Factors Analysis of a predict Dengue using Big Data Analytics – A Review

[1]Navpreet Kaur, [2]Dr. Sandeep Sharma
[1]Student, [2]Head of the department
Department of Computer Engineering and Technology
Guru Nanak Dev University, Amritsar, Punjab, India

Abstract: — Due to technology the term data is replaced by transforming big data in many fields. Rapidly advancements in the technology causes cultivated data enter into the era of big data. Analytical Data helps to find the various vectors of diseases (Dengue) outbreak in a particular area. Traditional tools and techniques are unable to store and analyze this massive amount of data. With the help of Big data analytic is easy to identify the particular required data from huge amount of data. To achieve these objective different tools has been used. The data is collected, cleansed and normalized. Cleansing of data is done that is important information is extracted from unstructured redundant data. The main goal of big data Analytics with dengue is easy to find factors that are causing for outbreaks the disease; we have studied and identified some variable such as weather, rainfall, temperature and humidity for the cause of this disease.

Index Terms: — **Big data, analytics, Dengue, Stages of dengue, Tools, Literature review.**

I. INTRODUCTION

Big Data Analytics largely involves collecting data from different sources, mange it in a way that it becomes available to be consumed by analysts and finally deliver data products useful in different fields. Big data analytics is a latest analytics standard which is used to examine a collection of data, which cannot be managed or processed with the presently accessible technologies. Big Data mining is used to extract significant and valuable information from the vast datasets. Big data analytics has emerged from two distinct concepts big data and analytics. Big data is huge amount of collection of data. Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making. It derives the need for technological infrastructure and tools that can capture, store, analyze and visualize huge amounts of structured and unstructured data. The process of converting large amounts of unstructured raw data, retrieved from different sources to a data product useful for organizations forms the core of Big Data Analytics. The potential of the Big Data shows the wide range of data can be translated into valuable information. Google web search query, social media also considered as important source of Big Data.
Data is about vast amounts of information. The volume of data that one has to deal has exploded to unimaginable levels in the past decade, and at the same time, the price of data storage has systematically reduced. Specifically, it focuses on information sets that are too large to handle in the usual manner. Private companies and research institutions capture terabytes of data about their users' interactions, business, social media, and also sensors from devices such as mobile phones and automobiles. The challenge of this era is to make sense of this sea of data. This is where big data analytics comes into picture. Big Data Analytics helps you to understand your field better. With the use of big data analytics, one can make the informed decisions without blindly relying on guesses. Discovering relations and understanding patterns within the data, big data analytics has the potential to perk up care, save lives at lesser costs. Thus, big data analytics applications in healthcare take advantage of the explosion in data to extract insight for making better informed decisions. Healthcare is a prime example of how the three Vs of data, velocity (speed of generation of data), variety, and volume, are an innate aspect of the data it produces. This data is spread among multiple healthcare systems, health insurers, researchers, government entities, and so forth. Furthermore, each of these data repositories is soloed and inherently incapable of providing a platform for global data transparency.

Analytics can be classified in to three types they are: Predictive Analytics, Descriptive Analytics and Prescriptive analytics.

a) **Descriptive analytics:** The simplest class of analytics, "one that allows you to condense big data into smaller, more useful nuggets of information".

b) **Predictive analytics:** It is the next step up in data reduction. It utilizes a variety of statistical, modeling, data mining, and machine learning techniques to study recent and historical data, thereby allowing analysts to make predictions about the future.

c) **Prescriptive analytics:** It is a type of predictive analytics. It's basically when we need to prescribe an action, so the business decision-maker can take this information and act.

II. DENGUE

Dengue is a mosquito-borne virus disease of humans. The disease is mainly concentrated in tropical and subtropical regions, putting, nearly a third of the human population, worldwide, at risk of infection. Dengue is a viral infection where human acts as the reservoir with any of four one serotype dengue virus. Infection are caused by any of four virus serotypes (DEN-1, DEN-2, DEN-3, DEN-4), it does not necessarily protect against a secondary infection with a heterologous serotype. Dengue is occurring in two forms:
(1) Dengue fever (DF) is caused by an arbovirus and spread by Aedes mosquitoes
(2) Dengue Hemorrhagic Fever (DHF).

Dengue fever (DF) is a critical issue for the most countries across the world. It is an acute, mosquito-transmitted viral of diseases characterized by fever, nausea, headache, arthralgia, malign and vomiting. Some infections results in Dengue Hemorrhagic fever and in its severe from Dengue shock syndrome (DSS) can threaten the patient's life primarily through increased vascular permeability and shock [1, 2]. Infection with serotypes results in varying degrees of pathological conditions, ranging from mild asymptomatic dengue fever to severe dengue hemorrhagic fever and dengue shock syndrome (DSS) which may turn fatal. Dengue has become one of the most widespread reemerging mosquito-borne diseases globally. There has been global increase in the frequency of DF, DHF and its epidemics, with a concomitant increase in disease incidence. The incidence of dengue is increasing in most tropical areas throughout the world [6, 7].
Dengue is one of the most important arthropod-borne viral diseases in terms of human morbidity and mortality. The percentage of serologically confirmed cases at the time of notification is relatively low due to lack of convalescent samples (second blood specimen) sent for confirmation. The incidence rate of dengue is highest among the working and school-going age groups [9].

The reason for this increased dengue virus activity is complex, but basically three factors are responsible:
(1) Lack of effective, long-term mosquito control in most tropical countries
(2) Increased urbanization in those same countries
(3) A marked increase in air travel which provided an ideal mechanism for transporting dengue viruses between tropical population centers.

All of which has created condition ensuring that dengue viruses will be introduced into areas suitable for epidemic transmission. Temporal changes in the environment are driven mostly by meteorological variable that determine the relationship between hosts, parasites and vectors. Variables linked to the human population such as density and socioeconomic status have also been considered important. In terms of number of individuals infected, it is by far the most divesting of all the recognized arthropod-transmitted virus diseases. It is estimated that more than 3 billion humans live in dengue endemic regions of the world and currently more than 50 million infections occur annually with at least 50,000 individuals requiring hospitalization of these, tens of thousands have a high risk of developing hemorrhagic diseases.
The disease is mainly concentrated in tropical and subtropical regions, putting nearly a third of the human population, worldwide, at risk of infection. A dramatic worldwide expansion of the DENV has occurred due to rapid urbanization, increase in international travel, lack of effective mosquito control measures, and globalization.

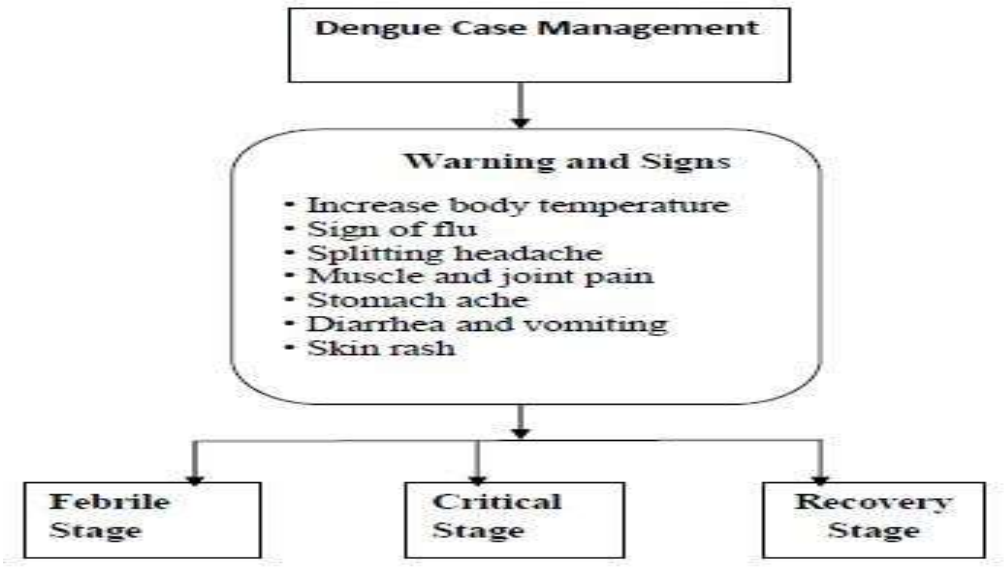

Fig 1: Dengue case management

III. STAGES OF DENGUE

Dengue is a systemic and dynamic disease and presented in three phases which are:

- Febrile
- Critical
- Recovery

Febrile Stage: During febrile phase an infected person suffers from higher fever shooting up to 104F and a pain in the head which lasts for 2 to 7 days. Mild hemorrhagic manifestations like positive tourniquet test or petechial and mucosal membrane bleeding may be seen in DF and Dengue Hemorrhagic Fever (DHF).

Critical Stage: In a critical stage, the condition of the patient worsens. The fever gets back to the normal temperature but, leakage of plasma from blood vessels begins and lasts for 1 to 2 days this may lead to accumulation of fluid in the chest and restricted flow of blood to all the important organs. People who each this stage of infection are at a higher risk of dying due to the disease. Varying circulatory disturbances can develop. In less severe case, patient recovers spontaneously, or after a short period of fluid or electrolyte therapy. In more severe forms of plasma leakage, the patient may sweat, have cool extremities and prolonged capillary refill time. The pulse rate and diastolic blood pressure increase, and the pulse pressure narrows.

Recovery Stage: The recovery stage show sudden improvement in the health of the patient and has symptoms like a decreased consciousness also occur in this stage. The stage of the infection may be different in order to avoid any complication it is best treat it at its initial stage.

IV. IMPACT OF BIG DATA ANALYTICS ON DENGUE

Big Data Analytics could be useful in improving health. It is also helping consume acquire more reliable and timely information about the cost and quality care. Data is an important Tool in developing new types of personalized healthcare:

a) **Right living:** Patients can build value by taking an active role in their own treatment, including disease prevention. The right-living pathway focuses on encouraging patients to make lifestyle choices that help them remain healthy, such as proper diet and exercise and take an active role in their own care if they become sick.

b) **Right care:** This pathway involves ensuring that patients get the most timely, appropriate treatment available. In addition to relying heavily on protocols, right care requires a coordinated approach across settings and providers, all caregivers should have the same information and work toward the same goal to avoid duplication of effort and suboptimal strategies.

c) **Right provider:** This pathway proposes that patients should always be treated by high-performing professionals that are best matched to the task and will achieve the best outcome. “Right provider” therefore has two meanings: the right match of provider skill set to the complexity of the assignment for instance, nurses or physicians’ assistants performing tasks that do not require a doctor but also the specific selection of the provider with the best proven outcomes.

d) **Right value:** To fulfill the goals of this pathway, providers and payers will continuously enhance healthcare value while preserving or improving its quality. This pathway could involve multiple measures for ensuring cost-effectiveness of care, such as tying provider reimbursement to patient outcomes, or eliminating fraud, waste, or abuse in the system.

e) **Right innovation:** This pathway involves the identification of new therapies and approaches to delivering care, across all aspects of the system, and improving the innovation engines themselves. They could also use the data to find opportunities to improve clinical trials and traditional treatment protocols, including those for births and inpatient surgeries.

ADVANTAGES OF BIG DATA ANALYTICS IN DENGUE:

The main benefits can be detecting diseases at earlier stages, detecting diseases abuse and fraud faster, and reducing costs.

- **Benefits to Patients**: Big data analytics can help patients make the right decision in a timely manner. From patient data, analytics can be applied to identify individuals that need “proactive care” or need change in their lifestyle to avoid health condition degradation. For example, patients in early stages of some diseases should be able to benefit from preventive care thanks to big data.

- **Benefits to Researchers and Developers (R & D)**: Collecting different data from different sources can help improving research about new diseases and therapies. R & D contribute to new algorithms and tools, such as the algorithms by Google, Face book, and Twitter that define what we find about our health system. Google R & D can also enhance predictive models to produce more devices and treatment for the market.

- **Benefits to healthcare Providers**: Providers may recognize high risk population and act appropriately (i.e. propose preventive acts).Therefore, they can enhance patient experience. Moreover, approximately 54% of US hospitals are members in local or regional Health-Information Exchanges (HIEs) or try to be in the future. These developments give the power to access a large array of information. For example, the HIE in Indiana connects currently 80 hospitals and possesses information of more than ten million patients.

V. TOOLS

There are many techniques being used to analyze datasets. Analytics are structured, or formalized, approaches to manipulating information. It covers activities like calculations, deriving new information, and documenting results, all with an eye to a particular theme. But more to the point, it does these things using a set of standardized tools. This has a couple of benefits:

a) The tools act as a guide for investigation. This is particularly useful in situations where you are unfamiliar with the information. Basic conclusions can be quickly drawn, which lead to more significant derivations.
b) The tool set is known and easy to understand. This gets you up-to-speed quickly with new information sets, and allows you to progress to the next level of investigation.
c) The results produced by the tools act as a baseline, and can be compared to external information and results. This, in turn, gives you confidence about your results, and points you to more complex activities.

Tools for Big data Analytics: There are various tools used in the big data namely Apache Hadoop, Apache Storm, Mongo DB, HPCC, R-Programming. It is used to improve the various factors in the development of big data and functionality of a computer system.

a) **Hadoop** is a project which is being developed by the Apache Software Foundation that supports huge data sets in a scattered location. This powerful system is known for its ease of use and its ability to process extremely large data in both, structured and unstructured formats, as well as replicating chunks of data to nodes and making it available on the local processing machine. This Hadoop Tool is a platform independent tool since it is developed in JAVA Framework which is used to process several applications or nodes at a same time with a speed ranging in terabytes. Apache has also introduced other technologies that accentuate Hadoop's capabilities such as Apache Cassandra, Apache Pig, and Apache Spark.

b) **Apache Storm** was originally developed by Nathan Marz and team at Back Type. It is a social analytics company. Apache Storm can be used with or without Hadoop, and is an open source distributed real time computation system. It makes it easier to process unbounded streams of data, especially for real-time processing. It is extremely simple and easy to use and can be configured with any programming language that the user is comfortable with. Storm is great for using in cases such as real time analytics, continuous computation, online machine learning, etc. Storm is scalable and fast, making it perfect for companies that want fast and efficient results. This is a very fast processing tool which is used to process million tuples per second per node.

c) **Mongo-DB** is also a great tool to help store and analyze big data, as well as help make applications. It was originally designed to support humongous databases, with its name Mongo-DB, actually derived from the word humongous. Mongo-DB is a good resource to manage data that is frequently changing or data that is semi-structured or unstructured. Most often, it is used to store data in mobile apps, product catalogs, real-time personalization, content management, and applications that deliver a single view across multiple systems. Mongo-DB is a no SQL database that is written in C++ with document-oriented storage, full index support, replication and high availability, etc.

d) **HPCC (High Performance Computing Cluster)** which is developed by LexisNexis Risk Solutions which is used to increase the performance factor of the System. This is a brilliant platform for manipulating, transforming, querying and data warehousing. It is used for parallel and batch based processing of application using big data. A great alternative to Hadoop, HPCC delivers superior performance, agility, and scalability. This technology has been used effectively in production environments longer than Hadoop, and offers features such as built-in distributed file system, scalability thousands of nodes, powerful development IDE, fault resilient, etc.

e) **R-Programming** is not just a software, but also a programming language. Project R is the software that has been designed as a data mining tool, while R programming language is a high-level statistical language that is used for analysis. An open source language and tool, Project R is written is R language and is widely used among data miners for developing statistical software and data analysis. In addition to data mining it provides statistical and graphical techniques, including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, and others.

RELATED WORK

Reference[1] In the research of disease (dengue) detection using big data analytics the various detection and diffusion model, the selection of researchers in different domain in order to distinguish parameters are (e.g. moisture, rainfall, map, transportation, trade, floating population, etc.) These parameters had been used as the key factors and be the base of research. People who live near to the mosquito breeding are likely to get the fever. It is a viral disease transmitted by the infective bite of female Aedes mosquito. Dengue fever and dengue hemorrhagic fever are emerging as major public health problems. But the proposed conceptual dengue active surveillance system framework introduced a new method of predetermine and early detection of dengue cases. This framework is combination of multiple secondary data from batch and real time data retrieval. A new approach to active surveillance outlined to develop an early warning surveillance system framework that can predict epidemic dengue to improve current passive surveillance system, the framework introduced data harvesting process from multiple sources as input, data pre-processing using data aggregator and filtering engine, storing large data in repository, analytic engine for analysis and processing the large data, and presentation of the information to the users. The data aggregator will aggregate the data from three different types of data such as structured, semi-structured and unstructured data to be stored into the semi-structured database such as MongoDB and NoSQL. The large data will be processed and analyzed using algorithm or mathematical calculations to determine the expected dengue cases. Then, the processed information will be presented to the users in a form of web or mobile application. This framework may become basis and foundation of this research to align and support with the existing passive surveillance system. The Outcomes of this framework are to become alternative method for early detection of dengue cases. This framework utilizing data in action from social media and weather information provider that available online in real time to replace the data collection from passive system using the result of laboratories and clinical test collected from hospital or health practitioners' reports. The quality and accuracy of the system depends on the machine learning analytics and features selection. System is useful for the authority to be alert with the situation that uncontrolled in future due to widespread of pandemic with early detection and surveillance system [1].

For Dengue Fever, traditional vector control is often conducted after the occurrence of dengue cases. Analytical model of dengue vector acted as an early warning system is crucial for prediction of dengue outbreaks. with the analytical model to see how changes in some specific variables such as rainfall, temperature, and humidity can dramatically affect the population of mosquito vectors, in order to provide early warnings of dengue outbreaks. We collected online data sensing, then combined the historical big data as training datasets for analytical computations. Amount of dataset can help improve the accuracy of forecasts, but if we utilize more detailed and real time information provided by social sensors, such as percentage of mosquitoes, pounding of road for decision making. The accuracy of warning can be effectively improved [2].

Data mining techniques are widely used to their capability to represent population and real time factor. Data mining is a process of identifying underlying patterns for seemingly uninformative data. It is observed that the propagation of the Influenza like diseases heavily depend on human interactions, where an infected human travels to a vulnerable area and mosquitoes of that area will bite him, and contract and spread the virus. It is observed that the Human mobility patterns can be recreated similar to infection patterns which are encountered in Dengue endemics. The potential of incorporating human mobility, derived through mobile network data in predicting Dengue propagation [3].

Based on the analysis, climate variables have a strong relationship to the dengue incidences. Machine learning techniques such as clustering and regression is done to compare the sum square of residual (SSE) to conclude which climate variable is giving a big impact on dengue cases. Using Dual climate shows how the variables affect the dengue incidence. These climate variables are clustered to seek the clear pattern which correlates with dengue occurrences. The data are clustered using K-means algorithm, and regression model is built for each cluster. Weather variables are independent variables and dengue incidence as the dependent variable. Averaged silhouette width method is used to define the number of K group [4].

Higher temperature and higher humidity makes the dengue transmission at peak level. It is started that an increase in temperature will lead to increase of dengue incidences. The association between dengue incidence and weather factors also apparently varies by locality, suggesting that a prospective dengue early warning system would likely be best implemented at a local or regional scale, rather than nation-wide. Such spatial down-scaling would also enable dengue control measures to be better targeted, timed and implemented. The knowledge gained from the current study is also potentially applicable to neighboring countries, which share many of Cambodia's weather, environmental conditions and social conditions. [7, 8].

Regression method is used for mathematical model and also for future predictions purposes. High accuracy can be obtained by using a regression method with weather data as independent variables. Regression model has always ended up in low accuracy model, as it is not updated with latest transformation [9]
The internet has become global communication network that allows connect and exchange information worldwide. In 2009, dengue monitoring system based on Google web search queries has been developed by researchers in Google. They were analyzing web query data using five methods of analysis to get a prediction system of influenza and dengue active system based on their product called Google Trends. [12, 13].

Table 1: Data Types, Sources and Factors in Dengue Fever Outbreak Prediction

Title of Paper	Data Type	Data Sources	Factors	Technique
Proposed Conceptual Framework of Dengue Active Surveillance System in Malaysia. May 2016 IEEE	Structured, Semi-structured & unstructured	Internet Online, and Real data	Weather or flood information, Social media Batch processing and data aggregator	Framework
Incorporating Big Data and Social Sensors in a Novel Early Warning System of Dengue Outbreak. 2015 IEEE	Structured & unstructured	Open data Social sensors	Rainfall Temperature Humidity Population density	System Framework
Cluster Based Regression Model on Dengue Incidence Using Dual Climate Variables. December 2016 /IEEE	Structured & unstructured	Open data	Rainfall and Humidity Temperature and Humidity variable Rainfall and Temperature variable	Clustering and regression technique
Dengue Outbreaks Using Local Weather And Regional Climate for a Tropical Environment in Colombia. 2014	Structured and unstructured	Epidemiological data	Rainfall Temperature Weather Humidity	Intraseasonal prediction
Effects of Weather Factors on Dengue Fever Incidence and Implication for interventions in Cambodia, 2016	Structured and Unstructured	Statistical data	Rainfall Temperature Humidity Weather factors	Framework
Dengue disease mapping in Malaysia based on Stochastic SIR Models in Human Populations	Structured and Unstructured data	Statistical data Health and Medical Research data	Rainfall Temperature Humidity	Stochastic SIR model
Dengue Propagation Prediction using Human Mobility, 2016 IEEE	Structured and Unstructured data	Past Dengue cases data Population data Mobile Network data	Rainfall Temperature Humidity	Mobility

VI. CONCLUSION

In this review paper based on analysis information and gathered sign from real-world environment estimating that main key factors of dengue outbreaks is weather, rainfall, temperature, humidity and population density. Temperature is often used because of its effects on biological parameters such as the extrinsic incubation period of mosquito. Humidity and optimal rainfall, it leads to more dengue cases. Because the mosquito vector requires water for completion of its life cycle, previous investigators had examined the use of rainfall data for disease outbreak prediction. But With the help of different framework, prediction models, social sensors and others it is easily to get the information and warn where the dengue actual outbreaks and can control with using different predictions.

REFERENCES

[1] MohdKhalit Othman and MohdShahrul NizamMohdDanuri, "Proposed Conceptual Framework of Dengue Active Surveillance System (DASS) in Malaysia," (ICICTM) 16- 17 May 2016 IEEE
[2] Chung-Hong, Hsin-Chang, Shih, "Incorporating Big Data and Social Sensors in a Novel Early Warning System of Dengue Outbreak," 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mminning.
[3] M.P.A.R.Abeyrathna, D.A.Abeygunawrdane, DanajaMaldeniya, KaushalyaMadhawa, "Dengue Propagation Prediction using Human Mobility 2016 IEEE.
[4] Shermon S. MAthulamuthu1, Vijanth S. Asirvadam1, Sarat C. Dass2, Balvinder S. Gill3, "Cluster Based Regression Mmodel on Dengue Incidence Using Dual Climate Variables," 2016 IEEE Conference on System, Process and Control, 16-18 December 2016, Malaysia.
[5] VibhaGanjir 1, Dr. B.K Sarkar 2, and Ravi Ranjan Kumar 3, "Big Data Analytics for Healthcare" International Research in Enginerring, Technology and Science, July 2016
[6] A. Chakravarti, R.Arora, and C. Luxemburger, "Fifty years of dengue in India," Trans. R. Soc. Trop. Med. Hyg, vol. 106, no. 5, pp.273-282z, 2012.
[7] Y. choi, C. S. Tang, L. Melver, M. Hashizume, V. Chan, R. R. Abeyasinghe, S. iddings and R. Huy, "Effects of Weather Factors on Dengue Fever Incidence and Implication for Interventions in Cambodia", BMC public Health20016, 16; 241
[8] S. Polwiang, "The correction of climate factors on dengue transmission in urban area: Bangkok and Singapore cases," 27july 2016
[9] R.Chandran and P.A.Azeez, "Outbreak of Dengue in Tamil Nadu, India", current science, vol. 109, no. 1, 10 July 2015.
[10] O.J. Brady, D.L. Smith, T.W. Scott, and S. I. Hay, "Dengue disease outbreak definitions are implicitly variable," Epidemics, vol. 11,pp. 92-102, 2015.
[11] S. Chinikar, S.M. ghiasi, N. Shah-Hosseini, E. Mostafavi, M. Moradi, S. Khakifirouz, F.S. RasiVarai, M.Zainali, and A.R. Fooks, "Preliminary study of dengue virus infection in Iran," Travel Med. Infect. Dis., vol.11, no. 3, pp. 166-169, 2013
[12] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M.S. Smolinski, and L. Brillant , "Detecting influenza epidemic using search engine query data," Nature, vol. 457 , no. 7232,pp. 2009
[13] E. H. Chan, V. Sahai, C. Conrad, and J.S. Brownstein, "Using we search query data to monitor dengue epidemic: A new model for neglected tropical disease surveillance," PLoSNegl. Trop. Dis, vol. 5, no. 5, 2011
[14] Vikas Yadav, Monica Verma and Vandana Dixit Kaushik, "Big Data Analytics for Health System," 2015 IEEE
[15] Nor AzahSamat and David F. Percy, "Dengue Disease Mapping in Malaysia based on Stochastic SIR Models in Human Populations".

[1] Ali, A. 2001.Macroeconomic variables as common pervasive risk factors and the empirical content of the Arbitrage Pricing Theory. Journal of Empirical finance, 5(3): 221–240.
[2] Basu, S. 1997. The Investment Performance of Common Stocks in Relation to their Price to Earnings Ratio: A Test of the Efficient Markets Hypothesis. Journal of Finance, 33(3): 663-682.
[3] Bhatti, U. and Hanif. M. 2010. Validity of Capital Assets Pricing Model.Evidence from KSE-Pakistan.European Journal of Economics, Finance and Administrative Science, 3 (20).

s