. This is usually a recognition of some deviation of your data happening at regular intervals. For example, you might see that your sales of a certain product seem to spike just before the holidays, or notice that specific disease is emerging in a specific area of a country.

2. *Classification:*

Is a more complex data mining technique that enable you to collect various attributes together into discernible categories, which you can then use to draw further conclusions, or serve some function. For example, if you're evaluating data on individual customers' financial backgrounds and purchase histories, you might be able to classify them as "low," "medium," or "high" credit risks. You could then use these classifications to learn even more about those customers.

3. *Association:*

Is related to tracking patterns, but is more specific to dependently linked variables. In this case, you'll look for specific events or attributes that are highly correlated with another event or attribute; for example, you might notice that when your customers buy a specific item, they also often buy a second, related item. This is usually what's used to populate "people also bought" sections of online stores.

4. *Outlier detection:*

In many cases, simply recognizing the overarching pattern can't give you a clear understanding of your data set. You also need to be able to identify anomalies, or outliers in your data. For example, if your purchasers are almost exclusively male, but during one strange week in July, there's a huge spike in female purchasers, you'll want to investigate the spike and see what drove it, so you can either replicate it or better understand your audience in the process.

5. *Clustering:*

Is very similar to classification, but involves grouping chunks of data together based on their similarities. For example, you might choose to cluster different demographics of your audience into different packets

based on how much disposable income they have, or how often they tend to shop at your store.

6. Regression:

Used primarily as a form of planning and modelling, is used to identify the likelihood of a certain variable, given the presence of other variables. For example, you could use it to project a certain price, based on other factors like availability, consumer demand, and competition. More specifically, regression's main focus is to help you uncover the exact relationship between two (or more) variables in a given data set.

7. Prediction:

Is one of the most valuable data mining techniques, since it is used to project the types of data you will see in the future. In many cases, just recognizing and understanding historical trends is enough to chart a somewhat accurate prediction of what will happen in the future. For example, you might review consumers' credit histories and past purchases to predict whether they will be a credit risk in the future.

The following two examples illustrate how **iMhoteb** electronic data management platform is applying the data mining tools and machine learning to ensure better utilizing of the data available:

Application 1: Heart disease prediction using artificial neural network:

Medical diagnosis is considered as a significant yet intricate task that needs to be carried out precisely and efficiently. Clinical decisions are often made based on doctor's intuition and experience rather than on the knowledge rich data hidden in the database. The automation of the medical diagnosis process, including prescriptions, would admittedly mitigate the human error and costs of the health insurance sector. This would have a radical impact on global public health eventually.

Heart disease is the leading cause of death, and experts estimate that approximately half of all heart attacks and strokes occur in people who have not been flagged as 'at risk.'

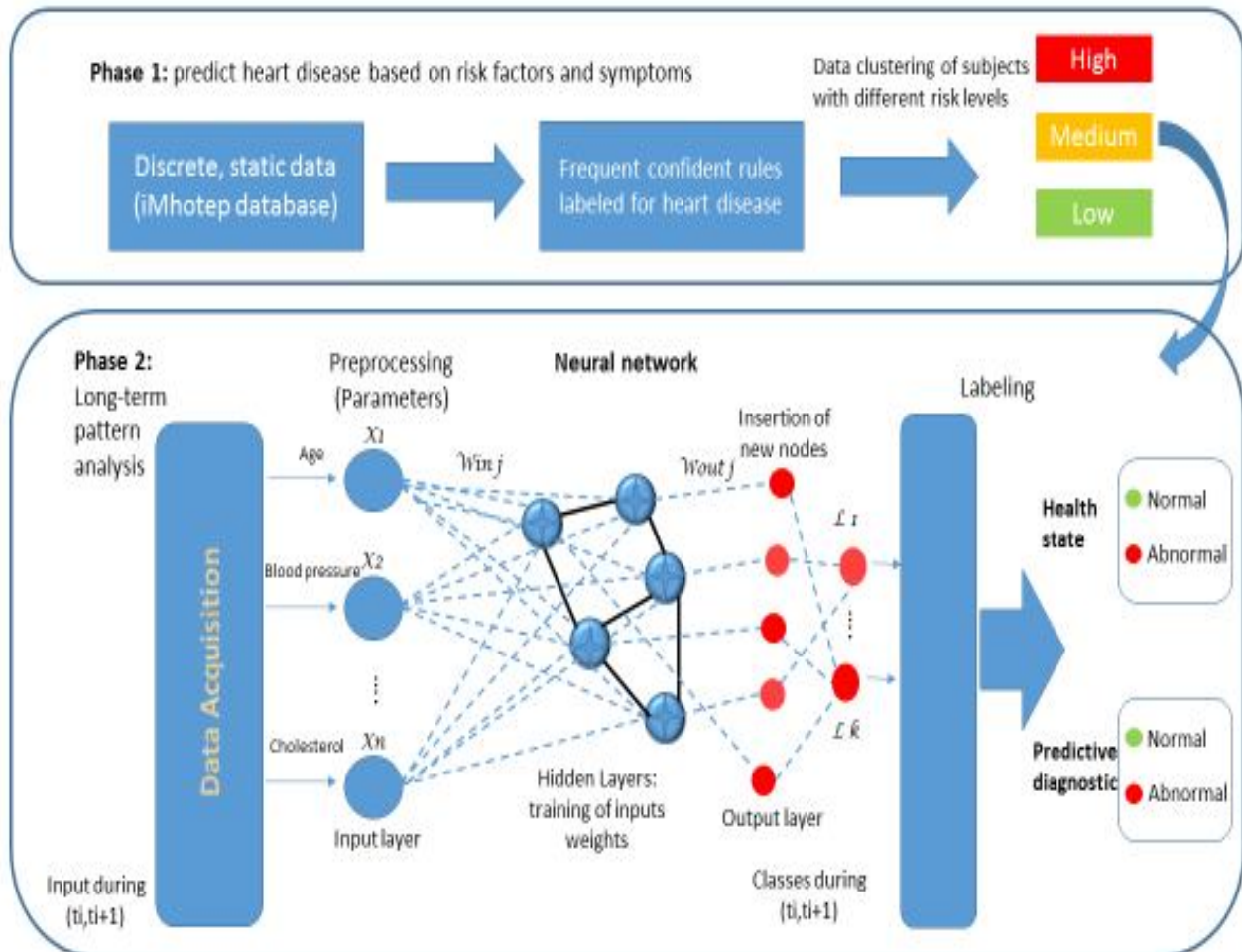


Figure 1: Heart disease prediction applying artificial neural network

Application 2: Fraud detection using data neural clustering:

Clustering can be used to find more fraud cases that are similar to known cases. Predictive can predict potential fraudulent claims and providers. However, it is difficult to obtain training historical data as most fraudulent claims may have not been discovered. Therefore, it is better to depend on large number of historical records.

Large networks, especially with small training datasets, can remember individual records. A better way is to decompose problems into multiple smaller neural networks with a small number of input and hidden layer nodes, each network predicting one type of frauds. Then integrate them with higher level neural networks or use rules to combine them. Supervised learning entails training an algorithm using labelled historical data. In this case, existing datasets already have target variables marked, and the goal of training is to make the system predict these variables in future data.

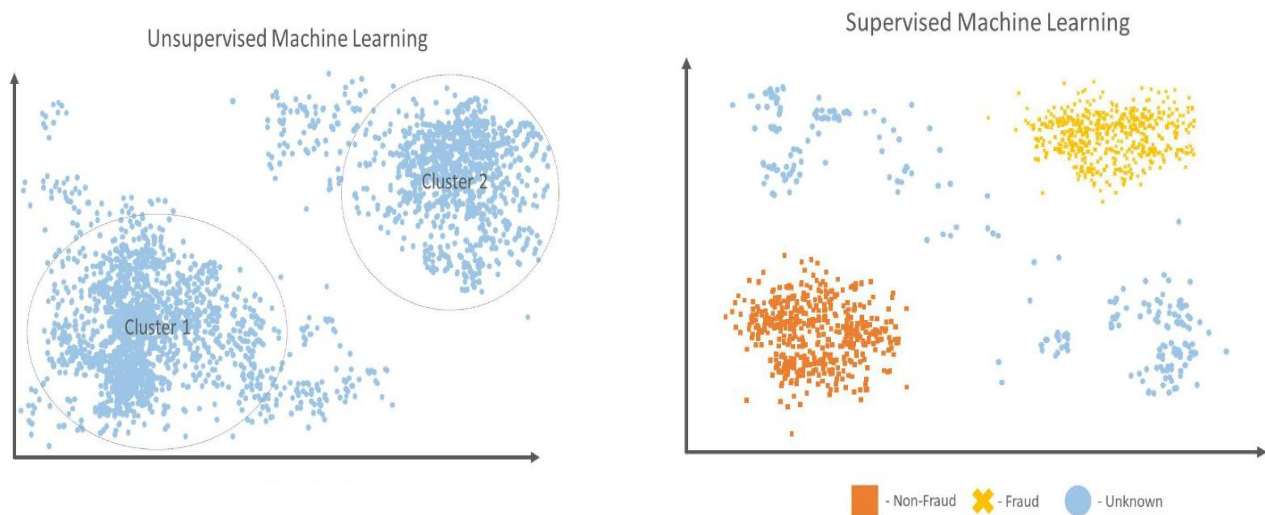


Figure 2: data clustering method used to detect heal care fraudulent activities

VI. Conclusions and Recommendations:

In conclusion, insurers and health service providers can take advantage of growth opportunities in data mining, service enhancement, and cost reduction. Also, **iMhotep** health system will give plenty of room for controlling fraud in the health insurance, increase medical awareness between patients, and open a new research area in big data sharing and embracing data ethics.

iMhotep health system and the National Health Association could be a guide to the government in developing its own healthcare policy and regulate funds. Also, it will interact with the main leading Information Technology entities.

iMhotep central database in medical/health records can be improved by ensuring the following:

- All patient identification data are accurately recorded on the first sheet of the medical/health record and the patient's name and medical/health record number are clearly shown on subsequent pages.
- The main condition and other diagnoses, problems and procedures are clearly written on the front sheet, along with the signature of the attending health care provider.
- The history of past and present illnesses/problems is recorded clearly, and the entry dated and signed.
- Consent forms are signed, dated and witnessed.
- Progress notes, whether for an inpatient or outpatient, are recorded daily or each time the doctor sees the patient and are clearly written, legible, signed and dated.
- For surgical patients, either as an inpatient or at a day surgery, operation forms should be completed with all relevant information, as well as anaesthetic forms and recovery room report, signed and dated.
- Nursing notes for inpatients should be completed daily, written clearly, and each entry dated and signed.

In a nutshell, the following are among the recommendations to be considered as a top priority in the agenda of the National Health Association and Health Insurer:

- Employ sufficient staff to meet the needs of the data collection area.
- Include continuous audit by independent parties under supervision of National Health Association.
- Involving volunteers, interns and part-time employees for the data entry role, to ensure cost and time saving.
- Conduct regular sample review on database to ensure data completeness and alignment with the data privacy policy.

- Engage in health awareness campaigns to promote awareness of people in the rural parts of the country and enhance the public health.
- Encourage the healthcare data scientists and actuaries to be more involved in the latest technological tools in data mining and artificial intelligence algorithms. This will enhance healthcare business analytics.

References:

- 1) Ball, J., Balogh, E. and Miller, B.T. eds., (2015). *Improving diagnosis in health care*. National Academies Press.
- 2) Boukenze, B., Mousannif, H. and Haqiq, A., (2016). *Predictive analytics in healthcare system using data mining techniques*. *Comput Sci Inf Technol*, 1, pp.1-9.
- 3) Copeland, B.; Raynor, M.; Shah, S., (2014) *Top 10 health care innovations: Achieving more for less*, Deloitte.
- 4) Donaldson, M.S. and Lohr, K.N., (1994). *Health data in the information age: use, disclosure, and privacy*. National Academies.
- 5) Hariharan, S. and Dey, P.K., (2010). *A comprehensive approach to quality management of intensive care services*. *International journal of health care quality assurance*, 23(3), pp.287-300.
- 6) Institute of Medicine (US). Committee on Assuring the Health of the Public in the 21st Century, (2003). *The Future of the Public's Health in the 21st Century*. National Academy Press.
- 7) Ionita, I. and Ionita, L., (2016). *Applying data mining techniques in healthcare*. *Studies in Informatics and Control*, 25(3), pp.385-394.
- 8) Pacurariu, A., Plueschke, K., McGettigan, P., Morales, D.R., Slattery, J., Vogl, D., Goedecke, T., Kurz, X. and Cave, A., (2018). *Electronic healthcare databases in Europe: descriptive analysis of characteristics and potential for use in medicines regulation*. *BMJ open*, 8(9).
- 9) Raghupathi, W. and Raghupathi, V., (2014). Big data analytics in healthcare: promise and potential. *Health information science and systems*, 2(1), p.3.
- 10) Spaan, E., Mathijssen, J., Tromp, N., McBain, F., Have, A.T. and Baltussen, R., (2012). *The impact of health insurance in Africa and Asia: a systematic review*. *Bulletin of the World Health Organization*, 90, pp.685-692.
- 11) Srinivasan, U. and Arunasalam, B., (2013). *Leveraging big data analytics to reduce healthcare costs*. *IT professional*, 15(6), pp.21-28.
- 12) Syed Danish Ali, (2016). *The great reach of big data in healthcare*. The magazine of the Institute & Faculty of Actuaries.
- 13) Tomar, D. and Agarwal, S., (2013). *A survey on Data Mining approaches for Healthcare*. *International Journal of Bio-Science and Bio-Technology*, 5(5), pp.241-266.

- 14) World Health Organization, (2003). *Improving data quality: a guide for developing countries*.
- 15) Organisation for Economic Co-operation and Development, (2017). *Technology and innovation in the insurance sector*.