A·P·U
ASIA PACIFIC UNIVERSITY
OF TECHNOLOGY & INNOVATION

JATI

Faculty of Computing, Engineering & Technology

JOURNAL OF APPLIED TECHNOLOGY AND INNOVATION

Volume 1, Issue 2
October 2017

eISSN 2600-7304

Journal of Applied Technology and Innovation

(E-ISSN: 2600-7304)

Volume 1, Issue 2 (2017)

Journal of Applied Technology and Innovation (JATI) is an electronic, open access, peer-reviewed journal that publishes original articles on novel theories, methods and applications in the field of electrical, electronics, mechatronics, telecommunication, computer, information technology, and software engineering. JATI e-journal reviews articles approximately in four (4) weeks period and publishes accepted articles, upon receiving the final versions on the forthcoming issue. It publishes 4 issues per year.

Journal of Applied Technology and Innovation (eISSN: 2600-7304)
vol. 1, no. 2, (2017), pp. 1-9

Determining Economic Growth with Trade Flow Analytics

Kalongo Hamusonde
Faculty of Computing, Engineering & Technology
Asia Pacific University of Technology & Innovation
57000 Kuala Lumpur, Malaysia
Email: kalongoh@gmail.com

Nirase Fathima Abubacker
Faculty of Computing, Engineering & Technology
Asia Pacific University of Technology & Innovation
57000 Kuala Lumpur, Malaysia
Email: dr.nirase.fathima@apu.edu.my

Abstract - This paper seeks to examine studies that have been conducted mainly in the period of 2015 to 2017 with regards to the analysis of trade flows in understanding their relationship with economic growth. Most studies examined causality relationships and employed co-integration tests and the Augmented Dickney-Fuller (ADF) model. The studies provided evidence that a relationship existed between trade flows and economic growth and this relationship was either bidirectional or unidirectional and existed both in the short-run and long-run depending on other factors other than trade flows. Other studies sought to compare statistical methods to determine which model produced statistically significant results in explaining and predicting trade flows with regards to economic growth.

Index Terms – Trade flow analytics, Economic growth, Gross Domestic Product (GDP)

1. Introduction

Of the many indicators needed for a country to achieve potent economic growth, trade flows are one of those indicators that cannot be ignored. Trade flows refer to the buying and selling of goods from one country to another. Trade flows measure the balance of trade which is simply the exports minus the imports of goods and services. Trade flows play a major role in in influencing the growth of a country. If suppose a country is a major exporter of certain goods, the demand for that country's currency hence increasing the currency's value and in turn a rise in economic growth. With the importance of trade flows been put to light, this paper seeks to investigate studies that have been conducted on the analysis of trade flows i.e. exports and imsport. Quite a substantial amount of studies has been done with regards to analytics on trade flows in

the region and most of these studies have used traditional statistical methods in their analysis. These studies have analysed exports or imports separately and sometimes collectively. The sole purpose of these analysis has been to determine the relationship between trade flows and at times Gross Domestic Product (GDP) with regards economic growth.

2. Materials, Methods and Discussion

With trade being one of the important indicators of economic growth as discussed above, many studies have devoted time to determining the relationship that exists between trade flows and economic growth. Tapsin (2015) conducted a study to ascertain what kind of relationship existed between foreign trade and economic growth in Turkey. To achieve this objective, the Augmented Dickney-Fuller (ADF) and unit root test were performed. The results showed that export and GDP had a bidirectional causality relationship while imports and GDP had a unidirectional causality relationship. The study finally concluded that imports and exports are important factors to consider when one seeks to measure economic growth in Turkey because the variables share a positive and significant relationship. Similarly, Ucan et al. (2016) conducted a study to determine the relationship between exports and economic growth in Turkey. The results from this study were similar to the ones found by Tapsin (2015) though the only notable difference was relationship between exports and GDP had a unidirectional causality relationship. The two results were not consistent because Tapsin (2015) had not considered other factors in the model like exchange rates which could affect trade flows but were considered in the study by Ucan et al. (2016).

Saaed & Hussain (2015) main hypothesis was to test for causality and co-integration between GDP, exports and imports in Tunisia. To test this hypothesis, Tunis annual data from the year 1977 to 2012 with GDP, imports and exports as the attributes was used. The Vector Error Correction Model (VECM) framework which employs the Granger-causality test was used. The study results showed that imports, exports and GDP have a unidirectional causality with imports being the major source of economic growth in Tunisia. The results further showed that imports increased economic growth in the long-run whilst exports did not. The models or techniques used for analysis in this study were traditional methods which have their own drawbacks. Like many traditional statistical analytics, the methods used in this study employed a lot of assumptions. The ADF procedure which was used to test for integration assumes that the error term of Autoregressive (p) process to be white noise which is a less strong assumption than Gaussian noise stationarity. Running the procedure under this assumption is erroneous and leads to less reliable results. Another study conducted in India by Mehta (2015) also primarily focused on testing for causality and co-integration among exports, imports and GDP. Here, time series data from the year 1976 to 2014 was used and the ADF procedure was implemented to test for causality while the VECM was used to test for co-integration.

The results showed that for the ADF to be implemented, the data had to be differenced at least once to attain stationarity. Once the data was stationary, further tests were conducted and it was found that long term GDP lead to increased exports but exports did not lead to a better GDP. It was also found that no causality exists between GDP and imports which implied that GDP does not lead to imports and imports do not lead to a greater GDP.

Similarly, Albiman & Suleiman (2016) conducted a study to determine if a relationship existed among exports, imports and domestic investment. This study used annual time series data from the year 1967 to 2010 for Malaysia and the VECM framework was used to test for causality and co-integration. Unlike the two studies discussed above, Saeed et al. (2015) & Mehta (2015), this study introduced capital formation into the analysis and the results differed in such a way that capital formation affected exports in the long-run while it affected imports in the short-run in determining the levels of domestic growth in Malaysia. Another study conducted by Altaee et al. (2016) to determine the effects of trade flows on economic growth in the Kingdom of Saudi Arabia showed that a fixed capital formation affected imports and exports in both the short-run and long-run. On the other hand, the study results also showed that financial development affected economic growth negatively in the short-run while turning out to have a positive impact in the long-run.

Bakari & Mabrouki (2017) sought to investigate the relationship between exports, imports and economic growth in Panama. Data ranging from 1980 to 2015 was tested for co-integration using the Vector Auto Regressive (VAR) Model and for causality using the Granger-Causality tests. A few adjustments were made to the analysis like the introduction of a second test called the Phillips Perron (PP) test to validate the results produced by the Granger-Causality tests. Another alteration was the introduction of the augmented function which was the aggregate production function of exports and imports expressed in logarithms. The study results from the VAR model showed that there was an absence of co-integration among exports and economic growth in Panama. This indicated that exports had no effect on the economic growth of Panama. On the other hand, a bidirectional causality existed among the variables imports and economic growth, hence imports were seen as a source of economic growth in Panama. With the same model and techniques used in the analysis of Panama trade flows conducted by Bakari & Mabrouki (2017), this time, Bakari (2017), used Germany annual data ranging from 1985 to 2015. In this study, one alteration was made, and this was the differencing of the export and imports data by using logarithms in order to make the model stationary before testing for causality and co-integration. The results were similar to the ones found in the Panama study though the only difference was that this study found a unidirectional causality between exports and imports. These results still provided evidence that exports and imports where a source of economic growth in Germany. A unidirectional causality between exports and imports was also found in another study and the results proved that

exports and imports were a major source of economic growth in Pakistan, Raza & Ying (2017).

Bakari (2016), again conducted an empirical study using Egypt annual data ranging from 1965 to 2015 to ascertain the relationship between exports, imports and economic growth. He used the VAR to test for co-integration and the Granger-Causality tests to test for causality. In this study, the variable domestic investment was introduced in the analysis to test against each of the three variables export, import and economic growth. The results showed that the introduction of domestic investment in the analysis significantly affected each variable differently. The co-integration results provided evidence that domestic investment, import and exports had no effect on economic growth while the causality results indicated that imports and domestic investment had a significant effect on the economic growth in Egypt. With Libya being one of the major exporters of petroleum in the world, Abdulhakim & Tarek (2016) conducted a study to determine the relationship between Foreign Direct Investment (FDI) and economic growth in Libya. This study focused solely on how foreign investments affect petroleum exports in relation to economic growth. Note that imports were not considered in this study. Annual export data ranging from 1992 to 2010 was gathered and the VAR model was used in the testing of this data to achieve the desired objectives of the study. Co-integration and causality between the variables was tested in this study. The results indicated that a long-term relationship existed between FDI and petroleum exports. The other results also showed that a unidirectional causality existed between FDI and petroleum exports. Though, it should be noted that one of the results provided evidence that there was no significant relationship between FDI and economic growth. This result was different from one of the deliverables of this study for it was expected that a significant relationship would exist between FDI and economic growth. One could assume that this output was a result of the researchers not controlling or putting into consideration other factors that could affect FDI like tax rate, wage rate, exchange rate, political situation to mention but a few.

In addition to analysis on exports, Verter & Becvarova (2016) conducted an analysis to determine the impact of agricultural exports on economic growth in Nigeria. Granger Causality, Odinary Least Square (OLS) regression, Impulse Response Function (IRF) and Variance Decomposition analysis were the statistical techniques employed in this study whilst considering annual time series data ranging from 1980 to 2012. Since the data was a time series data, ADF and unit tests were additional tests which were run to make the data stationary and this was achieved after the first difference. The results from the IRF where not constant thus showing a fluctuation of an upward and backward shock of agriculture exports in relation to economic growth. A shock to agricultural exports affecting economic growth was also seen in the results from the Variance Decomposition analysis. Both results from the OLS and the Granger Causality tests provided evidence that agricultural exports did lead to economic growth in Nigeria. Nevertheless, these results contradicted to the previous works done by Ojide et al. (2014) who conducted a

study to evaluate the impact on non-oil exports on economic growth in Nigeria. By performing the Autoregressive Distributed Lag (ARDL) model and co-integration tests on annual time series data ranging from 1970 to 2011, the study found that an inverse relationship existed between non-oil exports and economic growth. This difference between the results of the two studies can be attributed to the trade deficit in agricultural products which Nigeria which has incurred for the past six years. At the time Ojide et al. (2014) conducted the study, Nigeria Imported more goods than it exported hence the value for Nigerian currency was not high and this affected its exchange rate and in return resulting in an inverse relationship between non-oil exports and economic growth.

With exports playing an important role in influencing Uganda's economic growth, Karamuriro & Karukuza (2015) conducted a study to determine if the influence on economic growth by exports was statistically significant using the Gravity Model analysis. Since Uganda is a member state in the international organisations Common Market for Eastern and Southern Africa (COMESA) and East African Community (EAC), the two organisations were added to the model as dummy variables to determine if a significant relationship existed between Uganda's exports and its affiliation to these organisations. Other variables included in the model were exchange rate and common language. Annual panel data ranging from 1980 to 2012 was used. Since the data was of time series nature, the ADF test had to be conducted first in order to test for stationarity before any other analysis could be done. The data was found to be stationary after the second difference. The results of this study showed that exports had a statistically significant effect on economic growth. The variable common language as well showed a statistically significant effect on the exports. This meant that exporting goods from Swahili speaking countries like Kenya and Tanzania had a major effect on Uganda's economy. The results further showed that Uganda's affiliation to the two international organisations had a significant positive effect on the exports because the two organisations provided intra-trade among member states which came with prized benefits such as reduced tax rates. Another study to determine Egypt's intra-trade intensity with COMESA member states was conducted by Elmorsy (2015). The study objective was to determine which variables play a major role in promoting trade in Egypt with relation to COMESA member states. To achieve the objective, the Gravity Model analysis was employed and data from 2005 to 2011 regarding COMESA total imports, Egypt's total exports and world total exports was used. The results gave evidence that intra-trade agreement among COMESA member states contributed to the economic growth in Egypt. Nevertheless, a few obstacles such as political, social and infra-structural issues were also identified as potential threats that would affect trade flows negatively in the long-run. Alkhateeb et al. (2016) further conducted a study on Egypt's intra-trade agreements and their effects on agricultural trade flows. The study results were similar to the ones achieved by Elmorsy (2015) but added that population size and exchange rate also had a significant contribution on economic growth in Egypt.

Additionally, Abedini & Darabi (2015) conducted a study in which they examined the relationship between GDP, exports, imports, inflation and insurance and how all these variables affect economic growth. This study used annual data ranging from 2003 to 20111 and focused only on Organization of Petroleum Exporting Countries (OPEC) member states. The regression analysis was the statistical technique used in achieving the desired objective of this study. The results indicated that all the variables in the study had a positive and significant relationship with each other and thus implied that whsen productivity increases, insurance would increase too and that lead to a rise in economic growth. Similarly, Belke et al. (2014) sought to determine the relationship between exports and domestic demand in six European countries which are part of the European Union. A smoothing transition regression model was the main statistical technique used but other models were also employed like the sunk-cost model used in capturing non-linear hysteresis dynamics and the unit root tests to test for stationarity. The data was sourced from national statistical offices of each of the six countries in the study ranging from 1980 to 2012. The results from the smooth transition regression provided evidence that domestic demand is a necessity in the short-run to improving exports. The results further indicated that the non-linear relationship between domestic trade and exports was extreme during some stages of the business cycle. This indicated that paying sunk costs for shifting sales would give rise to the performance of the export market in the long-run and would still remain high even in an economic turmoil.

Not all trade analysis studies have been based on determining the relationship between trade flows and economic growth. Some of the studies have been based on analysing and comparing models to determine which one is better at achieving the desired objectives or testing how effective a statistical technique can perform at analysing trade flows. Milad et al. (2015) conducted a study to determine which model was better in forecasting Malaysia's imports of crude material by comparing two composite models. The first composite model was one with regression processing whilst the second model was one without regression processing. Data ranging from 1991 to 2013 was used and unit root tests were performed on the dataset to make it stationary. The results indicated that the model with regression processing outperformed the model without regression processing. The model with regression processing was able to reduce forecasting errors better than the other model. This is because the model with a regression process had a lower percentage of U-statistics which meant that it provided a best fit with highly significant p-values than the other model. Additionally, Kompas & Che (2016) conducted a study to determine if the structural and stochastic optimal model was significant in forecasting imports and exports of liquefied natural gas (LNG) in Asia-Pacific. The results from this study indicated that LNG exports in Asia-Pacific would increase by 90% in a period of 15 years starting from 2015 onwards. The results also projected an increase in LNG demand that would in turn increase imports of LNG within Asia-Pacific. Therefore, the study provided evidence that the structural and stochastic optimal model was reliable in forecasting trade flows. Similarly, Mladenovic et al. (2016)

conducted a study to determine which Artificial Neural Network (ANN) algorithm between Extreme Learning Machine (ELM) and the Back Propagation (BP) was more accurate in predicting GDP based on trade flows. Data from 28 European countries acquired from EUROSTAT was used in this study. The results showed that the ELM algorithm outperformed the BP in prediction accuracy and it had very small number of underestimated values as compared to the BP algorithm.

3. Conclusions

The studies have shown that trade flows can be used in determining economic growth in the short-run and long-run. When it comes to modelling, causality and stationarity tests should be the very first tests one should undertake when analysing trade flows for the data is of time series nature. This helps in the reduction of errors and provides very reliable results. It is very important to also take note that other variables like exchange rate, domestic investments, sunk costs etc. should be considered when modelling for they tend to affect trade flows differently hence omitting them would lead to erroneous results that would not be generalizable. It was also noted from some of the studies that affiliating to international organisations proves helpful for the intra-trade agreements tend to improve economic growth for affiliated countries by providing certain incentives like reduced tax rates. It can be further noted that not many works have been done where big data analytics and the use of machine learning algorithms is incorporated. Future studies would do well to venture in that direction in order to describe, predict or prescribe trade flows with a much greater accuracy and minimised errors.

References

Abdulhakim, A. A. & Tarek, Z. (2016) The effects of foreign direct investment on economic growth in Libya: A causality analysis. Open Science Journal. [Online] 1 (2). P. 1-15. Available from: https://osjournal.org/ojs/index.php/OSJ/article/view/62 [Accessed: 29 March 2017].

Abedini, R. & Darabi, R. (2015) The effect of private investment, exports, imports, inflation and GDP on per capita premium: Evidence from members of OPEC countries. *Management Science Letters.* [Online] 5 (1). P. 657-662. Available from: http://www.growingscience.com/msl/Vol5/msl_2015_63.pdf [Accessed: 2 April 2017].

Albiman, M. & Suleiman, N. (2016) The relationship among export, import, capital formation and economic growth in Malaysia. *Journal of Global Economics.* [Online] 4 (2). p. 1-6. Available from: http://dx.doi.org/10.4172/2375-4389.1000186 [Accessed: 2 April 2017].

Alkhateeb, Y. T. T., Mahmood, H. & Maalel, N. (2016) Egyptian intra agriculture trade with Common Market for Eastern and Southern Africa trading partners: A gravity model. *International Journal of Economics and Financial Issues.* [Online] 6 (6). p. 177-182.

Available from: https://www.econjournals.com/index.php/ijefi/article/download/4145/pdf [Accessed: 29 March 2017].

Altaee, A. H. H., Al-Jafari, K. M. & Khalid, A. M. (2016) Determinants of economic growth in the Kingdom of Saudi Arabia: An application of autoregressive distributed lag model. *Applied Economics and Finance*. [Online] 3 (1). p. 83-92. Available from: http://redfame.com/journal/index.php/aef/article/view/1200 [Accessed: 29March 2017].

Bakari, S. (2016) The relationship between export, import domestic investment and economic growth in Egypt: Empirical analysis. *Munich Personal RePEc Archive.* [Online] Available from: https://mpra.ub.uni-muenchen.de/id/eprint/76627 [Accessed: 2 April 2017].

Bakari, S. (2017) Trade and economic growth in Germany. *Munich Personal RePEc Archive.* [Online] Available from: https://mpra.ub.uni-muenchen.de/id/eprint/77404 [Accessed: 2 April 2017].

Bakari, S. & Mabrouki, M. (2017) Impact of exports and imports on economic growth: New evidence from Panama. *Journal of Smart Economic Growth.* [Online] 2 (1). p. 67-79. Available from: http://jseg.ro/ojs/index.php/jseg/article/download/18/pdf [Accessed: 29 March 2017].

Belke, A., Oeking, A. & Setzer, R. (2014) Exports and capacity constraints: A smooth transition regression model for six Euro area countries. *Working Paper Series.* Available from: https://www.ceps.eu/system/files/WD395%20Belke%20et%20al%20Exports%20and%20Capacity%20Constraints.pdf [Accessed: 2 April 2017].

Elmorsy, S. (2015) Determinants of trade intensity of Egypt with COMESA countries. *Journal of the Global South.* [Online] 2 (5). p. 1-25. Available from: http://dx.doi.org/10.1186/s40728-014-0002 [Accessed: 30 March 2017].

Karamuriro, T. H. & Karukuza, W. (2015) Determinants of Uganda's export performance: A gravity model analysis. *International Journal of Business and Economic Research.* [Online] 4 (2). p. 45-54. Available from: http://www.sciencepublishinggroup.com/journal/paperinfo?journalid=178&doi=10.11648/j.ijber.20150402.14 [Accessed: 30 March 2017].

Kompas, T. & Che, N. T. (2016) A structural and stochastic optimal model for projections of LNG imports and exports in Asia-Pacific. *Heliyon.* [Online] 2 (6). p. 1-35. Available from: http://www.sciencedirect.com/science/article/pii/S2405844015304837 [Accessed: 2 April 2017].

Mehta, N. S. (2015) The dynamics of relationship between exports, imports and economic growth in India. *International Journal of Research in Humanities & Social Sciences.* [Online] 3 (7). p. 39-47. Available from: http://raijmr.com/wp-content/uploads/2015/11/9_39-47-Dr.-Sachin-N.-Mehta.pdf [Accessed: 29 March 2017].

Milad, A. H. M., Ibrahim, I. R. & Marappan, S. (2016) A comparison among two composite models (without regression processing) and (with regression processing), applied on Malaysian imports. *Applied Mathematical Sciences.* [Online] 9 (116). p. 5757-5767. Available from: http://dx.doi.org/10.12988/ams.2015.58528 [Accessed: 2 April 2017].

Mladenovic, S. S., Milovancevic, M., Mladenovic, I. & Alizamir, M. (2016) Economic growth forecasting by artificial neural network with extreme learning machine based on trade, import and export parameters. *Computers in Human Behaviour.* [Online] 65 (2016). p. 43-45. Available from: http://doi.org/10.1016/j.chb.2016.08.014 [Accessed: 29 March 2017].

Ojide, G. M., Ojide, C. K. & Ogbodo, C. J. (2014) Export-led growth hypothesis in Nigeria: Applications of ARDL model and co-integration analysis. *Global Journal of Emerging Market Economies.* [Online] 6 (1). p. 5-13. Available from: http://journals.sagepub.com/doi/10.1177/0974910113511190 [Accessed: 3 April 2017].

Raza, M. & Ying, X. Z. (2017) The causal relationship between export and economic growth of Pakistan. *International Journal of Economics, Commerce and Management.* [Online] 5 (2). p. 210-231. Available from: http://ijecm.co.uk/wp-content/uploads/2017/02/5213.pdf [Accessed: 30 March 2017].

Saaed, A. A. & Hussain, A. M. (2015) Impact of exports and imports on economic growth: Evidence from Tunisia. *Journal of Emerging Trends in Economics and Management Sciences.* [Online] 6 (1). p. 13-21. Available from: http://jetems.scholarlinkresearch.com/articles/Impact%20of%20Exports%20and%20Imports.pdf [Accessed: 2 April 2017].

Tapsin, G. (2015) The relationship between foreign trade and economic growth in Turkey. *International Review of Research in Emerging Markets and the Global Economy.* [Online] 1 (3). p. 417-429. Available from: http://globalbizresearch.org/files/6028_irrem_gulcin-tapsin-153797.pdf [Accessed: 30 March 2017].

Ucan, O., Akyildiz, A. & Maimaitimansuer, M. (2016) The relationship between export and economic growth in Turkey. *European Scientific Journal.* [Online] 1 (6). p. 61-70. Available from: http://eujournal.org/index.php/esj/article/view/7532/7262 [Accessed: 30 March 2017].

Verter, N. & Becvarova, V. (2016) The impact of agricultural exports on economic growth in Nigeria. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis.* [Online] 64 (2). p. 691-700. Available from: https://acta.mendelu.cz/media/pdf/actaun_2016064020691.pdf [Accessed: 2 April 2017].

Journal of Applied Technology and Innovation (eISSN: 2600-7304)
vol. 1, no. 2, (2017), pp. 10-27

Data Mining Techniques in Diagnosis of Chronic Diseases

Keerthana Rajendran
Faculty of Computing, Engineering & Technology
Asia Pacific University of Technology & Innovation
57000 Kuala Lumpur, Malaysia
Email: keer.abhitham@gmail.com

Abstract - Chronic diseases and cancer are raising health concerns globally due to lower chances of survival when encountered with any of these diseases. The need to implement automated data mining techniques to enable cost-effective and early diagnosis of various diseases is fast becoming a trend in healthcare industry. The optimal techniques for prediction and diagnosis vary between different chronic diseases and the disease related-parameters under study. This review article provides a holistic view of the types of machine learning techniques that can be used in diagnosis and prediction of several chronic diseases such as diabetes, cardiovascular and brain diseases, chronic kidney disease and a few types of cancers, namely breast, lung and brain cancers. Overall, the computer-aided, automatic data mining techniques that are commonly employed in diagnosis and prognosis of chronic diseases include decision tree algorithms, Naïve Bayes, association rule, multilayer perceptron (MLP), Random Forest and support vector machines (SVM), among others. As the accuracy and overall performance of the classifiers differ for every disease, this article provides a mean to understand the ideal machine learning techniques for prediction of several well-known chronic diseases.

Index Terms - Data mining, Healthcare systems, Machine Learning, Big data analytics

1. Introduction

The evolution of healthcare industry from the traditional healthcare system to the utilization of Electronic Health Records (EHR) system has introduced the concept of big data in the healthcare sector. Big data is defined in terms of 4Vs, which represent the volume, variety, velocity and veracity. The large amount of data generated through omics data such as genomic, proteomic, transcriptomic, epigenomic and metabolomic, as well as EHR data from clinical records, administrative records, charts and laboratory test results, contribute to the copious volume of data in the healthcare industry. Recently, social media data are also being integrated into the EHR system to analyse the patient behaviour. The data are produced in variety of formats from unstructured, semi-structured to structured data with errors such as missing values. Different data have different velocity of generation, so their acquisition time and frequency are largely

varied. Moreover, these data are obtained from diverse sources whose reliability is not authenticated (Raghupathi & Raghupathi, 2014; Auffray et al., 2016). By employing data mining methods into big healthcare data (BHD), several patients can be assessed at the same time and better care can be given based on improved understanding of patient medical profile. Some of the benefits of applying data mining in healthcare are as follows (Durairaj & Ranjani, 2013):

- Optimized management of hospital resources
- Better understanding of patients to improve customer relation
- Detection of fraud and abuses found in insurance and medical claims
- Control the widespread of hospital infections and identify high-risk patients
- Enhanced patient care and treatments through healthcare decision support system

Data mining is defined as the process of identifying unknown patterns, relationships and potentially valuable information from huge datasets with the use of statistical and computational approaches. The primary tasks of data mining are to build predictive and descriptive models as illustrated in Figure 1 (Durairaj & Ranjani, 2013). Data mining techniques that are used commonly in healthcare include classification, decision tree, k-Nearest Neighbour (k-NN), support vector machine (SVM), neural network, Bayesian methods, regression, clustering, association rule mining and Apriori algorithm. These machine learning techniques are helpful in assessing the risk factors such as socioeconomic and environmental behaviour of individuals besides their medical profiles in diseases, especially chronic illnesses and cancer (Tomar & Agarwal, 2013; Dey & Rautaray, 2014; Ahmad, Qamar & Rizvi, 2015).

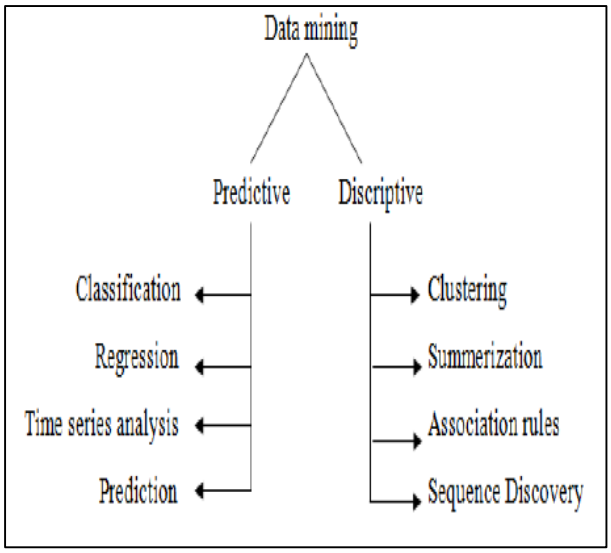

Figure1: Predictive and descriptive data mining techniques (Durairaj & Ranjani, 2013)

This review article focuses on the application of data mining techniques in chronic diseases, with a central focus on different cancer subtypes. The sections are segmented as follows: Section 2 highlights the importance of data analytics in healthcare in general, the current trends of data analytics and some of the data mining methods used in health

informatics. Section 3 elucidates on the existing data mining techniques used in identification of most known chronic diseases such as diabetes disorder, cardiovascular disease, brain disorder and chronic kidney disease (CKD). Each of these disease is explained as a sub-section of Section 3. Section 4 converges its attention to the employment of machine learning techniques in diagnosis, prognosis and prediction of various cancer types which are breast cancer, lung cancer and brain cancer, divided into sub-sections. Section 5 discusses the overall review of the involvement of data mining approaches in chronic diseases and cancer in terms of their limitations and benefits. Section 6 is the conclusion which provides an insight on the challenges of machine learning application in healthcare industry.

2. Data Analytics in Healthcare

Data analytics has a profound use in healthcare, especially in machine learning for the application of descriptive, prescriptive and predictive analytics. Medical data source ranges from EHR, genomic profiles to administrative and financial data, resulting in surplus of data which require extensive application of data mining algorithms to extract valuable output and make informed clinical decisions. In healthcare, the key function of data mining techniques involves the determination of rates of mortality due to disease risk factors and forecast of diseases at an early stage. In a book chapter by Hersh (2017), employment of data-driven measures for diagnosis and tailored treatment of diseases in precision medicine have also been highlighted. Genome-wide association studies (GWAS) incorporates genomic data into the EHR to identify disease disorders and obtain genome-level information. This information is processed and transformed into structured formats for computational analysis using statistical techniques and machine learning algorithms which results in insights gained from the disease prediction models. Inevitably analytics in healthcare data comes with several barriers such as misinterpretation of the transformed data due to coding leading to false positive results and ethical issues raised on the privacy and security of personal data as well as the rights of data access and sharing.

Another review article by Raghupathi & Raghupathi (2014) has carried a similar view point on the involvement of big data in healthcare along with its benefits and features, and conceptual frameworks and tools for data analytics used in healthcare. The vast amount of data generated in various structures and formats are growing exponentially and inability to detect accuracy of these data call in for the need of big data analytics in healthcare. Focusing more into the benefits of data mining in healthcare, the authors highlighted that data analytics can curtail the cost of treatment and diagnosis of patients, reduce trial and error practice in clinical trials by using data analytical tools and algorithms, estimate population health trends, recognize patients who are prone for re-admission, mitigate fraudulence and misuse, enable real-time update of patient conditions and perform genome-based analytics for precision medicine. The

architectural skeleton highlighted how multitude of data obtained from various sources in a raw form can be placed in a data warehouse to allow data transformation. These transformed data then go into a selection of tools/platforms for applied analytics. Hadoop is a prominent platform to analyse and process big data. Steps involved to apply big data analytics in healthcare are conception of project, manifestation of proposal, implementation of methods such as data compilation and processing, and lastly, deployment of the results. Obstacles such as privacy, data security, real-time data evaluation, quality and governance regarding big healthcare data should be brought to light.

The article by Dinov (2016) showcased the possible techniques to conquer these barriers so that convoluted data can be transformed into comprehensible format for analysis. To achieve an automatically processed decision, quantitative and structured format of BHD are vital. To derive statistically valuable data, some measures that can be used include visual mining, text analytics, information retrieval, data standardization and predictive modeling markup language (PMML) to interpret and analyse medical information. Social network analytics, which can be displayed in terms of nodes and edges, are used to identify relationships and patterns in the datasets. Multitude techniques such as k-NN, Gaussian mixture modeling (GMM) as well as un-supervised, semi-supervised supervised machine learning algorithms are utilized to segment, group and organize complex data. Incomplete records which might have missing values occurring at random or not random can be handled via logistic regression. Exploratory and explanatory analyses can be used to analyse and display incongruent data using cloud-based dashboards. Along this, predictive analytics is the crucial task of data mining in BHD. Programming languages such as SQL and NoSQL besides cloud computing are advances to refine BHD. Open-source platforms like Apache, Hadoop, MapReduce and Spark are freely accessible to analyse large data in the healthcare sector.

As healthcare information are known to exist in copious amount, it is necessary to construct a standardized series of steps to analyse these data and interpret them into knowledgeable output. A well-accepted process that is followed in healthcare data mining is Knowledge Discovery (KDD) which involves the interpretation of large volume of data and identifying a pattern in the data to attain an insight that improves decision-making, as shown in Figure 2 (Ahmad, Qamar & Rizvi, 2015). In KDD, selection of target data from the database is the first step, followed by data pre-processing where any unwanted data are filtered and the noisy data are eliminated. The raw and unstructured data are then transformed into a structured format for analysis. Data mining is the key process in KDD which involves descriptive and predictive analytics incorporating numerous algorithms and statistical measures to discover trends and build prediction models. Data mining is primarily assigned to carry out two approaches known as static end-point prediction and temporal data mining which consists of classification, regression, association rule learning, cluster analysis, hidden Markov model (HMM) and temporal association rule mining (TARM) on the transformed datasets. The interpreted

outcomes allow informed decision-making. Examples of data mining application in healthcare are artificial neural network of human brain, besides decision tree and nearest neighbour for classification and prediction. In precision medicine, genomic data incorporated into the EHR were analysed using clustering to identify cancer subtypes and provide tailored treatments (Taranu, 2015; Wu et al., 2017).

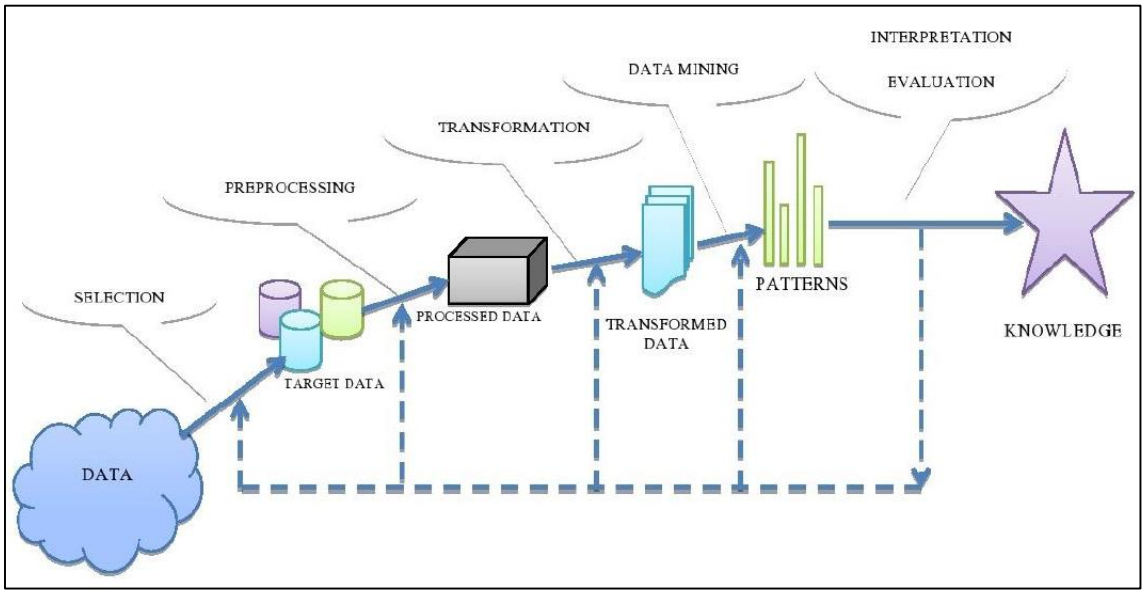

Figure 2: KDD Process (Ahmad, Qamar & Rizvi, 2015)

Application of computational knowledge in healthcare is called health informatics which is transpiring into a demanding field requiring data experts due to the evolution of big data in healthcare. Within this field, there are several subdomains such as bioinformatics, medical image informatics, clinical informatics and public health informatics which employs various levels of data generation from molecular, tissue, patient and population level data. These data are utilized to address the research questions posed in order to find answers for clinical, human biological and epidemic queries. Herland, Khoshgoftaar & Wald (2014) have elucidated on the various data mining techniques applied at each data levels in health informatics. At molecular level, patient cancer gene expression profiles were analysed to group leukaemia sub-types and to forecast the recurrence of colorectal cancer (CRC) among the subjects at the initial phase using classification and support vector machines (SVM) techniques. Tissue level involve methods like feature abstraction and selection applied on highly dimensional human brain images and MRI of brain tissue samples to develop an extensive neural network. Besides, Fuzzy Decision Tree (FDT) and classification were used to predict the occurrence of Alzheimer's disease at distinct stages. The third level uses the patient records tested using various scoring systems under classification technique and logistic regression to build prediction models for patient readmission rate, fatality estimate and life span of patient. Moreover, Alternating Decision Tree (ADT) and Principal Component Analysis (PCA) were also used for prediction based on patients' physiological

parameters, derived from EHR. At population level, text analytics were employed on social media data such as Twitter, internet search engines and messaging applications to provide patients with facts on illnesses. Real-time epidemics tracking and prognosis of a population health were determined using data mining techniques such as decision trees, SVM, Naïve Bayesian and logistic regression analysis.

3. Data Mining in Chronic Diseases Identification

The current trends in data mining field that are available to diagnose different chronic diseases that are the common causes of death worldwide such as diabetes, cardiovascular disease, brain disease and chronic kidney disease are illuminated in this section.

3.1 Diabetes Disorder

Diabetes mellitus is a condition where the body is unable to produce sufficient insulin, resulting in increased blood sugar level. Various risk factors contribute to diabetes which can be used as measures to predict diabetes. The article by Renuka Devi & Maria Shyla (2016) highlighted on diabetes mellitus discussed various data mining algorithms used on the Pima Indian Diabetes Dataset to determine the occurrence of diabetes. The data was cleaned to eliminate noise and replace missing instances. A total of six physiological variables were used to diagnose type 1 and type 2 diabetes, besides gestational diabetes. Some of the techniques used to classify the diabetes-related attributes include Naïve Bayes, Random Forest, Modified J48 Classifier, SVM, k-NN, genetic algorithm, etc. Software tools such as Weka, MATLAB, Tanagara, RapidMiner, etc. were used to perform data analytics operations containing all the machine learning techniques and statistical algorithms. Upon comparison across all the techniques, it was found that the highest prediction accuracy of 99.87% was achieved using Modified J48 Classifier with the aid of Weka and MATLAB tool.

Another study on type 2 diabetes mellitus by Hu et al. (2016) explained that impaired glucose tolerance (IGT) is a causative factor of this disease as well as atherosclerosis which can lead to cardiovascular disease. Hence, data mining techniques were employed to identify the patterns among IGT patients with an elevated risk of atherosclerosis. ACT NOW, a clinical trial dataset, was used as the training dataset where it was processed using imputation and categorical attributes were created for the disparate variables. First, feature selection using Fisher score was assigned to choose the attributes with most significance. Probabilistic Bayesian classifiers were used to generate the prediction model which has been proven to work well on small-scale, multimodal training and test datasets in other reported studies. Two more classification techniques known as multilayer perceptron (MLP) and random forest (RF) were adopted to compare the accuracy of the prediction, which was determined using Brier score and receiver

operating characteristic curve (AUC). The best predictor was found to be the Naïve Bayes with feature selection holding a better performance with 89.23% accuracy in comparison to the other two approaches with about 88% accuracy level each. Naïve Bayes technique proved to enable the determination and prognosis of IGT subjects who encounter rapid progression of atherosclerosis.

3.2 Cardiovascular Disease

Heart disease is one the most common chronic diseases that causes fatality in adults. Heart disease prediction and their risk factors analyses have been reported in several articles using machine learning algorithms such as SVM, decision tree, genetic algorithm, neural networks, Naïve Bayesian classifiers and Iterative Dichotomized 3 (ID3). But, association rule technique in cardiovascular disease prognosis has been barely explored. Hence, Khare & Gupta (2016) have incorporated association rule into their study to identify heart disease risk factors. The data was obtained from UCI open data repository. Training dataset was prepared by removing the missing observations and shortlisting the key attributes that are related to the analysis. The numerical attributes were nominalized to categorical attributes since Apriori algorithm of the association rule mining was used. This algorithm utilized candidate generation approach to discover the recurring (frequent) variable set. Rules were generated based on the presence of heart disease by focusing on the causative determinants for this disease, where classification association rules (CAR) were applied. Accuracy of the rule was validated in the training stage. The schematic representation of association rule method is shown in Figure 3. The results showed at least 85% confidence for all the rules generated. Attributes such as gender (male), older generation, increased serum cholesterol level, presence of asymptomatic chest pain and defective thalassemia are associated with the exposure to heart disease. While blood sugar level, an indirect measure of diabetes, was found to be negatively correlated to heart disease. This technique was found to be useful in focusing only on the primary risk factors of heart disease, which enables cost-effective diagnosis and time-efficient treatments.

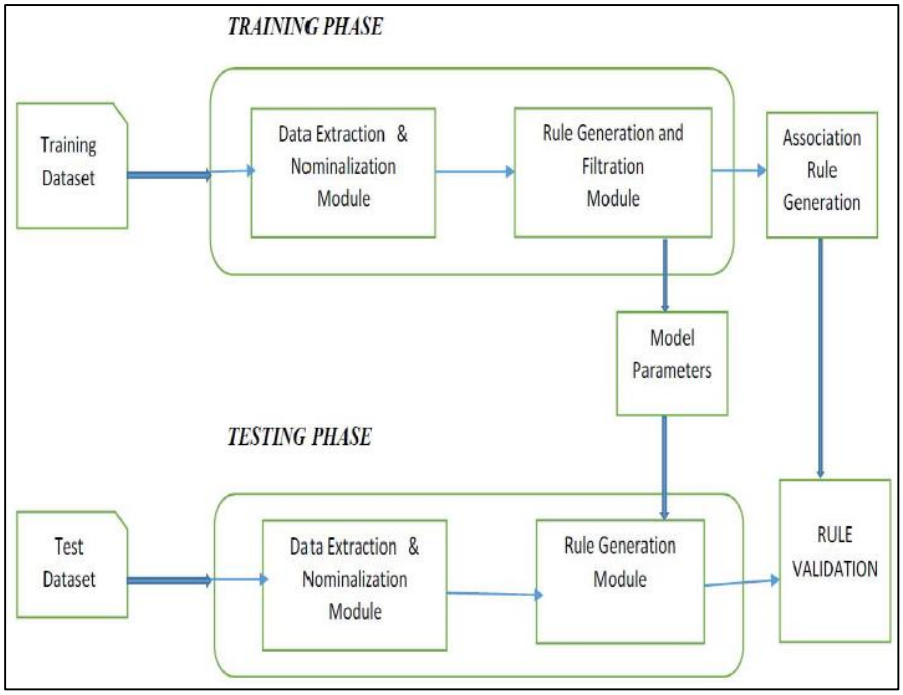

Figure 3: Association rule technique (Khare & Gupta, 2016)

3.3 Brain Disease

One of the well-known type of dementia among older people is Alzheimer's disease (AD). This neurodegenerative illness is one of the primary causes of fatality in the USA. Besides other external factors such as lifestyle and medical circumstances, genetic factors appear to have greater influence in the progression of AD. A study was conducted by Kumar & Singh (2016) to determine the participation of AD-related genes in the pathogenic pathway and diagnosis using decision tree method. The AD gene datasets were obtained from several online repositories and 2111 genes significant to the disease were selected. Feature selection of the variables were done using Chi-squared attribute evaluation and gain ratio (GR) methods. J48 algorithm found in Weka and C4.5 algorithm enabled in RapidMiner were used to achieve classification of the dataset and clustering of the genes were done through enrichment analysis. Mini Mental State Examination (MMSE) scores, one of the vital attributes, obtained through classification algorithm showed different values for distinct phases of AD. Upon classification of similar features, decision tree for the gene dataset was built using MMSE score as the root node. The roles of the genes were determined via enrichment analysis and 7 genes were found to be highly correlated with AD based on the association scores. The C4.5 algorithm proved to generate a better prediction accuracy with minimal error rate for prior AD diagnosis.

An early diagnosis of chronic diseases can reduce fatality greatly and this is a demanding area of data mining application in healthcare. But, errors in diagnosis often lead to delayed treatments or administration of wrong medications. Thus, a study was done by Chase et al. (2017) on patients with multiple sclerosis (MS) to recognize the signs

and symptoms of the disease in an early stage using natural language processing (NLP) of the unstructured medical notes in the EHR. Medical records of patients with MS and a randomly selected population in a clinic at Columbia University Medical Center (CUMC) were used as datasets for this study. The MS-positive population data were classified to build a prediction model for pre-recognition of MS, while the random patient group data were used to determine unidentified MS. A classifier was trained to recognize the terms that are linked and not linked to MS. Employing Naïve Bayes algorithm under classification technique from Weka tool, differences between patients with MS and those without MS were determined. Terms with similar features or category were classed into a bucket. The model generated a sensitivity of 75% and specificity of 91% in MS-identified patients, while in the random group 81% sensitivity and 87% specificity of classification were achieved. This shows that the method is feasible for pre-diagnosis of MS prior to the detection of ICD9 code, which is a gene marker for MS. The classification of the random population proved to have assisted in identifying patients who may have MS, based on the signs and symptoms but are missing the neurological characteristics of MS. The limitations of the study include restricted number of patients, the classifier was developed based on majority of Hispanic female patients' population, and only one NLP system was used to identify the MS terms.

3.4 Chronic Kidney Disease

Chronic kidney disease (CKD) has recently become an increasing health concern worldwide, but there are yet to be many research papers on computerized diagnosis for this disease. UCI Machine Learning warehouse dataset was used in a study by Subasi, Alickovic & Kevric (2017) and well-studied techniques such as artificial neural network (ANN), SVM, k-NN, C4.5 decision tree and random forest (RF) were used to build prediction models for CKD analysis. Under ANN, MLP approach was used to identify the correlation between the 24 variables and 2 potential output results (with CKD or without CKD) via the network of neurons and nodes. Supervised SVM was used to assess classification and regression analyses by segregating the training group using hyperplanes. An improvisation of ID3 known as C4.5 decision tree was used on this CKD dataset as it can analyse numeric variables, while Classification and Regression Tree (CART) under random forest was implemented. The training set comprised of 90% of the dataset while 10% was used as test dataset. The outputs were displayed using a Confusion Matrix for various data mining techniques. The classification models were generated using Weka tool and the performances of the machine learning algorithms were statistically validated using total precision, F-measure and overall classification accuracy. All the techniques used yielded exceptional performance accuracies but RF classifier proved to have the highest classification performance rate, even compared to other techniques used in previous literature studies. Thus, RF was proposed as an ideal data mining technique to evaluate for an early diagnosis of CKD at a faster time.

4. Machine Learning in Cancer Diagnosis and Prediction

The data mining techniques and algorithms used to predict the occurrence of several types of the most common cancers such as breast, lung and brain cancers are explored in detail in this section.

4.1 Breast Cancer

Oftentimes, the gene expression arrays, also known as microarray profiling, were utilized predominantly in the diagnosis of cancer disease as genetic compositions have greater influence in the cause of cancer. But, by incorporating patient clinical records such as ultrasound images and laboratory results into the microarray data, this will enhance the decision-making process during cancer diagnosis. In the study by Gevaert et al. (2006), Bayesian networks are employed where it treats the clinical data and microarray analysis on the same level of importance in the prediction of breast cancer where the output is evaluated as either a poor or good prognosis. Patient dataset consists of the training set and testing set where each set has poor and good forecasts of breast cancer. Physical and biological parameters of patients were obtained from the clinical notes and were combined with microarray analysis dataset. The Bayesian network software used the pre-processed data as the input source. Three methods of combining clinical and microarray data were administered which were full, decision and partial integration and their performances were validated using Area Under the ROC (Receiver Operator Characteristics) Curve (AUC). Decision and partial integration showed a significantly varied ROC AUC compared to the other methods, thus they were utilized to build the models for the training dataset. Best Partial Integration Model (BPIM) was found to have better performance than Best Decision Integration Model (BDIM) as the former requires lesser number of genes when incorporating clinical data for forecasting the disease prognosis. The results also portrayed that the clinical and microarray variables in the Markov blanket enhances the model performance. Thus, BPIM under the Bayesian networks approach provides a mean for a cost-effective diagnosis of breast cancer while retaining the molecular level data.

Another research by Thomas et al. (2014) on breast cancer was done by incorporating the clinical and microarray data using a suggested data mining technique called weighted Least Square SVM (LS-SVM) classifier to result in an improved prognostic application in breast cancer therapy. The five datasets were obtained from the Integrated Tumour Transcriptome Array and Clinical data Analysis (ITTACA) repository, where 2/3rd of the population was used as training set and the rest for testing. Transformation of each dataset into a kernel matrix was done and an integration framework was developed. The paper elucidated on the Generalized Eigenvalue Decomposition (GEVD) and LS-SVM formulations, where they recommended a new machine learning technique, called weighted LS-SVM classifier, for data integration and classifications. The

performance rate of each dataset was compared for GEVD, kernel GEVD and weighted LS-SVM classifier using test AUC and Leave-One-Out Cross Validation (LOO-CV). The proposed classifier approach introduced an optimized single framework to resolve the issues of excessive cost and classification using heterogenous datasets and improved the prediction and treatment efficiencies for each patient.

Breast cancer has evolved as one of the most frequent cancer types globally among female and has contributed to major death rates. To subdue this concern, data mining tools can be implemented into the clinical system of cancer diagnosis at an early stage so that ideal treatments can be administered. Alickovic & Subasi (2015) explained that the traditional method of clinical diagnosis of breast cancer such as mammography can be supported with automatic diagnostic tools to assist in distinction between benign and malignant breast tumours. Two distinct datasets were acquired from UCI Machine Learning repository consisting of benign and malignant tumours data. The machine learning approaches used in this study are RF, MLP, SVM, C4.5 Decision Tree, Logistic Regression, Bayesian Network, Radial Basis Function Networks (RBFN) and Rotation Forest. Rotation Forest creates classifier groups by segmenting features into subsets and applying Principal Component Analysis (PCA) on each subset. Distinct rotations are formed from different feature set splits, resulting in variant classifiers with diversity and precision. Genetic Algorithms (GA) are used to select the key features from the breast cancer datasets as inputs to classifiers to enhance the classification accuracy of the multitude data mining techniques. Weka tool was used to execute these algorithms. AUC ROC was used to represent the classifiers' performances. The results portrayed that Rotation Forest, which is a Multiple Classifier System (MCS), with GA-based feature selection produced the greatest classification accuracy of 99.48% in the breast cancer dataset compared to other classification algorithms. This study recommended the employment of this novel technique by clinicians to correctly assess breast cancer diagnosis and enhance decision-making.

4.2 Lung Cancer

Lung cancer is another cancer type that is listed as the most occurring chronic disease worldwide. It is vital to infer an early diagnosis, identify the right type of lung cancer and provide accurate treatment to reduce the fatality rate among the carriers of this disease. An optimal solution to this is by employing various machine learning techniques as done by Podolsky et al. (2016) in their study to confirm an early diagnosis. This research used four open datasets that contain gene expression levels for various lung cancer types. The data mining methods used to analyse the data were k-NN with increasing degrees of freedom, Naïve Bayes classifier, SVM and C4.5 decision tree. Two of the datasets used 10-fold cross validation to be used as inputs into the algorithms while the other two datasets had pre-prepared training and testing sets. ROC AUC and Matthews Correlation Coefficient (MCC) were calculated to validate the performance of the algorithms. The

authors discussed that the SVM algorithm showed high performance values for two of the gene datasets which can be used to categorize lung cancer based on their histological variants at a great accuracy. The third analysed dataset was precisely distinguished between healthy lung and adenocarcinoma via all the data mining techniques except C4.5 decision tree. But, this technique was found to be compatible on the fourth dataset. This study concluded that SVM is the most ideal machine learning algorithm for an effective diagnosis of lung cancer and can predict tumour development and its metastasis.

4.3 Brain Cancer

Cancer of the brain is a very critical and life-threatening chronic disease as brain is part of the central nervous system of the body and any damages to brain activities will also affect the other parts of the body. Brain cancer exists as primary tumours (originated from brain cells) and secondary or malignant tumours (cancer cells that originate from other parts of the body and metastasize in the brain). A study was done by George, Jehlol & Oleiwi (2015) to classify Magnetic Resonance Images (MRI) to identify the brain cancer types and its subtypes using supervised machine learning approaches. C4.5 decision tree and MLP algorithms were used to perform automated classification of a type of benign cancer tumour and five variants of malignant cancer. Figure 4 illustrates the framework of the classification method used in this study. The MRI data were acquired from hospitals and from internet due to the lack of brain cancer data in open databases. Image pre-processing was done to overcome the problem of high dimensionality of data through sigma filtering to eradicate noise from MRI. The images were segmented using adaptive threshold method which partition the data based on grey or coloured images to generate binary image representations. Region detection was utilized to identify and separate the tumour region from the other objects in the MRI. The features were then extracted based on six shape properties related to brain images using MATLAB program. The data was split into 55% for training set and the remaining was used for test validation. The C4.5 algorithm displayed 91% accuracy while MLP yielded 95% accuracy for the overall brain tumour types. MLP was found to take more time to build the model compared to C4.5 decision tree algorithm. The study proposed to use a larger dataset with more features to improve the accuracy of diagnosis and performance of the classifiers.

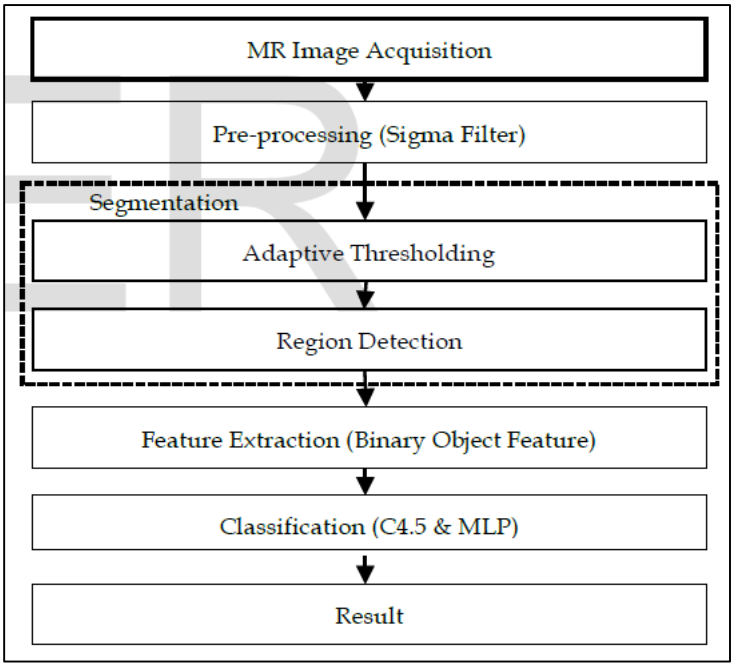

Figure 4: Brain cancer classification architecture (George, Jehlol & Oleiwi, 2015)

5. Discussion

The papers that have been discussed in the previous sections addressed few limitations while employing the data mining techniques. One of the issues encountered is that the size of the dataset is small, hence it was difficult to validate the performance and accuracy of prediction and diagnosis. The algorithm that gave the best result may work on a smaller dataset but the same technique might show different performance in a larger dataset for the same disease conditions. Thus, a larger sample size can yield better prediction accuracies with less errors (George, Jehlol & Oleiwi, 2015; Khare & Gupta, 2016; Hu et al., 2016).

Secondly, heterogenous datasets with different attributes were compatible with different algorithms. Even for the same disease type, the performance of the algorithms was found to be varied for distinct datasets. This could be due to the criteria of the algorithms such as feature selection and filtering, size of the dataset, percentage of training and testing datasets, types of statistical measures used to evaluate the performance and high dimensionality such as medical visual images in the dataset. Hence, it is crucial to compare across the multitude machine learning techniques to attain an optimum result and reduce cost of diagnosis and treatment (Renuka Devi & Maria Shyla, 2016; Podolsky et al., 2016). Another limitation in a study highlighted that the population ethnicity and gender used in the dataset may have showed a biased result of the ideal data mining technique. The accuracy of the technique may be different if applied on different race or gender (Chase et al., 2017).

But majority of the articles have emphasized the primary benefits of using data mining techniques in assessing chronic diseases. The main advantage was that there was substantial reduction in cost incurred for diagnosis and treatment of the diseases by using machine learning methods. Time acquired for training and validating the algorithms were much lesser compared to clinical diagnosis. Further, the prediction models were generated in shorter time so the necessary treatments can be provided without any delay. Data mining also optimized the number of diagnosis required to medically assess the chronic diseases, thus the treatment costs would also be reduced significantly. Indirectly, this creates an affordable and high quality medical care for all and reduces the fatality rate globally (Gevaert et al., 2006; Khare & Gupta, 2016; Subasi, Alickovic & Kevric, 2017).

To overcome the limitations of the techniques, some articles have highlighted that two or more algorithms can be merged to produce new algorithms, which yielded better performance and precision results. More such approaches can be further explored for different chronic diseases, especially for cancer prediction and prognosis (Thomas et al., 2014; Alickovic & Subasi, 2015). Open-software tools that are free of charge such as Weka, MATLAB and RapidMiner have multiple in-built algorithms which were used in most of the papers to generate prediction models and to identify relationships and trends in chronic diseases. These tools can also be used by non-technical people who do not have IT background but have domain knowledge such as doctors or clinicians to make informed, evidence-based healthcare decisions (Kumar & Singh, 2016; Renuka Devi & Maria Shyla, 2016). Table 1 highlights the comparison of distinct data mining techniques in various chronic diseases as well as application of different algorithms for the same disease type.

Table 1: Comparison of data mining techniques and different algorithms in the literature

Chronic Disease	Author & Year	Contribution	Data Mining Technique
Type 1 & 2 diabetes mellitus & gestational diabetes	Renuka Devi & Maria Shyla (2016)	Applied data mining methods to classify diabetes-related attributes and predict occurrence of diabetes	Modified J48 classifier
Type 2 diabetes mellitus	Hu et al. (2016)	Enabled prognosis of IGT patients who encountered rapid progression of atherosclerosis	Naïve Bayes
Heart disease	Khare & Gupta (2016)	Identified primary heart disease risk	Association rule

		factors using machine learning techniques	
Alzheimer's disease	Kumar & Singh (2016)	Determined Alzheimer's correlated genes using decision tree for accurate prediction and prior diagnosis of disease	C4.5 algorithm
Multiple sclerosis (MS)	Chase et al. (2017)	Built a prediction model for pre-recognition of MS based on natural language processing of unstructured medical notes	Naïve Bayes
Chronic kidney disease (CKD)	Subasi, Alickovic & Kevric (2017)	Established a better classification and prediction technique for early diagnosis of CKD compared to previous works	Random Forest classifier
Breast cancer	Gevaert et al. (2006)	Identified a cost-effective diagnosis technique using minimum number of genes and clinical data	Bayesian networks
	Thomas et al. (2014)	Introduced an optimized single framework to improve disease prognosis in breast cancer therapy	Least Square Support Vector Machines (LS-SVM) classifier
	Alickovic & Subasi (2015)	Distinguished between benign and malignant cancer using classifier system with GA-based feature selection	Rotation Forest
Lung cancer	Podolsky et al. (2016)	Identified a technique for an effective diagnosis of lung cancer and can predict tumour	Support Vector Machines (SVM)

		development and its metastasis	
Brain cancer	George, Jehlol & Oleiwi (2015)	Built an automated system to classify medical that enabled identification of brain cancer types and its subtypes	Multilayer Perceptron (MLP) and C4.5

6. Conclusions

Big data analytics in the healthcare industry have given an insight and awareness towards the importance of data mining applications in diagnosis, prediction and prognosis of various illnesses and disorders. Past few years have seen an increase in the research for discovering ideal machine learning algorithms and tools for disease modelling. Chronic diseases such as diabetes, cardiovascular diseases, brain disease, CKD and various cancer types like breast, brain and lung cancers have caused many public health concerns globally and with the influence of machine learning approaches in the diagnosis of these diseases, the studies have proven to yield a reduction in the mortality rate. Yet, the full functionality of machine learning utilization remains a challenge due to the concerns in data security, privacy, integrity and exchange. Government legislations and encryption techniques are being employed actively on data analytics nowadays to mitigate these social and ethical issues.

Acknowledgements

I would like to express my deepest gratitude and appreciation to my lecturers, Prof. Dr. Logeswaran and Mr Manoj Jayabalan for their guidance and suggestions during this review writing. I would also like to acknowledge with much appreciation to all those who have helped me in completing this report successfully.

References

Ahmad, P., Qamar, S. & Rizvi, S. (2015). Techniques of data mining in healthcare: a review. International Journal of Computer Applications. 120(15). p. 38-50.

Alickovic, E. & Subasi, A. (2015). Breast cancer diagnosis using GA feature selection and rotation forest. Neural Computing and Applications. 28(4). p. 753-763.

Auffray, C., Balling, R., Barroso, I., Bencze, L., Benson, M., Bergeron, J., Bernal-Delgado, E., Blomberg, N., Bock, C., Conesa, A., Del Signore, S., Delogne, C., Devilee, P., Di Meglio, A., Eijkemans, M., Flicek, P., Graf, N., Grimm, V., Guchelaar, H., Guo, Y., Gut, I., Hanbury, A., Hanif, S., Hilgers, R., Honrado, Á., Hose, D., Houwing-Duistermaat, J., Hubbard, T.,

Janacek, S., Karanikas, H., Kievits, T., Kohler, M., Kremer, A., Lanfear, J., Lengauer, T., Maes, E., Meert, T., Müller, W., Nickel, D., Oledzki, P., Pedersen, B., Petkovic, M., Pliakos, K., Rattray, M., i Màs, J., Schneider, R., Sengstag, T., Serra-Picamal, X., Spek, W., Vaas, L., van Batenburg, O., Vandelaer, M., Varnai, P., Villoslada, P., Vizcaíno, J., Wubbe, J. & Zanetti, G. (2016). Making sense of big data in health research: towards an EU action plan. Genome Medicine. 8(1).

Chase, H., Mitrani, L., Lu, G. and Fulgieri, D. (2017). Early Recognition of Multiple Sclerosis Using Natural Language Processing of the Electronic Health Record. BMC Medical Informatics and Decision Making, 17(1).

Dey, M. & Rautaray, S. (2014). Study and analysis of data mining algorithms for healthcare decision support system. International Journal of Computer Science and Information Technologies. 5(1). p. 470-477.

Dinov, I. (2016). Methodological challenges and analytic opportunities for modeling and interpreting big healthcare data. GigaScience. 5(1).

Durairaj, M. & Ranjani, V. (2013). Data mining applications in healthcare sector: a study. International Journal of Scientific & Technology Research. 2(10). p. 29-35.

George, D., Jehlol, H. & Oleiwi, A. (2015). Brain tumor detection using shape features and machine learning algorithms. International Journal of Scientific and Engineering Research. 6(12). p. 454-459.

Gevaert, O., Smet, F., Timmerman, D., Moreau, Y. & Moor, B. (2006). Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. Bioinformatics. 22(14). p. 184-190.

Herland, M., Khoshgoftaar, T. & Wald, R. (2014). A review of data mining using big data in health informatics. Journal of Big Data. 1(1). p. 2.

Hersh, W. (2017). Healthcare data analytics. Health Informatics: Practical Guide for Healthcare and Information Technology Professionals, 6th Ed. Florida: Informatics Education. p. 62-75.

Hu, X., Reaven, P., Saremi, A., Liu, N., Abbasi, M., Liu, H. & Migrino, R. (2016). Machine learning to predict rapid progression of carotid atherosclerosis in patients with impaired glucose tolerance. EURASIP Journal on Bioinformatics and Systems Biology. 2016(1).

Khare, S. & Gupta, D. (2016). Association Rule Analysis in Cardiovascular Disease. In Proceedings of the 2016 Second International Conference on Cognitive Computing and Information Processing (CCIP). Mysore, India: IEEE. p. 1-6.

Kumar, A. & Singh, T. (2016). A new decision tree to solve the puzzle of Alzheimer's disease pathogenesis through standard diagnosis scoring system. Interdisciplinary Sciences: Computational Life Sciences. 9(1). p. 107-115.

Podolsky, M., Barchuk, A., Kuznetcov, V., Gusarova, N., Gaidukov, V. & Tarakanov, S. (2016). Evaluation of machine learning algorithm utilization for lung cancer

classification based on gene expression levels. Asian Pacific Journal of Cancer Prevention. 17(2). p. 835-838.

Raghupathi, W. & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. Health Information Science and Systems. 2(1).

Renuka Devi, M. & Maria Shyla, J. (2016). Analysis of various data mining techniques to predict diabetes mellitus. International Journal of Applied Engineering Research. 11(1). p. 727-730.

Subasi, A., Alickovic, E. & Kevric, J. (2017). Diagnosis of Chronic Kidney Disease by Using Random Forest. In Proceedings of the International Conference on Medical and Biological Engineering 2017 (IFMBE Proceedings, Vol 62). Singapore: Springer. p. 589-594.

Taranu, I. (2015). Data mining in healthcare: decision making and precision. Database Systems Journal. 4(4). p. 33-40.

Thomas, M., Brabanter, K., Suykens, J. & Moor, B. (2014). Predicting breast cancer using an expression values weighted clinical classifier. BMC Bioinformatics. 15(1).

Tomar, D. & Agarwal, S. (2013). A survey on data mining approaches for healthcare. International Journal of Bio-Science and Bio-Technology. 5(5). p. 241-266.

Wu, P., Cheng, C., Kaddi, C., Venugopalan, J., Hoffman, R. & Wang, M. (2017). -Omic and electronic health record big data analytics for precision medicine. IEEE Transactions on Biomedical Engineering. 64(2). p. 263-273.

Journal of Applied Technology and Innovation (eISSN: 2600-7304)
vol. 1, no. 2, (2017), pp. 28-40

Literature Review of Data Mining Techniques in Customer Churn Prediction for Telecommunications Industry

Sherendeep Kaur
Faculty of Computing, Engineering & Technology
Asia Pacific University of Technology & Innovation
57000 Kuala Lumpur, Malaysia
Email: TP019638@mail.apu.edu.my

Abstract **-** Customer churn is one of the most critical issues faced by the telecommunications industry. In the telecommunications industry, it is more expensive to acquire a new customer as compared to retaining the current one. Hence, customer churn prediction is currently the main mechanism employed by the industry in order to prevent customers from churning. The objective of churn prediction is to identify customers that are going to leave the telecommunications service provider in advance. Customer churn prediction would allow the telecommunications service provider to plan their customer retention strategy. The high volume of data generated by the industry, with the help of data mining techniques implementation, becomes the main asset for predicting customer churn. Due to this reason, recent literature of different data mining techniques and most popular data mining algorithms for customer churn prediction are reviewed in this paper. Additionally, recent literature on newly developed algorithms based on the popular algorithms are also reviewed.

Index Terms **–** Data Mining, Big data analytics, Churn prediction

1. Introduction

The telecommunications industry in today's world is dealing with a major deficit towards their generated revenue due to the aggressive market competition (Umayaparvathi & Iyakutti, 2012). The telecommunications industry as any other service providers' industries consider customers to be the most crucial resource for them. However, the aggressive competition, such as lucrative retention polices offered by telecommunications service providers to attract new customers, has resulted major loss of customers in this industry as customers tend to leave one telecommunications service provider to another for this reason or another (Adwan et al., 2014). It is possible for telecommunications companies to focus more on acquiring new customers, however customer acquisition usually costs more compared to customer retention, and this will eventually lead to lower revenue (Chen, Fan & Sun, 2012). Hence, due to this reason, the

telecommunications industry has shifted their focus to customer retention (Adwan et al., 2014).

The act of customers leaving one service provider to another service provider is called customer churn (Umayaparvathi & Iyakutti, 2012). Recently, customer churn prediction has become a highly-discussed domain. Many studies have analysed customer churn problem from many different viewpoints to design and recommend the best solution to the telecommunications industry (Shaaban et al., 2012).

Data mining techniques have been widely used as the solution to predict customer churn, by identifying the factors that are most likely contributing into customer churn in order to enable telecommunications service providers to take immediate action to prevent churning (Umayaparvathi & Iyakutti, 2012). The large volume of data, such as demographic data, billing information, call details, network details, among others help to ensure the accuracy of the data mining technique's application in the telecommunications industry (Umayaparvathi & Iyakutti, 2012). Among many data mining techniques for customer churn prediction, supervised data mining techniques are the most extensively explored. Supervised data mining techniques are appropriate when models to be developed can learn from labelled training data. Supervised data mining techniques consist of varied algorithms such as linear regression, neural networks, decision trees, k-nearest neighbours, genetic algorithms, Naïve Bayes, support vector machines (SVM) and others. (Shaaban et al., 2012)

In section 2, customer churn prediction is explained in detail and the objective of customer churn prediction as well as the types of churners are also discussed. In section 3, data mining is defined and the relation between data mining and knowledge discovery in database (KDD) is explained. Additionally, the steps required in KDD process are also explained in this section. In section 4, various data mining techniques and algorithms for customer churn prediction in telecommunications industry is presented. Lastly, the conclusion is presented in section 5.

2. Customer Churn Prediction

The objective of customer churn prediction is to predict the impending churners based on the predefined forecast horizon, assuming the data related with each subscriber in the network. According to Umayaparvathi & Iyakutti (2012), the customer churn prediction problem is normally characterized into three major stages, namely, training stage, test stage and prediction stage. In the training phase, the contribution for customer churn problem is from the historical data such as call details and personal and/or business customers' data, which has been obtained and retained by the telecommunications service providers. Furthermore, in the training stage, the labels are structured in the list of churners' records. In the test stage, the trained model with highest accuracy is tested to predict the churners' records from the actual dataset which does not contain any churn

label. Lastly, in the prediction stage, which is also known as the knowledge discovery process, the problem is classified as predictive modelling or predictive mining.

Customer churn prediction helps the customer relationship management (CRM) to avoid customers who are expected to churn in future by proposing retention policies and offering better incentives or packages to attract the potential churners in order to retain them. Hence, the possible loss of the company's revenue can be prevented. (Umayaparvathi & Iyakutti, 2012)

Shaaban et al. (2012) stated that there are two types of churners, namely, involuntary and voluntary. Involuntary churners are the list of customers that are removed by the telecommunications service provider, itself, due to non-payment status, deception and non-usage of phone. Meanwhile, voluntary churners are the customers that decide to terminate their service with the respective telecommunications service provider. Involuntary churners are easy to be recognized; however, the voluntary churners are more difficult to be identified. Generally, the customer churn problem in the telecommunications industry is voluntary.

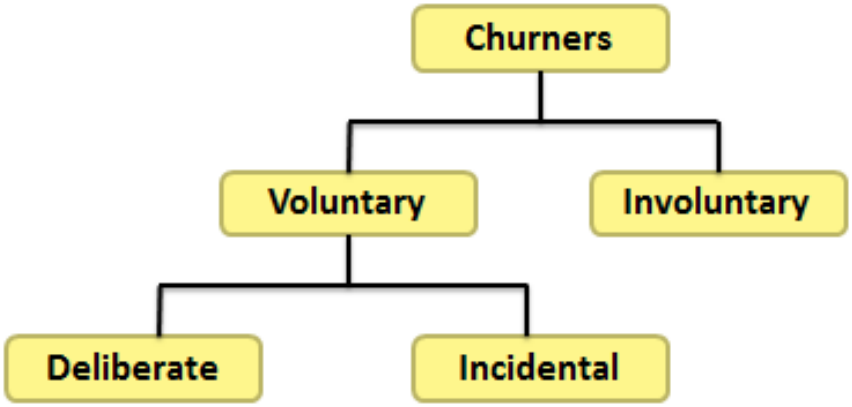

Figure 1: Types of Churners (Shaaban et al., 2012)

It can be seen in Figure 1 that voluntary churn is separated into two sub-categories which are deliberate churn and incidental churn. Deliberate churn is resulted from factors such as economic factors (example: price sensitivity), technology factors (example: more innovative technology is offered by another telecommunications service provider), poor customer service factors and others inconvenience related factors. Incidental churn is not caused by customers' plan, but because of sudden changes occurred in the customers' lives, such as changes in financial situation, geographical or relocation changes and others. (Shaaban et al., 2012)

3. Data Mining

Data mining is described as the process of determining and extracting valuable information in large databases (Verbeke, 2012). Data mining is an essential element of knowledge discovery in databases (KDD) process. KDD is a process which explains the

steps that must be taken to ensure a thorough data analysis. KDD process contains five steps as illustrated in Figure 2.

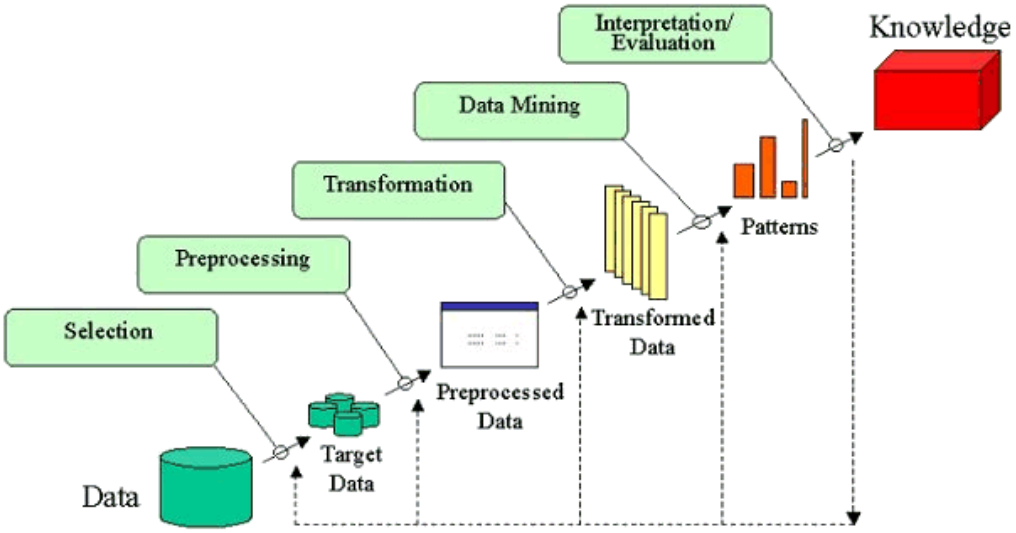

Figure 2: Steps of KDD Process (Tomar & Agarwal, 2014)

The first step involves the selection of the right data from numerous heterogenous data sources to create the target dataset. The right selection of attributes and variables is very important at this step in order to produce an accurate analysis. In data preprocessing step, missing values, outliers and inconsistency are removed and/or cleaned. This step is required to enhance the quality of the target dataset as well as the mining results. The transformation step focuses on transforming the variables from one format to another to ease the implementation of the data mining algorithms. The transformation step involves changing variables from one format to another to perform data mining. The application of data mining step has to be aligned with the objective of the entire KDD process, in this case, customer churn prediction. The last step is the interpretation and evaluation of the mined data, this step determines the success level of the purposed model and documents the knowledge discovery for future references. (Tomar & Agarwal, 2014). To effectively manage the churn prediction problem, different researches have implemented different data mining algorithms which is extensively reviewed in the following section.

4. Data Mining Techniques and Algorithms for Customer Churn Prediction

Azeem, Usman & Fong (2017) used fuzzy algorithm to predict customer churn in telecommunications industry. Recently, the fuzzy algorithm has gained popularity in telecommunications industry to conduct customer churn prediction as normally the data required for churn prediction is very noisy and fuzzy algorithm is very effective in

managing noisy data. The dataset for this research was obtained from a telecommunications company based in South Asia. SVM, linear regression, C4.5, gradient boosting, Artificial Neural Network (ANN), random forest, AdaBoost and neural network were used for performance comparison with fuzzy algorithm. The fussy algorithm utilized numerical variables and domain knowledge in the variable selection stage. The previous researches conducted on customer churn prediction, mainly focused only on post-paid customers. However, with the implementation of fuzzy algorithm, call details record variable of prepaid customers was also considered in this research and the effectiveness of predicting prepaid customer churn was highly measured. The performance evaluation results showed that the fuzzy algorithm was more accurate and effective as compared to other algorithms with Area Under the Curve (AUC) rate of 98% and True Positive score of 0.68.

Ismail et al. (2015) proposed the Multilayer Perceptron (MLP) algorithm to predict customer churn. The dataset for this research was obtained from one of the telecommunications company in Malaysia. The MLP algorithm was compared with two statistical algorithms which are logistic regression and multiple regression. The performance of the algorithms was measured based on sensitivity, accuracy and specificity. The variables used for this analysis were customer demographic variables, customer relationship variables, billing variables and usage variables. MINITAB software was utilized for constructing the multiple regression and logistic regression algorithm and MATLAB software was utilized for constructing the MLP neural network algorithm. The performance evaluation result showed MLP neural network as the best algorithm for customer churn prediction with an accuracy of 91.28%. Meanwhile, multiple regression produced accuracy prediction of 78.84% and logistic regression produced prediction accuracy of 75.19%. Additionally, the sensitivity showed 93.5% for MLP neural network, 83.98% for multiple regression and 71.80% for logistic regression. Furthermore, the specificity of showed 88.28% for MLP neural network, 71.90% for multiple regression and 79.66% for logistic regression. The overall finding result suggested that MLP neural network algorithm is the most suitable approach for customer churn prediction as compared to conventional statistical algorithms.

Vafeiadis et al. (2015) conducted customer churn prediction by comparing the commonly used data mining algorithms such as ANN, decision trees, SVM, Naïve Bayes and logistic regression. These algorithms were then compared with their enhanced versions in order to boost their performance further. The main objective of this study was to evaluate the suitability of the state of data mining algorithms for the customer churn problem. The telecommunications dataset for this research was obtained from UCI Machine Learning Repository. The dataset was transformed to MLC++ format in order to enable the performance comparison of the algorithms and their enhanced versions. The evaluation was performed using Monte Carlo simulation at various settings of each algorithm. The enhanced algorithm methodology used for this study was AdaBoost.M1. The objective of this methodology is to boost the performance of the algorithms'

performance through a combination of weak algorithms, such as decision tree and ANN, in order to obtain better decision. Monte Carlo simulation for this dataset revealed that the combination of ANN and decision tree algorithms as well as SVM achieved the highest accuracy compare to others. The enhanced algorithm methodology is proven to be very useful for telecommunications industry as it can improve the rate of accuracy for customer churn problems, especially for large customers' data.

Whilst most researches on customer churn prediction focused more on data mining algorithms than identifying the most important variables to be used for these algorithms, Adwan et al. (2014) conducted a research on customer churn prediction by applying a MLP algorithm. MLP neural network not only predicts customer churn, but also provides insights on the significance of each input variable used for customer churn prediction. The MLP neural network has an extensive range of implementation for classification and prediction problems in business and industrial domains. This research used real data which was provided by Umniah, a major telecommunications network in Jordan. Confusion matrix was used to assess the MLP neural network algorithm. Additionally, k-cross validation with k=5, was implemented in order to provide a better understanding on how well the algorithms would perform when new data is presented. Furthermore, to understand the importance of each variable towards the churn, change of error (COE) and weights contribution on the network were implemented. COE grades input variables, fitting to the variation of some quality measures when each input is erased from the dataset in the training process. Weight contribution in the network is applied in order to obtain comparative frequencies for the variables. There were three variables that appear on both COE and weight contribution in the network, namely, total monthly fees, total of international outgoing calls and 3G services. These findings can be valuable to the customer relationship management department in the telecommunications industry as it allows them to plan more effective strategies to manage customer churn. Nevertheless, there is a noticeably disadvantage of the COE method in which the noise and inconsistency in the training dataset can minimize its consistency.

Keramati et al. (2014) used a dataset obtained from an Iranian mobile company for their research. They compared the performance of four algorithms which are decision tree, ANN, k-nearest neighbours and support vector machine. After analysing the algorithms' performance and understanding their special structures, a new algorithm was proposed in order to improve the evaluation metric value. The proposed algorithm, which was named as "Best Hybrid Methodology" in this study, showed above 95% accuracy for recall and precision were easily attained. Additionally, compared to the four algorithms, the new proposed algorithm also has better tuning parameter that can be easily deployed to predict as per the analyst's requirement. If an analyst needs to analyse which customers are likely to churn and which customers are likely to have the least tendency to churn, the algorithm can be tuned to produce the desired analysis. Furthermore, this study also introduced a new dimensionality reduction techniques in order to extract the most important set of variables to be involved in data mining process.

Kim, Jun & Lee (2014) conducted customer churn prediction analysis by presenting an innovative variable, called the network variable, which was acquired from network analysis. The network variable was calculated through the propagation process. The spreading activation was employed in the propagation process. Additionally, the network variable as well other customers' personal details variables were used as the input. Unlike other researches on customer churn prediction, this research actually measured the numerous features of all existing churners who influence potential churners. The dataset was acquired from a telecommunications company, which involved the call detail records data and customers' personal detail data. Logistic regression and ANN were employed in this research. To accomplish better performance evaluation of the algorithms, two hypotheses were applied. The first hypothesis was that existing churners have an influence on the potential churners in the same group. The second hypothesis was that churn date and centrality influence the existing churners in leading potential churners. Group identification was employed to the network variable and the propagation process was conducted in the same group to investigate the first hypothesis. The result showed that the prediction performance of both algorithms after group detection was still similar to non-group detection result, but the propagation process time was faster. The second hypothesis performance was evaluated by adjusting the churners' existing energy considering the centrality and churn date. The result showed that adjusting the churners' existing energy enhanced the prediction, and both algorithms also performed better when the existing energy was adjusted.

Abbasimehr, Setak & Soroor (2013) proposed two stage structure for customer churn prediction. The first stage is the identification stage and the second stage is the data mining algorithm stage. In this research, social network based variables were included along with conventional variables from customer relationship management database. The neuro fuzzy algorithm was used in the data mining process. The effectiveness of the two types of neuro fuzzy algorithms which are locally linear neuro-fuzzy (LLNF) with locally linear model tree (LoLiMoT) and the adaptive neuro-fuzzy inference system (ANFIS) were explored. Additionally, the two neuro fuzzy algorithms were compared with two neural network algorithms, namely, MLP and radial basic function (RBF). The dataset used for this research was obtained from the Teradata Centre at Duke University. The customers' variable in the dataset was clustered by using k-means algorithm and the customers' data that were on the top cluster was used in the data mining algorithm stage. The result showed that neuro fuzzy algorithms performed better compared to the neural network algorithms.

Brandusoiu & Toderean (2013) performed a customer churn prediction analysis by utilizing dataset obtained from Department of Information and Computer Science of University of California. IBM SPSS (Statistical Product and Service Solutions) was used as the technology to mine the data. There were no missing values identified from the dataset and there was a perfect correlation (R=1) between some variables SVM algorithm was

used as a data mining algorithm for this research. As a requirement of SVM algorithm, The "yes-es" in the dataset was cloned by balancing and by boosting the training set in order to have an equal distribution with the "no-es". The SVM algorithm was trained by using four kernel functions which are RBF, Linear, Polynomial (POL) and Sigomid (SIG). The result showed that POL performed the best out of the four kernels with overall accuracy of 88.56%. However, overall the four kernels predicted around 80% percent accuracy. Based on the predictors that have higher significance in scoring the kernel performance, the customer relationship management can plan different marketing approach to retain potential churners. In spite of this, the study was only conducted with only one algorithm, comparison between few algorithms would reveal different results and higher accuracy rate could have been achieved.

Kirui et al. (2013) added a new subcategory of variables in order to enhance the accuracy of customer churn prediction in the telecommunications industry. The new subcategory of variables which are call pattern description variables, call pattern changes description variables, contract-related variables, were originated from the statistical traffic data and customers profile data. The dataset used for this study was obtained from European telecommunications company. Bayesian Network and Naïve Bayes were used to assess the performance of predictive significance of the added new variables. The evaluation results were then compared using C4.5 decision tree algorithm. True churn rate and false churn rate were achieved from the algorithms implementation. Bayesian Network and C4.5 decision tree showed a strong distinction of significance of every added subset of variables. However, Naïve Bayes showed that most of the added subset of variables performed almost evenly. In spite of this, C4.5 decision tree showed better accuracy performance on the dataset used. Yet, according to (Kirui, 2013), the false churn rate and the true churn rate are better measures as compared to accuracy rate for the customer churn prediction case. Nevertheless, this study did not address the class imbalance problem of the initial datasets. Hence, the minority class instances were not clearly identified even though they might actually achieve high overall accuracy.

Qureshi et al. (2013) tested logistic regression, decision tree (including CHI-squared Automatic Interaction Detector (CHAID), Exhaustive CHAID, Classification and Regression Trees (CART) and Quick, Unbiased and Efficient Statistical Tree (QUEST)), ANN and k-mean clustering on the data set acquired from an online source (http://www.customer-dna.com/). Additionally, re-sampling method was used in this research to tackle a very general problem in telecommunications industry which is class imbalance problem. Recall, precision and F-measures were used to evaluate the performance of different prediction algorithms. Furthermore, to determine the important variables as predictors, p-value below 0.05 and Pearson correlation were used. The re-sampled dataset gave unbiased result compare to the dataset with class imbalance. After implementation of all the algorithms, the results showed that Exhaustive CHAID, a variant of decision trees, was the most accurate for the dataset with accuracy of 70%. To further boost the accuracy, five new variables derived from some of

the existing variables, were introduced to the dataset. The overall accuracy after including the new variables was indeed increased to 75.4%.

Chen, Fan & Sun (2012) proposed a data mining algorithm named the Hierarchical Multiple Kernel Support Vector Machine (H-MK-SVM) for both statistical and longitudinal behavioural data. The telecommunications dataset used for this study was from a mobile services company provided by the Centre for Customer Relationship Management of Duke University. To train the H-MK-SVM, a three-stage algorithm were established and applied. The longitudinal behavioural data for this study was not transformed in the training process. It was used directly as an input for the data mining algorithm without any clustering step as commonly done in standard contexts. Additionally, the training procedure of the H-MK-SVM was also a variables selection procedure because the spare non-zero coefficient correlation to the selected variables. The H-MK-SVM algorithm constructed a classification formula by calculating the coefficients of both longitudinal and statistical data. To compare the H-MK-SVM algorithm's performance, ten other algorithms, namely, the MK-SVM, SVM, least squares SVM (LS-SVM), decision tree, logistic regression, feed-forward ANN, RBF Neural Network, random forest, AdaBoost, and the proportional hazard model (Cox), were used in this study. The performance result utilizing the Lift and the AUC measures revealed that the H-MK-SVM performed effectively on both imbalance as well as balance class data compare to other algorithms. A collaborative data mining algorithm such as H-MK-SVM is a developing framework especially for customer churn prediction in the telecommunications industry as it performs effectively with large volume of data.

Huang, Kechadi & Buckley (2012) focused on deriving new sets of variables from the initial variables to predict customer churn. The new sets of variables which are aggregated call details, account information, Henley segmentation, bill information, dial types, payment information, service information, line information, complaint information, among others, were compared to the existing variables. The dataset itself, was obtained from a telecommunications company in Ireland. The prediction was conducted by using seven different algorithms which are linear classifiers, decision trees, logistic regression, ANN, SVM, Data Mining by Evolutionary Learning (DMEL) and Naïve Bayes on both derived set of variables and the existing set of variables. The result revealed that the new proposed set of variables were more effective for the churn prediction as compared to the existing set of variables. Furthermore, the result showed that decision tree and SVM with a low ratio were compatible for predicting the true churn rate and the false churn rate. Additionally, DMEL algorithm was revealed to be the weakest algorithm for customer churn prediction as it was not compatible for large dataset with high dimension. However, this study did not include enough derived variables in the data mining process which could improve the prediction accuracy.

Phua et al. (2012) predicted customer churn as well as customers win-backs in the near future based on datasets obtained from a telecommunications company. The customers involved were individuals and small medium enterprises (SMEs). The

prediction objectives for this study was not only on identifying the likely churners with good accuracy but also identifying these churners within a short time period of consequent three months. An appropriate computational strategy was used in order to find fixed patterns to predict churners and possible win-back. Additionally, to attain dependencies for the stronger original variables, a few derived variables were created. The class imbalance correction was also performed by using under-sampling and over-sampling strategies in order to enhance the accuracy for minority class instances. Tree classifiers algorithms were used in this study, namely, ADTree, decision stump, RepTree, J48, TreeLMT, random forest, bagging + decision stump, bagging + simple cart, simple cart. Additionally, Naïve Bayes and classification via regression algorithms were also used as comparison. The performance evaluation result of the algorithms showed that random forest and simple cart performed the best with the highest accuracy prediction as compared to other algorithms.

Shaaban et al. (2012) proposed a model which consists of six stages which are identify problem domain, data selection, investigate data set, classification, clustering and knowledge usage to perform data mining for customer churn prediction, as shown in Figure 3; in which, the classification phase created two types of customers, namely, churners and non-churners. The clustering phase created three clusters which are utilized for evaluation of retention strategy for further research. The clustering step is not limited to three clusters only, the number of clusters is subject to the type of knowledge usage. The knowledge usage obtained the clusters in order to provide solution to retain each type of churners. Churners can be grouped according to many criteria based on customer dissatisfaction and/or profitability. The dataset used for this study was obtained from an anonymous mobile service provider. The dataset was separated into a training set (80%) and a test set (20%). decision tree, ANN, SVM were used in this research. The confusion matrix showed that ANN and SVM predicted 83.7% of accuracy while decision tree predicted 77.9% of accuracy for this dataset. (Shaaban et al., 2012).

Umayaparvathi & Iyakutti (2012) conducted churn prediction analysis with dataset obtained from a data mining competition, PAKDD-2006. The dataset was aggregated into training data and test data for six months' period, as prediction models need historical data of customer behaviour for a specific period of time in order to predict the customer behaviour in the future. Most relevant variables for algorithm implementation were selected. The variables later were categorized into 4 major groups which are customer demography, customer care service, bill and payment and call detail record. These variables are trained in order to implement the decision tree algorithm and ANN algorithm. The performance of the algorithms was tested based on the counts of test data. A confusion matrix was established for both algorithms based on demographic variables in order to find the predictive accuracy and error rate of the algorithms. The confusion matrix revealed that decision tree algorithm surpassed the ANN algorithms for customer churn prediction, as the error rate was lower by 0.4 % and the accuracy was higher by

0.4% as well compare to ANN algorithms based on the datasets used. However, this study did not combine other variable groups in the testing process. The result could have been different and could also predict better accuracy if other variable groups were tested as well.

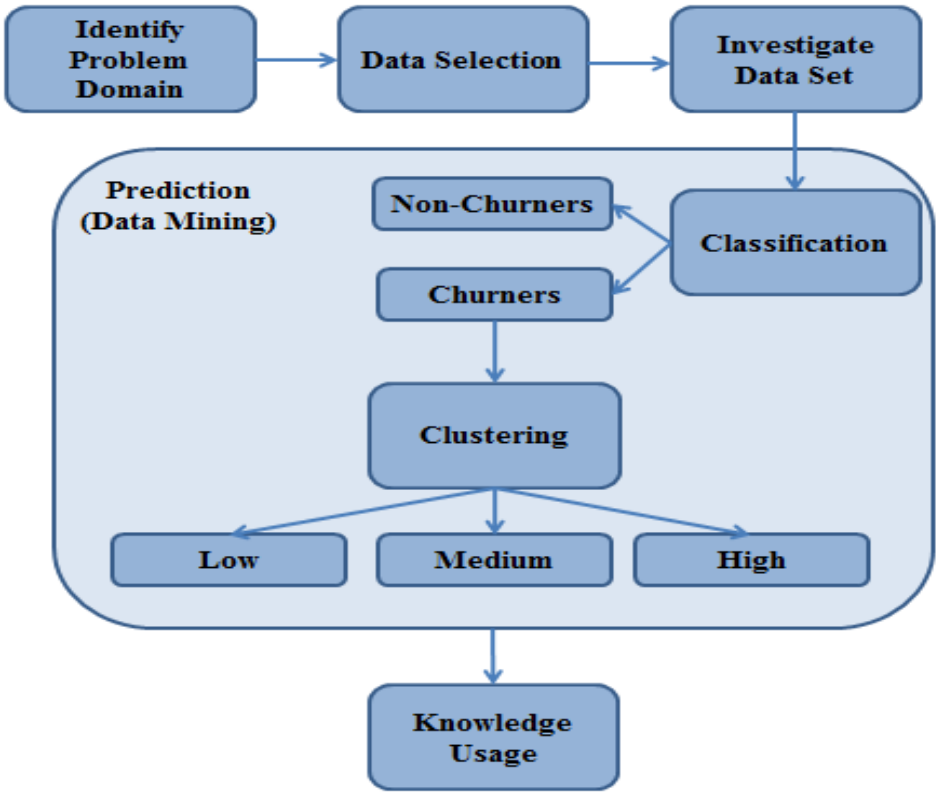

Figure 3: Proposed Model for Customer Churn Prediction (Shaaban et al., 2012)

5. Conclusion

Customer churn prediction is one of the most crucial missions in telecommunications industry. The aggressive market of telecommunications industry has forced the service providers to employ the best data mining algorithms which produce most accurate prediction in order for them to stay competitive in the market.

Many data mining algorithms have been reviewed and SVM, Bayes Network, decision tree, ANN, amongst others, were found to be the most popular algorithm in customer churn prediction. Some researchers also have combined few algorithms and established an innovative algorithm in order to produce better accuracy rate. Additionally, the enhanced algorithm methodology such as AdaBoost was found to be very valuable, as it can enhance the accuracy performance of weak algorithms. However, the accuracy performance of each algorithm differs in every research. This is due to the different dataset used and the different input variables chosen for the experiment.

Most of the literature focused more on data mining algorithms, but only a few of them focus on distinguishing the important input variables for churn prediction to be used for data mining algorithms implementation. Additionally, only noticeably one literature that had actually combined social network based variables in the input variables for data mining algorithms implementation. Moreover, the class imbalance problem was found to be not addressed on some of the literatures.

In future, this literature review can be used as a foundation for customer churn prediction analysis. Few of the most prominent algorithms can be tested on new dataset. Additionally, external environment factors such as social factors can be considered as one of the input variables for churn prediction as it is proved to have an impact on the accuracy result.

References

Abbasimehr, H., Setak, M. & Soroor, J. (2012) A Framework for Identification of High-Value Customers by Including Social Network Based Variables for Churn Prediction Using Neuro-Fuzzy Techniques. International Journal of Production Research. 51(4). p.1279-1294.

Adwan, O., Faris, H., Jaradat, K., Harfoushi, O. & Ghatasheh, N. (2014) Predicting Customer Churn in Telecom Industry Using Multilayer Preceptron Neural Networks: Modeling and Analysis. *Life Science Journal*. 11(3). p.1-7.

Azeem, M., Usman, M. & Fong, A.C.M. (2017) A Churn Prediction Model for Prepaid Customers in Telecom Using Fuzzy Classifiers. *Telecommunications Systems*. p.1-12.

Brandusoiu, I. & Toderean, G. (2013) Churn Prediction in the Telecommunications Sector Using Support Vector Machines. *Annals of the Oradea University Fascicle of Management and Technological Engineering*. 22(1). p.19-22.

Chen, Z.Y., Fan, Z.P. & Sun, M. (2012) A Hierarchical Multiple Kernel Support Vector Machine for Customer Churn Prediction Using Longitudinal Behavioral Data. *European Journal of Operational Research*. 223(2). p.461-472.

Huang, B., Kechadi, M.T. & Buckley, B. (2012) Customer Churn Prediction in Telecommunications. *Expert Systems with Applications*. 39(1). p.1414-1425.

Ismail, M.R., Awang, M.K., Rahman, M.N.A. & Makhtar, M. (2015) A Multi-Layer Perceptron Approach for Customer Churn Prediction. *International Journal of Multimedia and Ubiquitous Engineering*. 10(7). p.213-222.

Keramati, A., Marandi, R.J., Aliannejadi, M., Ahmadian, I., Mozzafari, M. & Abbasi, U. (2014) Improved Churn Prediction in Telecommunications Industry Using Datamining Techniques. *Applied Soft Computing*. 24. p.994-1012.

Kim, K., Jun, C.H. & Lee, J. (2014) Improved Churn Prediction in Telecommunications Industry by Analyzing a Large Network. *Expert Systems with Applications*. 41(15). p.6575-6584.

Kirui, C., Hong, L., Cheruiyot,W. & Kirui, H. (2013) Predicting Customer Churn in Mobile Telephony Industry Using Probabilistic Classifiers in Data Mining. *International Journal of Computer Science Issues*. 10(2). p.165-172.

Phua, C., Cao, H., Gomes, J.B. and Nguyen, M.N. (2012) Predicting Near-Future Churners and Win-Backs in the Telecommunications Industry. arXiv preprint arXiv:1210.6891.

Qureshi, S.A., Rehman, A.S., Qamar, A.M., Kamal, A. & Rehman, A. (2013) *Telecommunications Subscribers' Churn Prediction Model Using Machine Learning*. In Digital Information Management (ICDIM), 2013 Eighth International Conference on. 10th September 2013. IEEE. p. 131-136.

Shaaban, E., Helmy, Y., Khedr, A. & Nasr. M. (2012) A Proposed Churn Prediction Model. *International Journal of Engineering Research and Applications*. 2(4). p.693-697.

Tomar, D. & Agarwal, S. (2014) A Survey in Pre-Processing and Post-Processing Techniques in Data Mining. *International Journal of Database Theory and Application*. 7(4). p. 99-128.

Umayaparvathi, V. & Iyakutti, K. (2012) Applications of Data Mining Techniques in Telecom Churn Prediction. *International Journal of Computer Applications.* 42(20). p.1-5.

Vafeiadis, T., Diamantaras, K.I., Sarigiannidis, G. & Chatzisavvas, K.C. (2015) A Comparison of Machine Learning Techniques for Customer Churn Prediction. *Stimulation Modelling Practice and Theory*. 55. p.1-9.

Verbeke, W., Dejaeger, K., Martens, D., Hur, J. & Baesens, B., (2012) New Insights into Churn Prediction in the Telecommunications Sector: A Profit Driven Data Mining Approach. *European Journal of Operational Research*. 218. p.211-229.

Journal of Applied Technology and Innovation (eISSN: 2600-7304)
vol. 1, no. 2, (2017), pp. 41-48

A Review on Existing Active Noise Reduction for Industrial HVAC system

Krishna A/L Ravinchandra
School of Engineering
Asia Pacific University of Technology & Innovation
57000 Kuala Lumpur, Malaysia
E-mail: k9999_5407@hotmail.com

Thang Ka Fei
School of Engineering
Asia Pacific University of Technology & Innovation
57000 Kuala Lumpur, Malaysia
E-mail: dr.ka.fei@apu.edu.my

Lau Chee Young
School of Engineering
Asia Pacific University of Technology & Innovation
57000 Kuala Lumpur, Malaysia
E-mail: dr.laucheeyong@apu.edu.my

Abstract - The existing industrial active noise reduction (ANR) for heating ventilating and air conditioning (HVAC) system has increased in terms of demand. It is a concern due to the amount of noise generated by industrial HVAC system which can be a problem to the surrounding, where noise pollution is always taken into consideration for healthier environment. In this paper, the industrial HVAC system is appraised based on implementation of the system, placement of sensor and actuators in addition there is also review on algorithms. Moreover, the active noise reduction system is introduced to reduce the lower frequency noise range, this helps in reducing the noise pollution created by the industrial HVAC system. Besides, the placement of sensor and actuators plays a role in determining the suitable distance in achieving best noise reduction position. Furthermore, an evaluation is done on Filtered-XLMS and Kalman Filter active noise cancellation algorithm, the results show that the FxLMS algorithm has much better performance rate compared to Kalman filter in terms of convergence. Lastly, multiple ways of implementing the ANR system is also evaluated which indicates that there are many methods in reducing the noise pollution.

Index Terms - Active noise reduction (ANR), Heating ventilating and air conditioning (HVAC), Filtered least mean squared (FxLMS) filter, Kalman filter

1. Introduction

In this recent era of technology, the industrial companies have increased the manufacturing of the Heating, Ventilating, and Air Conditioning (HVAC) system. The HVAC system assists in cooling and heating the temperature in a room depending on a specific location where the system is placed. The the air-conditioning system comprises seven different processes which are heating, cooling, humidifying, dehumidifying, cleaning, ventilating and air movement, therefore the importance and necessities varies differently (Lopes, Gerald & Piedade, 2015). The usage of HVAC system is wide and is seen to be used in offices, household or even factories. The HVAC system creates an acoustical (noise) environment which affects the rooms surrounding. With the growth of the industrial tools, the noise problem has been promptly increasing. The conventional method used in reducing the noise is via passive noise reduction such as mechanical barriers, enclosures etc. Noises are constituted in different frequency range, the passive noise reduction eliminates in the range of frequency between 4 kHz to 10 kHz but unable to effectively cancel the lower frequency noise of less than 4 kHz because of its unrealistic size of barrier required for noise reduction (Kasbekar, Wisler & Panahi, 2011). The size of the barrier required for noise reduction can be re-evaluated where the thickness of the material can be increased because the low frequency has long wavelengths, but the material will be costly and impractical to implement (Deveneni, Panahi & Kasbekar, 2011).

Therefore, active noise control (ANC) is introduced to control the lower frequency noise effectively. The best way to define ANC is to destructively interfere with undesirable noise with generated secondary sound signal (Rajesh, Jeevamalar & Jancirani, 2012). The ANC is basically a system which assist in removing the undesirable noise coming from the surrounding.

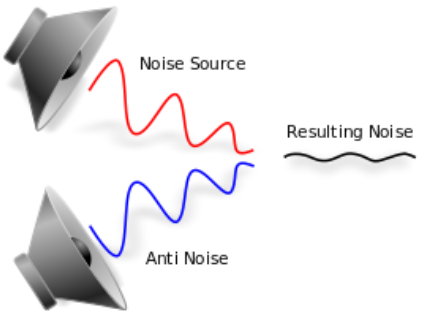

Figure 1: Active Noise Cancellation (Swain, 2013).

Figure 1 shows how the active noise cancellation works, where the noise source is represented with a red sine wave line and the anti-noise is represented with blue wave line. When both noises are colliding with each other it results to the minimal wave line which is represented in black. Fundamentally, before the anti-noise is generated, there is a process for computation, where a sensor will identify the noise source, then send it to the signal for a calculation which creates the anti-noise wave in an opposite phase of the noise source signal, then the anti-noise is projected using the loudspeaker to control the undesirable noise source signal. The ANC system cannot really produce a pin drop silence but it reduces to noise to a substantial amount. Thus, ANC system is best suited to be used in creating a more eco-friendly HVAC system. Moreover, it is also preferred that the efficiency of the system to be increased which allows to have a better current and power saving for the HVAC system.

2. Implementation of the system

The application of the ANR system here was studied based on three different papers, which basically evaluates the implementation for the ANC system using different approach to help in reducing the noise pollution. Where, in (Devineni, Panahi & Kasbekar, 2011) discussed the methods and algorithms applied is the Feedback ANC and Multiple Input & Multiple Output Period Aware Linear Prediction (MIMO-PALP), it is applied for the MIMO structure which enables a bigger and effective cancelling zone at the surrounding of the compressor surface and the results attained were compared between Noise Attenuation Level of MIMO PALP and Feedback ANC, this is to determine the performance and the MIMO Feedback ANC has a better performance compared to MIMO PALP.

Khan et al., proposed a system based on Virtual Instrument in Reality (VISIR) concept using National Instruments (NI) PXI system which concern the multi-channel measurement and control of the sound field. The approach proposed by the authors is a more comprehensive analysis method for studying the properties of the acoustic modes inside the air ducts in order to achieve better results (Khan et al., 2014).

In (Wisler & Panahi, 2013), the analyses and design of a real time active noise control system for cancelling the acoustic noise generated by an HVAC unit within a 3-dimensional enclosure, where the author used Open Media Application Platform as a controller with a feedback ANC algorithm programmed to it, it could achieve a stable and consistent attenuation of the noise from a commercially available HVAC unit in excess of 14 decibels.

3. Placement of Sensor & Actuators

The various position of sensor and actuators can be a great factor in reducing the noise in industrial HVAC system, therefore an evaluation is done on the placement of sensor and actuators. As proposed by Jung et al., a 2x2x2 multi-channel active noise control (ANC) technique is used for the active noise barriers (ANB), where the application is then placed at numerous location which applies a single-channel FXLMS to attain the best noise reduction performance, the outcome of the application is based on two different scenario where one application is tested for multi-frequency tone noise source and another is for an air compressor, Table 1 shows the results for the second scenario, which concludes that the closer the placing of the ANB the better the performance of ANC and improved noise reduction (Jung et al., 2014). Figure 2 shows the ANRS schematic block diagram and structure.

Table 1: The noise reduction performance at three different distance (Jung et al., 2014).

Distance / Condition	0.1m	1m	2m
ANC on (dB)	54.3	52.8	53.9
ANC off (dB)	63.8	57.9	57.4
Noise reduction (dB)	9.5	5.1	3.5

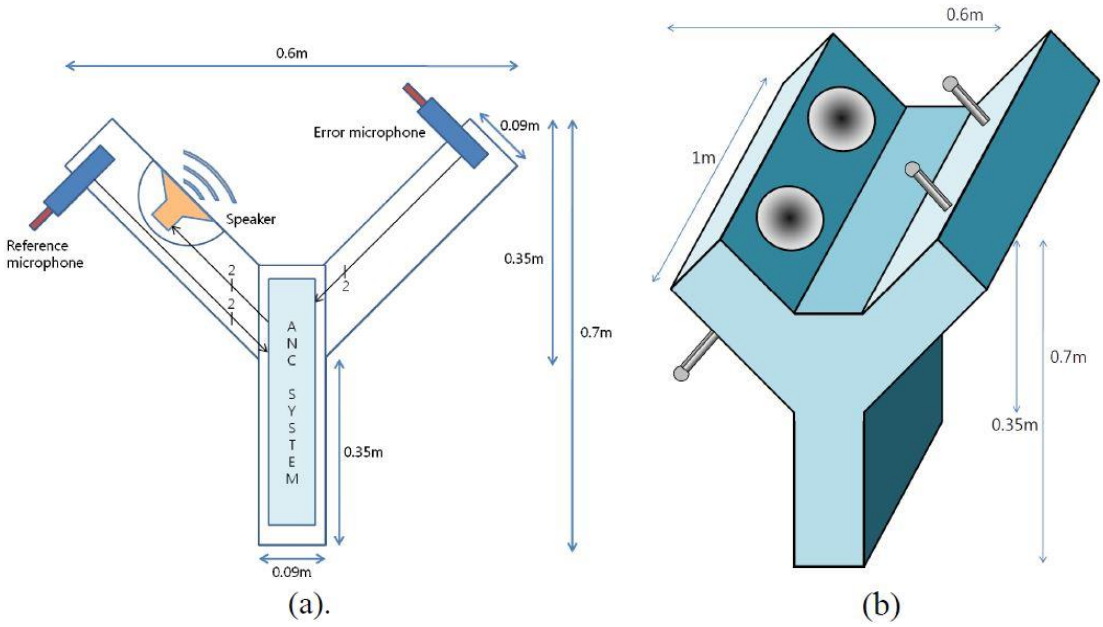

Figure 2: (a) Schematic block diagram of ANRS, (b) The structure of ANRS (Jung et al., 2014).

The available techniques that aims at the displacement of error microphone in ANC system was discussed by (Sheth & Ardekani, 2015). The author conducted a comparison about the various types of displacement, with three different types of approach towards the remote ANC, involving the moving microphone, remote ANC and virtual microphone, the conclusion from this approach was that numerous technique has its own advantages and limitations in creating an efficient quite zones by using ANC.

4. Algorithms

The comparison of FXLMS and Kalman filter is assessed in this paper. Using a soft threshold based approach on FXLMS algorithm is proposed by authors, MATLAB simulation is done to compare the convergence rate and stability of modified FXLMS and Akthar's algorithm, the reason of using soft thresholding is to enhance the robustness of the Sun's algorithm, the outcome showed that when large amplitude of impulse noise is present, the proposed algorithm improves the convergence rate and its stability (Saravanan & Santhiyakumari, 2015).

According to (Ardekani & Abdulla, 2011) a simple configuration for realization of ANC system is presented which shows how the silence zone is created, it is distributed to five fragments which are ANC physical mechanism, digital electronic control system, secondary path constraint, reference signal measurement constraint and adjustment of ANC controller. The authors also revised the adaptive ANC where it can be simply adjusted to a system identification framework without having the knowledge of the primary and secondary path. According to the research, the FXLMS algorithm is considered to be a vital adaption algorithm in ANC where it is usually considered just to evade mathematical difficulties.

In (Gaur & Gupta, 2016) analyses two parameter which is the improvement in FXLMS algorithm and second is application based review. Feed forward and feedback are the two basic approach for ANC system, the FXLMS algorithm and its application is done based on review which is shown in **Table 2**, where it discussed the improvement of the algorithms based on convergence rate, complexity, etc.

Table 2: Summary on Review of FxLMS Algorithm (Gaur & Gupta, 2016).

Summary No.	Algorithm	Findings
1	FxLMS	Simple computation and convergence slower than conventional LMS
2	MFxLMS	Convergence better than FxLMS but involves heavy computation.

3	MFxLMS1 and MFxLMS2	Convergence is better than FxLMS (equal to MFxLMS) but computation is lesser than MFxLMS
4	CFxLMS	Developed for broadband ANC. Here weights of filter are given specific upper and lower bound. Convergence rate has increased.
5	Variable threshold based FxLMS	Convergence rate has increased with little increase in computation
6	Convex combination based FxLMS	Convergence rate is high but with heavy computation, since it involves parallel combinations
7	VSS FxLMS	Developed for Narrow Band ANC, convergence rate has increased (nearly equal to FxRLS) with little computations. Good for both stationary and non-stationary noise environment. Cost and computation complexity lies
8	Data reusability based FxLMS	Used for impulse noise, it normalizes step size and so improves convergence rate
9	VSS FxLMS with variable tap length	Maintains good convergence rate even in case of long tap length applications
10	FxWLMS & FxLMLS	Developed for ANC application to hearing aid and found effective in feedback cancellation in presence of outliers.

Two analysis is done for Kalman Filter. The paper in (Gabrea, 2012) researches speech signal corrupted by an additive noise for processing where autoregressive (AR) is introduced as a way to overcome corrupted speech signals. According to the author, an algorithm for minimal realization of the model was proposed, by minimizing the mean squared error in the approximation of the speech signal by Kalman Filtering, outcome of this investigation is that the effectiveness of this method is tested using natural speech signal sampled at the frequency of 8kHz corrupted by a Gaussian white noise.

Another Kalman Filter paper was researched on an adaption of the KF to multi-channel ANC was done, the authors proposed an algorithm based on Kalman filter with a random walk space state model (Lopes, Gerald & Piedade, 2015). According to the author, MFx Multi-channel Kalman (MFxMK) comes from the combination of MFx structure with Kalman Filter, this algorithm is less complex to singular or close to singular autocorrelation matrices than the Recursive Least Square and its' results. However, it circulates the covariance matrix of the state approximation.

5. Conclusion

This paper conducted a review on active noise reduction for industrial HVAC system, where the implementation of the system, placement of sensor and actuators and algorithms is assessed. And based on this review, future works are taken into consideration in enhancing the ANR system to achieve the best performance for the ANR system for industrial HVAC system, which is also helps in keeping a healthier and noise pollution free environment. Therefore, the implementation of the system can be a standalone ANR system which can be easily implemented into the industrial HVAC system, the placement of sensor and actuators determines the best position for microphone and speakers to be placed in the industrial HVAC to achieve the best noise reduction and the algorithm will determine the performance of the ANR process. In conclusion, future recommendation for the ANR system could be implemented and approached using FXLMS algorithm and raspberry pi.

References

Ardekani, I. T. and Abdulla, W. H. (2011). FxLMS-based active noise control: A quick review. In Asia Pacific Signal and Information Processing Association Annual (APSIPA) Summit and Conference, Xi'an, China. pp. 1-11.

Deveneni, N., Panahi, I. and Kasbekar, P. (2011). *Predictive multi-channel feedback active noise control for HVAC systems*. In Electro/Information Technology (EIT), 2011 IEEE International Conference. pp. 1-5.

Gabrea, M. (2012). *A single microphone noise canceller based on an adaptive Kalman filter*. In Electrical & Computer Engineering (CCECE), 2012 25th IEEE Canadian Conference. Canada. pp. 1-4.

Gaur, S. and Gupta, V.K. (2011). A Review on Filtered-X LMS Algorithm*. International Journal of Signal Processing Systems*. Vol. 4, No. 2. pp. 172-176.

Jung, T.H., Kim, J.H., Kim, K.J. and Nam, S.W. (2011). *Active noise reduction system using multi-channel ANC*. In Control, Automation and Systems (ICCAS), 11th International Conference. Korea, Wednesday 26th October to Saturday 29th October 2011. pp. 36-39.

Kasbekar, P., Wisler, A. and Panahi, I. (2012). *Real time reduction of HVAC noise using a FPGA.* In Southeastcon Proceedings of IEEE. pp. 1-5.

Khan, I., Zmuda, M., Konopka, P., Gustavsson, I. and Hakansson, L. (2014). *Enhancement of remotely controlled laboratory for Active Noise Control and acoustic experiments*. In Remote Engineering and Virtual Instrumentation (REV), 11th International

Conference. Portugal, Wednesday 26th February to Friday 28th February 2014. pp. 285-290.

Lopes, P.A., Gerald, J.A. and Piedade, M.S. (2015). *The Random Walk Model Kalman Filter in Multichannel Active Noise Control*. In IEEE Signal Processing Letters, 22(12). pp. 2244-2248.

McDowall, R. (2007). Fundamentals of HVAC systems: SI edition. Academic Press.

Rajesh, M., Jeevamalar, J. and Jancirani, J. (2012). *Active noise reduction of automotive HVAC system using filtered LMS [Part-1 sound measurement].* In IEEE-International Conference on Advances in Engineering, Science and Management (ICAESM). India, Friday 30th to Saturday 31st March 2012. pp. 204-205.

Saravanan, V. and Santhiyakumari, N. (2015). *A modified FXLMS algorithm for active impulsive noise control*. In 2015 IEEE International WIE Conference on Electrical and Computer Engineering (WIECONECE). Bangladesh, Saturday 19 December to Sunday 20 December 2015. pp. 222-226.

Sheth, D., & Ardekani, I. T. (2015). Error signal measurement in active noise control systems. In APSIPA Newsletter. http://www.apsipa.org/doc/APSIPA_Newsletter_Issue_10.pdf

Swain, A. (2013). Active Noise Control: Basic Understanding. Research Gate. pp. 1-19.

Wisler, A. and Panahi, I.M. (2013). *Implementation of a real-time feedback active noise control system to cancel noise within a 3-dimensional enclosure*. In 2013 8th International Symposium on Image and Signal Processing and Analysis (ISPA). Italy, Wednesday 4th September 2013 to Friday 6th September 2013.pp. 651-654.

Journal of Applied Technology and Innovation (eISSN: 2600-7304)
vol. 1, no. 2, (2017), pp. 49-57

Real-Time Indoor Tracking

Fawwad Ahmed Shabbir
School of Engineering
Asia Pacific University of Technology & Innovation
57000 Kuala Lumpur, Malaysia
Email: fawwad93@gmail.com

Lai Nai Shyan
School of Engineering
Asia Pacific University of Technology & Innovation
57000 Kuala Lumpur, Malaysia
Email: nai.shyan@apu.edu.my

Veeraiyah Thangasamy
School of Engineering
Asia Pacific University of Technology & Innovation
57000 Kuala Lumpur, Malaysia
Email: dr.veeraiyah@apu.edu.my

Abstract – In this paper, a development of an autonomously navigating robot car, where the robot car would autonomously navigate towards a user-defined destination location, using a UWB based indoor localization system is described. For accurate 3-axis indoor localization, Pozyx indoor localization system was used, which uses Decawave DWM1000 chips for UWB communication. Time of Arrival distance calculation method, and Least Linear Square Algorithm were used for calculation of location co-ordinates of the Pozyx mobile tag, which was integrated with the robot car. A GUI was developed on Unreal Engine, to provide real-time visual representation of location of the robot car. The user was allowed to set destination co-ordinates using pick and place in the developed Unreal Engine GUI. The autonomous navigation robot was programmed, using an Arduino Uno microcontroller. Minimum accuracy for indoor static indoor localization achieved for the system for all the 3-axis was 90% in the 3D environment. Furthermore, a minimum accuracy for autonomous navigation towards destination co-ordinates achieved for the system in the x-axis and y-axis was 87%. A minimum navigation speed of 35cm/s was achieved by the robot car. For future works, autonomous navigation in 3-axis could be performed. Communication between the robot car and controlling PC could be performed using wireless methods, to improve range of operation of the robot.

Index Terms - Indoor tracking, Autonomous navigation, Indoor localization

1. Introduction

Location awareness is becoming highly important in daily lives of humans (Rainer, 2012). From exact location of one's children, to the exact co-ordinates of an Unmanned Aerial Vehicle, these have become integral for smooth and efficient performance of highly important daily tasks. GPS has revolutionized outdoor localization, and has ensured the possibility of self-navigating outdoor vehicles. Indoor localization is much more complex due to the highly variable indoor environment layouts (Faird, Nordin & Ismail, 2013). Overcoming this complexity can lead to increased accuracy and efficiency in indoor localization.

2. Materials and Methods

Real-time indoor localization co-ordinates in the 3D axes of the Pozyx tag, was achieved at first. Once the localization co-ordinates had been acquired, these co-ordinates were then communicated to Unreal Engine, for visual representation of the Pozyx tag a virtual environment, as shown in Figure 1. This communication was achieved via COM Port 3.

Figure 1: Virtual environment (GUI) developed to display real-time location of robot car using Unreal Engine

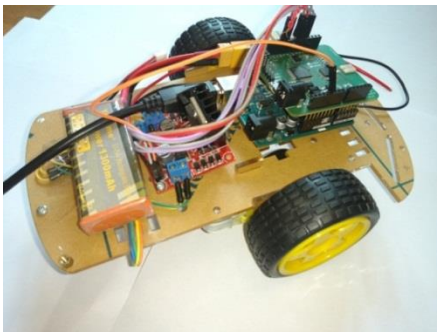

Figure 2: Robot car developed for autonomous navigation, with the help of indoor localization

Alongside achieving indoor localization co-ordinates, a 2-wheel drive robot car was developed, as shown in Figure 2, equipped with necessary components to allow it to turn, move in different directions, and adjust its speed etc., and to self-navigate to a desired destination, once integrated with the localization system. The indoor localization system and the 2-wheel robot car were then integrated using Arduino Uno as shown in Figure 3. Since the Pozyx localization system and the robot car both used Arduino Uno for their operations, integration with each other was achieved with minimal difficulty. The flowchart of Figure 4 highlights the overall operation for the whole project

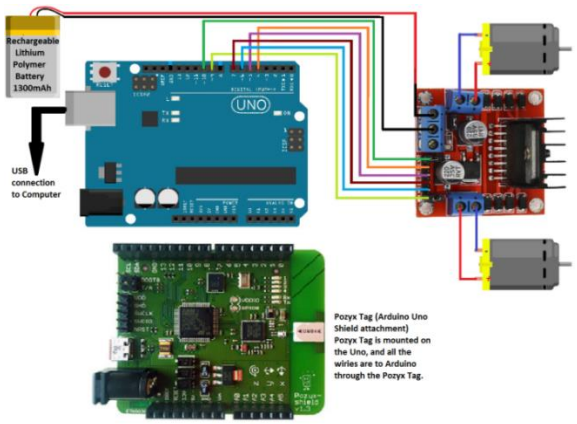

Figure 3: Wiring diagram for the developed robot car and localization system

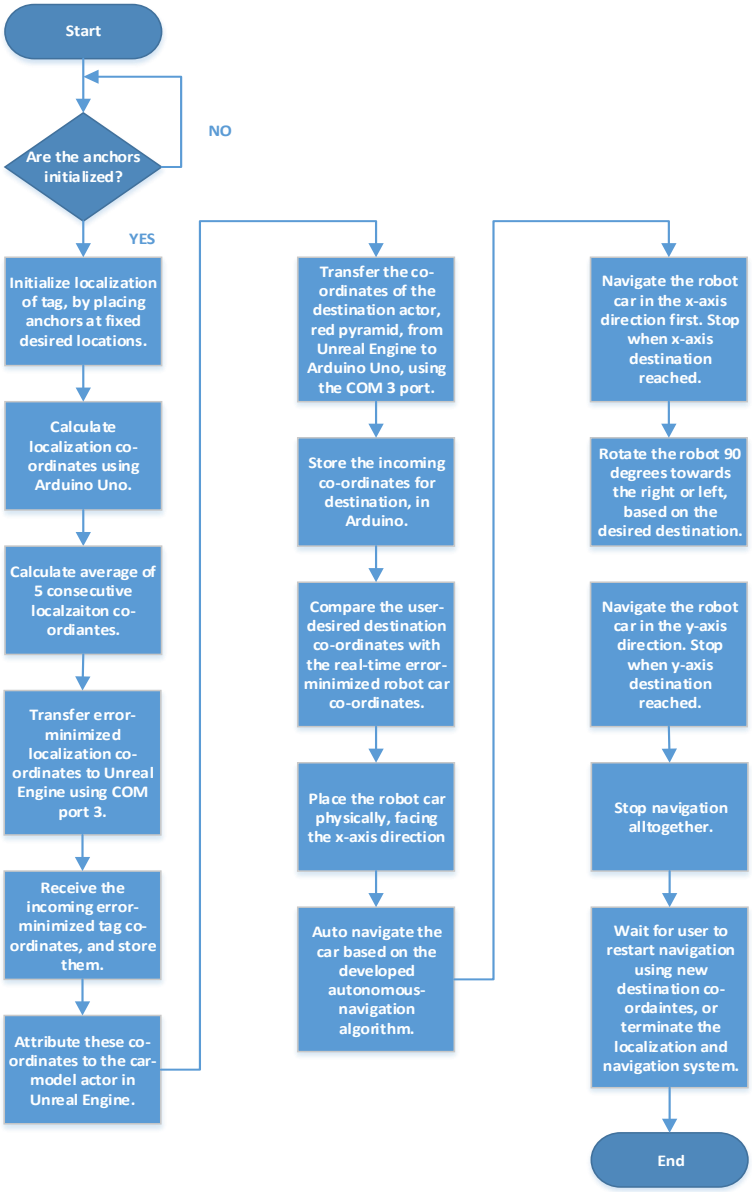

Figure 4: Overall operational sequence for the developed system

3. Results and Discussion

3.1 Accuracy Test for Pozyx Tag Using Arduino IDE

The indoor localization system was tested for several important parameters involved in localization using Arduino IDE and Unreal Engine; which includes visual representation of these localization co-ordinates using Unreal Engine, and autonomous navigation of the robot car using these localization co-ordinates in an indoor environment. For the initialization of the localization system, 4 Pozyx anchors were placed at fixed locations co-ordinates as shown in Figure 5.

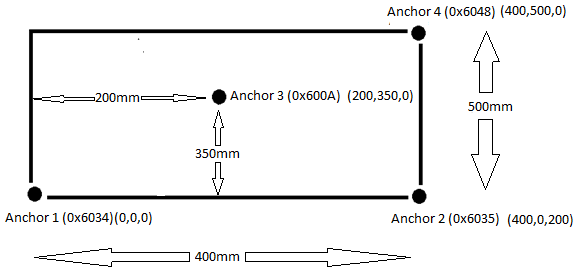

Figure 5: Experimental Setup 1

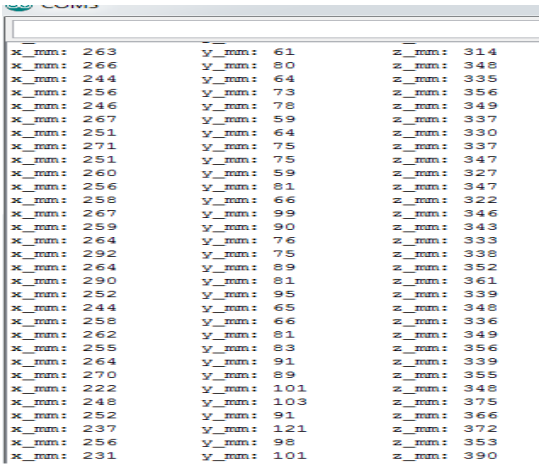

Figure 6: Trial 1 results using Arduino IDE

The test was performed with the Pozyx tag placed at 5 different locations, separately. The locations at which the tag was placed during the testing were (265, 80,350), (315,260,445), (160,280,500), (300,380,470), and (280,135,870). For each location, 25 values for the calculated localization co-ordinates, in millimetres, for the three different axes, namely x-axis, y-axis and z-axis, were collected by noting the values printed in the Serial Monitor, as shown in Figure 6.

The average percentage accuracy for each of the axis, indicate that the results obtained from localization system were satisfactorily accurate. Despite sporadic fluctuations in the obtained localization values, the average localization values were at least 94% accurate for the axis, at least 91.79% for the y-axis, and at least 97.42% for the z-axis, as shown in Table 1 and Figure 7.

Table 1: Results for accuracy test of localization system using Arduino IDE

	Average Percentage Accuracy (%)		
	X-axis	**Y-axis**	**Z-axis**
Trial No. 1	98.57	95.75	97.65
Trial No. 2	98.37	98.51	98.69
Trial No. 3	96.18	96.24	97.97
Trial No. 4	94	95.71	97.42
Trial No. 5	95	91.79	98.69

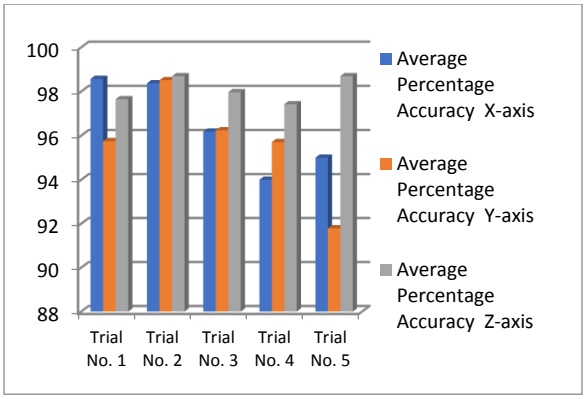

Figure 7: Accuracy results of localization system using Arduino IDE for 3-axis

3.2 Test for Accuracy of 3-axis Indoor Localization and Visual Representation of Localization Using Unreal Engine.

The test was performed with the Pozyx tag placed at 5 different locations, separately, with the Pozyx anchors placed at locations shown in Figure 8. The locations at which the tag was placed during the testing were (16,24,35), (10,16,53), (27,27,34), (23,38,54), and (23,27,39). For each location, 15 values for the calculated localization co-ordinates, in centimetres, for the three different axes, namely x-axis, y-axis and z-axis, were collected and tabulated. These calculated localization co-ordinate values were printed on the simulations screen of the Unreal Engine simulation program, as shown in Figure 9.

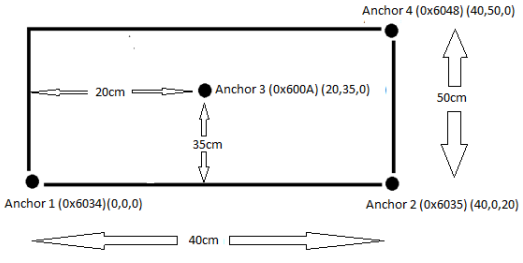

Figure 8: Experimental Setup 2

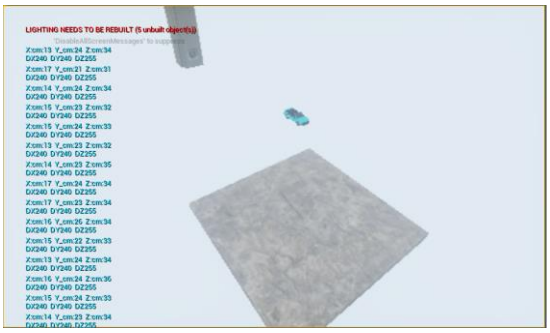

Figure 9: Trial 1 results using Unreal Engine

The percentage accuracy in each of the 5 trials, for each axis, is tabulated in Table 2. These values were graphed on a bar graph, to indicate the average localization accuracy obtained for each axis, for each of the 5 trials as shown in Figure 10. The average percentage accuracy for each of the axis, indicate that the results obtained from localization system were satisfactorily accurate. Despite sporadic fluctuations in the obtained localization values, the average localization values were at least 90.33% accurate for the axis, at least 91.85% for the y-axis, and at least 93.16% for the z-axis.

Table 2: Results for accuracy test of localization system using Unreal Engine

	Average Percentage Accuracy (%)		
	X-axis	**Y-axis**	**Z-axis**
Trial No. 1	93.33	97.78	95.81
Trial No. 2	90.67	92.08	94.84
Trial No. 3	93.58	94.32	97.97
Trial No. 4	95.36	95.96	96.79
Trial No. 5	97.68	91.85	93.16

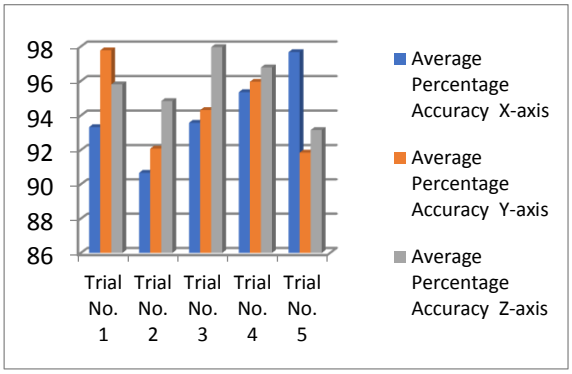

Figure 10: Accuracy results of localization system using Unreal Engine for 3-axis

3.3 Accuracy and Response Time of Robot Navigation Test

The test was performed with the robot car tasked with navigating to 5 different destination co-ordinates, separately, with the Pozyx anchors placed at locations shown in Figure 11. Since the robot car was only able to navigate on the x-axis and the y-axis, the test was only performed for the x-axis and the y-axis values, while neglecting the z-axis values. The desired destination co-ordinates during the testing were (150,150,0), (180,180,0), (200,200,0), (200,250,0), and (230,250,0). For each destination location, 10 trials were conducted, and the final destination coordinates reached, in the x-axis and the y-axis, were collected and tabulated. These final reached-destination co-ordinates were printed on the simulations screen of the Unreal Engine simulation program. Using the tabulated results for each axis, the average value for reached destination co-ordinates was calculated. The percentage accuracy for each of the 5 trials, for each axis is shown in Table 3. These values were graphed on a bar graph, shown in Figure 12 to indicate the average localization accuracy obtained for each axis, for each of the 5 trials.

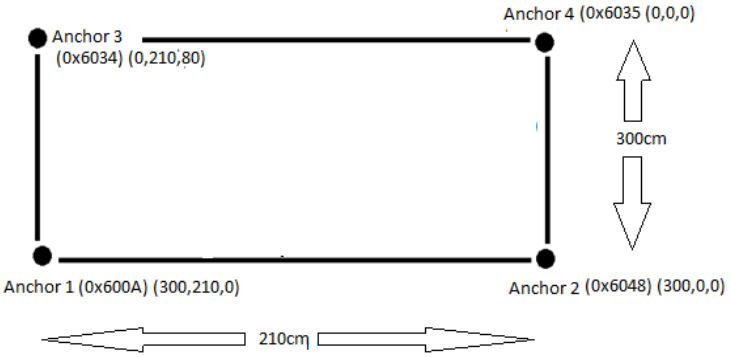

Figure 11: Experimental Setup 3

Table 3: Results for accuracy test of autonomous robot navigation in x-axis and y-axis

	Average Percentage Accuracy (%)		
	X-axis	**Y-axis**	**Z-axis**
Trial No. 1	98.07	96.20	0
Trial No. 2	94.56	92.28	0
Trial No. 3	95.80	87.10	0
Trial No. 4	87.20	97.24	0
Trial No. 5	95.48	97.64	0

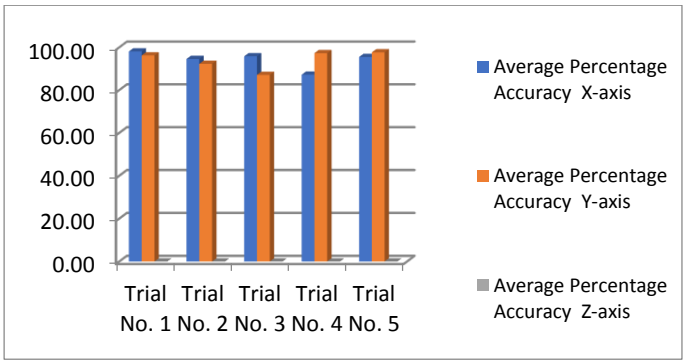

Figure 12: Accuracy results for autonomous robot navigation in x-axis and y-axis

The average time taken, in seconds, for the robot car to travel to the desired destination location, was obtained by calculating the average of time taken for all the 10 trials for each experiment. The average time taken was used to calculate the maximum and the minimum error in the values for the time taken. The average response time for the robot car was also calculated, and displayed in Figure 13, by dividing the total user-defined distance travelled by the time taken to reach the destination, as shown in Table 4, where the answer signified the average rate of travel, or the average speed of the robot in centimetres per second.

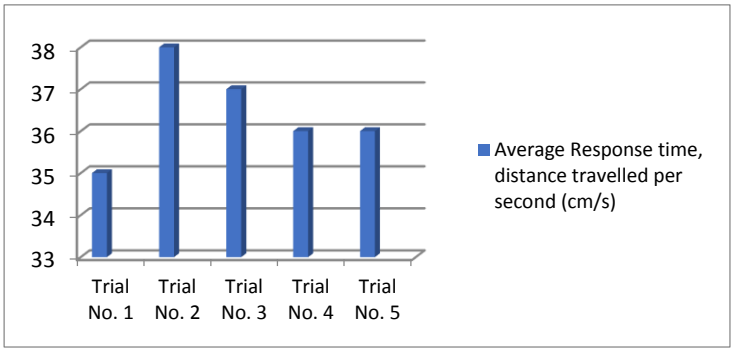

Figure 13: Average Response time results for autonomous robot navigation

Table 4: Results for response time of autonomous robot navigation

	Average Response time, distance travelled per second (cm/s)
Trial No. 1	35
Trial No. 2	38
Trial No. 3	37
Trial No. 4	36
Trial No. 5	36

The average percentage accuracy, shown in Figure 12, for each of the 2 axis indicates that the results obtained from robot navigation using the indoor localization system were satisfactorily accurate. Despite a deviation of the achieved destination from the desired destination location, the average achieved destination values were at least 87.20% accurate for the axis, at least 87.10% for the y-axis, as compared to the user-desired destination location co-ordinates

4. Conclusions

The effectiveness of the Decawave DWM1000 chip, for performing indoor localization in 3-axis, has been thoroughly tested, using the Pozyx tag. The results indicated an average static indoor localization accuracy of above 90%, in all the 3-axis of localization. Therefore, the use of Decawave DWM1000, which communicates using Ultra-wideband signals, in order to perform indoor localization has been proven to be highly effective and highly accurate.

The integration of a self-navigating robot, with the indoor localization system, in order to enable the robot to self-navigate in the indoor environment using the indoor localization system, was also successfully achieved. The robot car was developed and programmed for autonomous navigation. Thorough tests for the autonomous navigation of the robot car yielded a minimum of 87% accurate autonomous navigation results by

the robot car with respect to its desired destination location. Thus, the autonomous robot car and the indoor localization system were effectively integrated.

The virtual room-like environment, created in Unreal Engine, allowed the user to select the desired destination co-ordinates by picking and placing a red-pyramid object, created in Unreal Engine. The location co-ordinates of the red-pyramid were then transferred to the Arduino Uno, through COM 3 port, and the robot car then auto-navigated to the user-desired destination. Throughout the autonomous navigation, the real-time location of the robot car was visually displayed in the Unreal Engine software.

References

Rainer, M., (2012) Indoor Positioning Technologies: Habilitation thesis. ETH Library: ETH Zurich Research Collections

Faird, Z., Nordin, R., Ismail, M., (2013) Recent Advances in Wireless Indoor Localization Techniques and System. *J. Comput. Net. Communc.* pp.252-264.

Journal of Applied Technology and Innovation (eISSN: 2600-7304)
vol. 1, no. 2, (2017), pp. 58-66

Alternative Emergency Braking System for Vehicles

Salman Anwar Khan
School of Engineering
Asia Pacific University of Technology & Innovation
57000 Kuala Lumpur, Malaysia
Email: salman-khan92@hotmail.com

Lim Siong Chung
School of Engineering
Asia Pacific University of Technology & Innovation
57000 Kuala Lumpur, Malaysia
Email: siong.chung@apiit.edu.my

Dhakshyani Ratnadurai
School of Engineering
Asia Pacific University of Technology & Innovation
57000 Kuala Lumpur, Malaysia
Email: dhakshyani@apu.edu.my

Abstract **-** The main aim of this project is to develop an alternative emergency braking system and brake failure identifier for vehicles. In this proposed method, an eddy current braking is used for alternative emergency braking system and the brake failure is identified by placing the ultrasonic sensor opposite to brake pads of the vehicles, that detects the thickness of the brake pad and notifies the driver with a visual and audible alert if there is a tear or brake pads are detached. The performance of the alternative emergency braking system is evaluated with different tests the first test is using aluminium and copper disk; and second test is of different magnets that is neodymium and ferrite magnet. The time taken for the speed of the motor to reduce the speed is up to 65% in 5 seconds using the aluminium disk and neodymium magnet. However, the speed reduction of the motor is only 50% in 5 seconds of time using the copper disk and ferrite magnet. Overall, the system proved to be very accurate in terms of braking efficiency and braking force generated by the eddy currents when the aluminium disk and neodymium magnet is used.

Index Terms **–** Emergency braking system, Eddy current braking, Brake failure identification

1. Introduction

Due to the rapid advance of technology our society and environment has been influenced greatly. In this era of time, one of the most crucial areas that consist of the advancement of technology is the automobile industry which has made travelling easier than ever with such short span of time. Even with such an advanced technology in Automobile industry there is still certain things that need to be improved to make the travelling much safer and reliable.

Over the years in an automobile industry there are many incidents reported of vehicle accidents due to the main brake failure of the car. One of the scariest things that can happen while driving is the brake failure, we apply the brake pedal and it goes all the way down as nothing is there (Smit et al., 2015). The brake failure can occur without any warning to the driver and can lead to anxiety, a feeling of worry and losing the control on steering wheel due to the highly panic situation (Oduro, 2012).

Improvements to road designs continue to reduce the number and sovereignty of vehicle accidents on our roads, improvements are made through the use of new active safety technologies design to help prevent accidents. Many good vehicles are fitted with lane departure warning system, to alert driver if they stray out of their lanes, an advanced emergency braking system that warns the driver if their truck is on a collision course with a slow moving or stationary vehicle ahead, and if necessary the advanced emergency braking system will automatically apply the brakes (Smit et al., 2015).

Despite having these advanced systems in an automobile industry there are no alternative emergency braking systems which can be used to stop the car in case of main car brake failure, which can be due to the worn out of the brake shoe or cut in the brake liner of the car.

This paper focus on developing an alternative braking system for vehicles which will stop the vehicle when there is a failure in the main braking system of the vehicle. This system will have the ability to identify the brake failure using different sensors. The system is designed to give the driver a visual and audible alert of the brake failure so that the driver can be ready to use the alternative system when required or can avoid speeding more and limiting the chances of malfunction.

2. Materials and Methods

To develop and construct the brake failure identifying and activating the alternative brake system, certain analysis of the development and construction requirements of the system must be emphasized. The system to be developed and constructed following are the mandatory requirements.

- Ultrasonic Sensor
- Aluminum Disk sheets
- Neodymium Magnets

- Bearings attached to the rotating steel rods.
- Arduino IDE software
- DC motor
- Buzzer
- Arduino Uno Board
- New Brake pads
- Old brake pads (torn)
- LED

The software part of the system has been done for the construction of the brake failure identifier. Whereas, for the alternative braking system there is no software part and it only comprises of hardware components, however, for the brake failure identifier has both the software and hardware.

2.1 Methodology for Brake Failure Identifier

The software used to program the brake failure identifier is Arduino IDE. For the hardware part, it comprises of Arduino UNO, 220 ohm resistor, breadboard, Ultrasonic sensor, LED and a buzzer. The software Arduino IDE helps programming for many applications (Venkatachalapathi & Mallikarjuna, 2016). The programming done for this project is to identify and notify the drivers in case of brake failure chances.

The ultrasonic sensor is connected to breadboard which measures the distance and alerts the user if brake pad thickness is lesser than normal defined length. For notifying the user LED and buzzer is connected which gives an audible and visual alert to the driver in case of brake failure.

2.3 Methodology of Alternative Emergency Braking System

The construction of the alternative braking system has the gears, bearings rotating rod, neodymium magnet, aluminum sheets, 24V dc motor and external voltage supply for dc motor. The dc motor is connected to one of the gear and the other gear is connected to the rotating rod this provides the rotational movement.

The bearing has been connected to allow the smooth rotation of the rod, the placing of the bearing was in the center of the two wood blocks which allows the passing of the rod from the center. The tightening nuts were connected to the bearing that gives stability to the high speed rotation of the rod.

The braking pads has been placed in the same base as the alternative brake system. There are two braking pads that are used one is completely torn and other is the new braking pad. Both the braking pads are placed and checked if the system identifies and notify the user with audible and visual alert.

3. Results and Discussion

The objectives of the project are achieved successfully, the results obtained for both the systems alternative emergency braking and brake failure identifier will be discussed in detail.

3.1 Results of Brake Failure Identifier

Figure 1 shows the brake fail identifier system, when there are new brake pads placed the LED is off. Hence, there is no need for activating the alternative brake system.

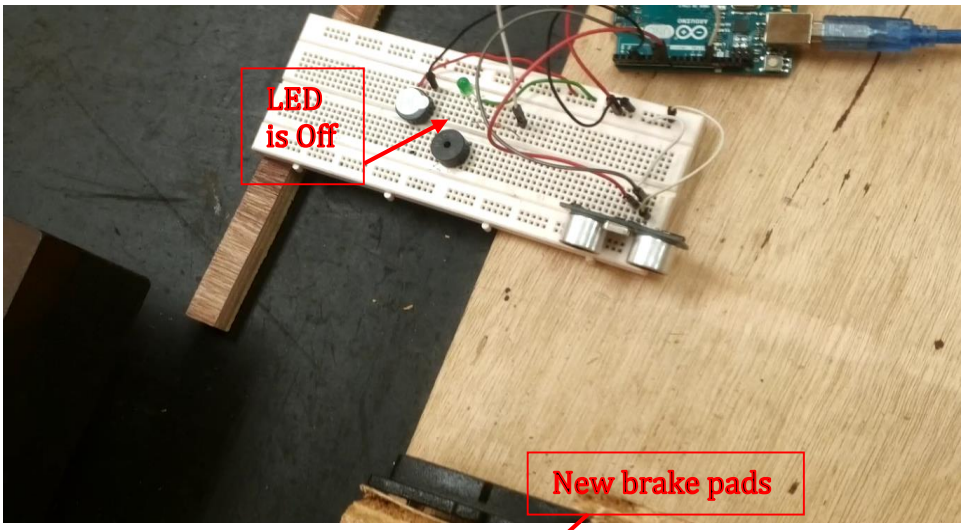

Figure 1: LED off with new brake pads

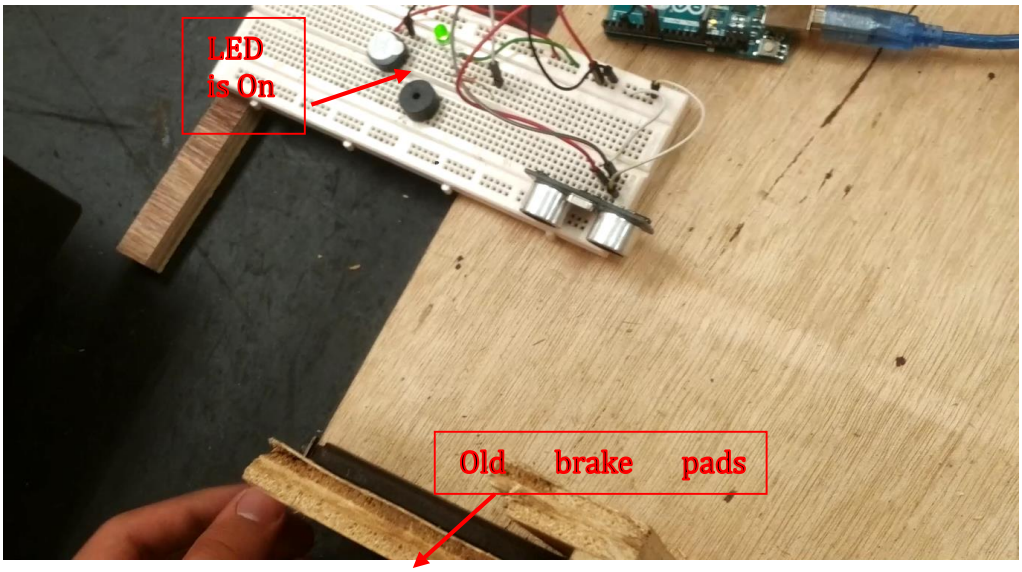

Figure 2: LED on with torn brake pads

Figure 2 shows the system has detected the danger and LED is on because the brake pads placed are the old ones with completely torn brake pad. This gives the user a visual and audible alert, since, this is just the picture the buzzer sound cannot be proved to be in working condition but a demonstration video has been recorded and will be demonstrated during the presentation.

3.2 Results of Alternative Emergency Braking System

The alternative brake system hardware is shown in the Figure 3, and it can be said that the higher the speed of the motor the higher the braking force is applied by the eddy currents.

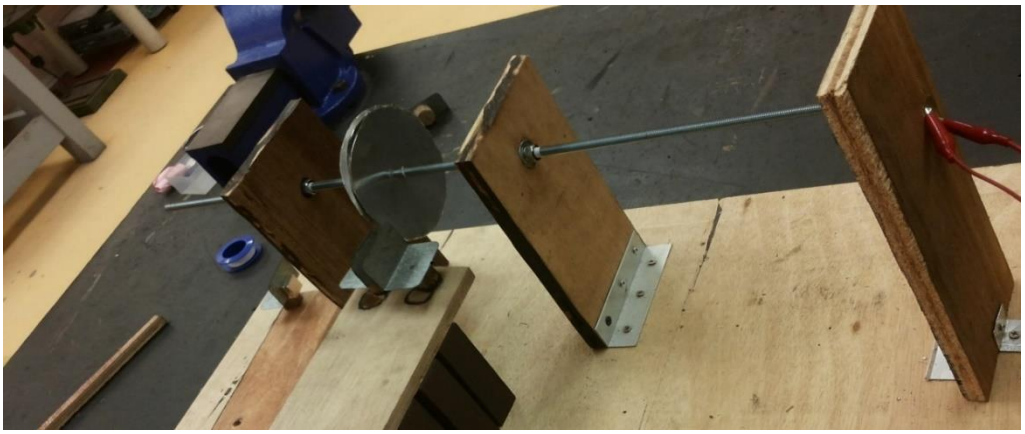

Figure 3: Hardware of alternative braking system

4. Testing of the System

To test the efficiency and effectiveness of the system five different tests are done with collected data. The data analysis of the results obtained are done with the help of graphs and experiment explanation is also given below of the two tests (Kumar, Ibraheem & Sharma, 2014).

4.1 Different non-conductive materials used for disk

The copper disk was attached to the steel rod and neodymium magnets were placed aligned at 2 cm apart and when the 24V dc motor was turned on the neodymium magnet was brought almost 0.5 cm closer to the copper disk, which is to determine the deceleration speed of the motor.

a) Data Collection

For the proposed method, the data collection is shown in Table 1.

Table 1: Braking force Copper versus Aluminum disk

Time (sec)	Copper (km/h)	Aluminum (km/h)
1	80	80
2	75	71
3	65	56
4	55	43
5	50	30

b) Data Analysis

As it can be seen in Figure 4 that the speed of the DC motor is reduced more than 50% that is 80km/h to 28 km/h in 5 seconds when aluminum is used and when copper disk is used the deceleration taken in 5 seconds is from 80 km/h to 48 km/h.

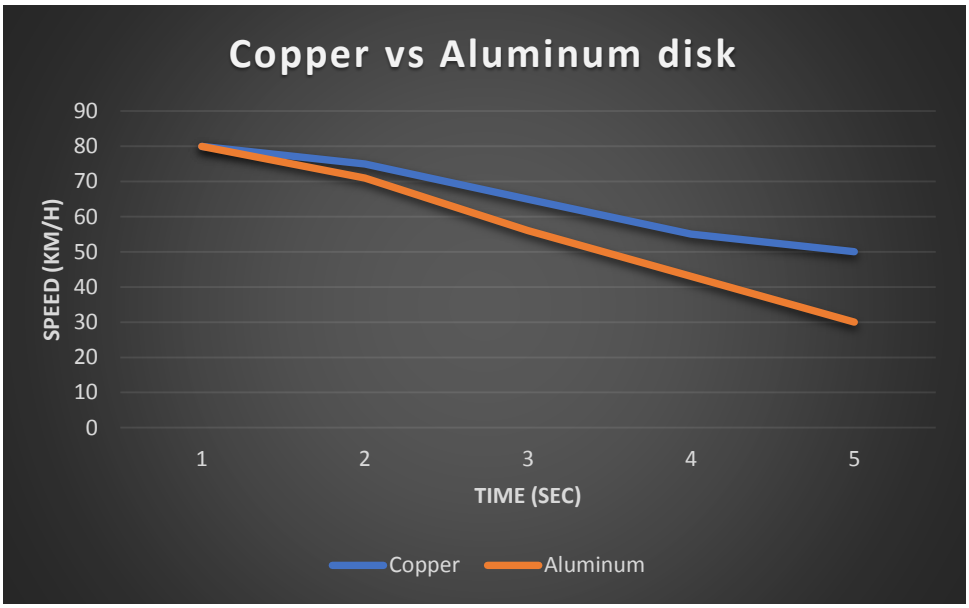

Figure 4: Speed variation with time (for Copper & Aluminum)

4.2 Distance between magnet and aluminum disk

The aluminum disk was attached to the steel rod and neodymium magnets were placed aligned and tested with three different distances apart. Then the 24V dc motor was switched on to determine the deceleration speed of the motor at different distances between disk and magnet

a) Data Collection

For the proposed method, the data collection is shown in Table 2.

Table 2: Distance between magnet and disk

Time (sec)	Distance 1.5 cm (km/h)	Distance 1 cm (km/h)	Distance 0.5 cm (km/h)
1	60	60	60
2	59	58	50
3	58	55	40
4	57	52	32
5	55	48	25

b) Data Analysis

Figure 5 shows the deceleration speed of the dc motor; the speed of the dc motor is monitored with three different distances between the disk and a magnet. The test with the closest distance that is 0.5cm gives the most efficient results in terms of braking torque generated. The result with 1.5cm distance shows only reduction of 10% speed within 5 seconds. Therefore, the closest the distance is between the disk and magnet, the higher the braking force will be generated to reduce the speed up to 60% within 5-6 seconds of time.

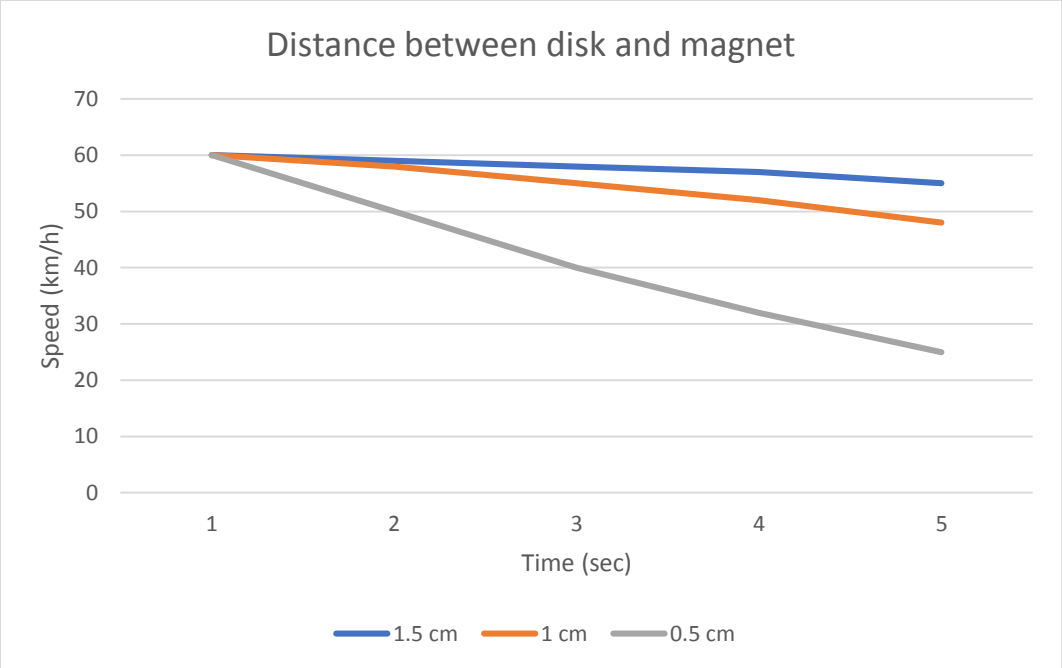

Figure 5: Distance between disk and magnet

4.3 Limitations of the Project

The magnetic brake method, which works on the principle of producing eddy currents is restricted to its magnetic braking force based on the speed of the motor, the lower the speed of the motor, the lower the braking force; and higher the speed of the motor, the higher the braking force will be applied. Therefore, the vehicle that is moving at low speed the less braking force will be applied and hence more chances of collision at lower speeds.

4.4 Recommendation to the Project

The auto-tire locking system can be installed to the tires as the magnetic brake reduce the speed of the car up to certain limit and then the auto-tire lock system will be activated (Christoph, Frank & Horst, 2015). Therefore, following and adhering the recommendations will have a positive impact to the overall braking system.

5. Conclusions

In conclusion, the alternative emergency braking system is being implemented to overcome the accidents caused due to the primary brake failure of the vehicles. However, the main reason of failure is due to the mechanical friction tear of the brake pads or cut in the brake liner. The system being implemented in this project is using the eddy current braking system that does not involve the mechanical friction, hence, the chances of accidents will be reduced.

For the driver to know the brake failure a program is developed in such a way that detects the torn of brake pads or cut in brake liner and notifies the driver with audible and visual alert. Statistical results show that the speed of the car is reduced up to 65% as soon as the alternative braking system is activated, however, further research and contribution from various researchers can improve the efficiency for generating the higher braking force and provide more reliability to the automobile sector.

References

Oduro, S. (2012) Brake Failure and its Effect on Road Traffic Accident. International Journal of Science and Technology. 9 (1), p. 448-454.

Smit, P., Meet, P., Anand, P. & Chetan, S. (2015) Development of the ElectroMagnetic Brake. *International Journal for Innovative Research in Science & Technology*. p.485-492.

Venkatachalapathi, N. & Mallikarjuna, V. (2016) Automatic Brake Failure Indicator and Over Heating Alarm. *International Journal of Emerging Technology and Advanced Engineering*. 6(7), p.8671-8681.

Christoph, H., Frank, R. & Horst, E., (2015) Simplified Model of eddy currents brakes and its use for optimization. In *Tenth International Conference on Ecological Vehicles and Renewable Energies*. p.324-332.

Kumar, A., Ibraheem & Sharma, A. (2014) Parameter Identification of Eddy Current Braking System for Various Applications. In International Conference on Innovative Applications of Computational Intelligence on Power, Energy and Controls with their Impact on Humanity. p.191-195.

Journal of Applied Technology and Innovation (eISSN: 2600-7304)
vol. 1, no. 2, (2017), pp. 67-78

Challenges in Software Design and Development of Cross-Platform Mobile Applications

Fitim Fejzullahu
School of Computing
Asia Pacific University of Technology & Innovation (APU)
Technology Park Malaysia, Bukit Jalil, Kuala Lumpur 57000 Malaysia
Email: fitafejzullahu24@gmail.com

Geetha Kanaparan
School of Computing
Asia Pacific University of Technology & Innovation (APU)
Technology Park Malaysia, Bukit Jalil, Kuala Lumpur 57000 Malaysia
Email: kgeetha@apu.edu.my

Abstract **–** The demand for developing cross-platform mobile applications has been growing in recent years. The attention towards this development approach is due to the ability to create mobile applications in a shorter time, with less cost and run cross-platforms using a single code base. This paper reviews the challenges of cross-platform developments and examines hybrid development as an alternative to native development. Emphasis is made on the examination of the latest mobile cross-platform frameworks by evaluating the strengths and weaknesses when using this developing approach. Findings from this research indicates that cross-platform development has been significantly improved (with the support of frameworks such as: Apache Cordova, React Native and Xamarin).

Index Terms **–** Cross-platform, issues, native, hybrid, frameworks

1. Introduction

Smartphones are small handy devices that can operate from anywhere at any time. The devices contain applications that support users in daily activities. According to Wireless Smartphone Strategies 44% of population uses smartphones in 2017 and the number will increase in the next 3 years up to 58% (Sui, 2016). Therefore, efficient mobile applications are necessary to keep up with latest market trends. As a result, it is imperative for business organizations and software companies to constantly seek for better approaches to bring the first and the best mobile applications in the marketplace. When developing mobile applications, the main concern for most organisations is time and cost. Native applications are only developed for a specific platform using proprietary SDK (Software Development Kit). Each platform requires a unique development

approach for its users. This process is complicated and costly because today there are many platforms out there, such as: iPhone, Android, Windows and Blackberry, which require different software development environment (Halidovic & Karli, 2014; Forni & Meulen, 2016). Thus, many companies struggle to develop applications to reach out to users across platforms. Thus, the need for hybrid application development has become necessary in order to ease the process by allowing to create cross-platforms mobile applications to be developed with a single codebase.

Initially, cross platform development started with HTML5, CSS and JavaScript, but this development approach was not able to support many features required by end users including camera, geolocation and notifications. Subsequently, developers found a better approach for cross-platform development using third-party support, which creates a container inside native application and allows web applications to be embedded inside this container. Presently, cross-platform development is used widely by many software companies. This development approach uses one codebase for multiple platforms which makes the development process shorter and low cost.

Several studies have been conducted to identify the best approach for creating cross-platform applications with a single base code. Positive improvements have been made with the support of new frameworks introduced by large companies such as Microsoft, Google and Facebook. The attention towards cross-platform developments have been increasing and software companies have realised the benefits of investing on this development technology. However, frameworks have their own benefits and drawbacks. The following section evaluates the latest frameworks native verses hybrid development that are significant to achieve reliable cross-platform applications.

2. Cross Platform Development (Hybrid vs Native)

Mobile applications may be categorized as native, web application, or hybrid. Native applications are developed using individual native languages and usually target one specific platform. For instance, Java programming language is used to develop Android applications while Swift based objective C is used to develop iOS applications. This approach allows creating compatible mobile application for only one target platform and is the most reliable way to develop mobile applications (Amatya & Kurti, 2013). However, using native development approach is not suitable solution for organizations that seek to exchange data among cross-platforms.

Table 1. Native Applications Development Requirement (Latif et al., 2016)

	iOS	**Android**	**Windows Phone**	**Blackberry OS**
Languages	Objective C, Swift	Java, (some C, C++)	C# and VB.Net	Java

Tools	xCode	Android SDK/ Android studio	Visual Studio	BB Java Eclipse
Format	.app	.apk	.xap	.cod
Hosting	App Store/ i-Tune	Google Play Store	Windows Phone	Blackberry World

Table 1 displays the programming languages, tools, formats, and the hosting environments for development of native mobile applications. Creating native applications is certainly complex since each platform requires unique way of development. The limitation to one platform is not convenient for companies that want to reach all users cross-platforms. They invest huge amount of money to create mobile applications for Android, iOS, Windows and website development on the other hand. The challenge is also in terms of cost, effort, team development skills, technology applied and time to build the application.

By contrast, web applications are developed using HTML5, CSS and Java Script. This development approach allows applications to run across web browsers using a single code base (Allen, Graupera & Lundrigan, 2010). However, HTML5 web applications do not support many functions in mobile devices including camera, geolocation, user interface and offline support (Heitkötter, Hanschke, Majchrzak, 2012).

As a result, hybrid mobile applications were introduced which are essentially web applications embedded inside a native container enabling the HTML5 pages to be displayed in mobile device in same behaviour as native applications (Halidovic, & Karli, 2014). Mobile hybrid applications are developed with HTML5, CSS and JavaScript and supported by frameworks (Halidovic, Dhupia & Rubab, 2015). Hybrid applications are developed in such a way that have look and feel of native applications. Generally, end users find it hard to distinguish the difference between hybrid and native applications. However, in the development side it is a huge difference due to the tools and technologies used in the development. Hybrid application have been facing issues in performance and user interface design.

All three development formats have their own benefits and drawbacks. Much discussion is ongoing about which approach is better. With the increased competition among companies, the demand for having mobile applications with low cost and fast development has also increased. There is no doubt that native application development is still preferred for creating reliable applications. However, in long term, hybrid applications appear to be a better solution since they are able to run across platforms. As an example, many applications had been developed to run on the Blackberry platform. However, in the recent years Blackberry brand does not have new mobile version and company is almost out of the market. This means, that applications should not be developed only for a particular platform but instead applications need to be cross-platform to sustain in long run. Business organizations are looking for applications that are 'developed once and run everywhere' (Charkaoui, Adraoui & Benlahmar, 2014).

Table 2 displays similarities and differences between native, hybrid and HTML5 web applications. The table shows the benefits of cross-platform development compare to web applications and native applications development. However, hybrid approach has issues with performance and user interface. The responsiveness of hybrid applications is slower compare to native applications. Nevertheless, choosing the right development framework enhances the hybrid performance, giving more powerful features with native feel and look. On the contrary, Table 2 shows that hybrid approach has extra benefits compare to native such as application flexibility, optimization, and multiplatform coverage.

Table 2. Comparison between Native, HTML5 and Hybrid (IBM Corporation, 2012; Korf & Oksman, 2016)

App Features	**Native**	**HTML5/ Web apps**	**Hybrid**
Development Languages	Only Native	Web only	Native and Web
Code portability and optimization	None	High	High
Access Device Specific Features	High	Low	Medium
Leverage existing knowledge	Low	High	High
Graphics	Native API's/ High	HTML, Canvas, SVG/ Medium	HTML, Canvas, SVG/ Medium
Upgrade flexibility	Low	High	Medium
Performance	fast	Slow	Medium
Native look and feel	Native	Emulated	Emulated/ closer to native
Distribution	App Store	Web	Play Store, App Store
Device Access			
Camera	Yes	No	Yes
Notifications	Yes	No	Yes
Contacts, Calendar	Yes	No	Yes
Offline storage	Secure file storage	Shared SQL	Secure file system, shared SQL
Geolocation	Yes	Yes	Yes
Gestures			
Swipe	Yes	Yes	Yes
Touch and Pinch	Yes	No	Yes
Connectivity	Online and Offline	Mostly online	Online and offline

2.1 Challenges of Cross-Platform Development

Developers face several challenges when creating cross-platform applications. Usually problems encountered are related to performance, functionality and design. These issues create uncertainty among business organisations and software companies when choosing cross-platform over native development. Cross-platform development is still improving and does not appear to be yet completely stable. According to IBM Corporation (2012), most common issues identified in hybrid development are HTML5 limitations, user interface design, lack support of animation and slow performance.

HTML5 has limitation on accessing (DOM) Document Object Models in web pages (Eisenman, 2016). DOM is an application programming interface (API) which support programmers to build documents, define logical structure of HTML files allowing to add, modify and delete elements. According to Halidovic & Karli (2014) the use of plugins in developments of hybrid applications may not be compatible with the new versions of frameworks. Hence, switching to the latest framework version will affect the application, so developers need to constantly solve issues occurred whenever new update is applied.

In many instances, the user interface design needs to be adjusted based on the target platform. This usually takes extra time for the development team to test the interface design for multiple platforms ensuring application works good for different mobile screen sizes. Yet, another issue is the slow responsiveness of application when using camera and animations. This can cause performance issues including slow response, battery drain, and freezing phone (Korf & Oksman, 2016). Another, issue is the lag on touch screen which may delay response time up to 300ms compared to native. As observed by Natili (2013), HTML5 also has weakness when displaying long list of data, causing delay or freezing on data loading. However, this issue may be resolved with the support of plugins.

Table 3: HTML5 features supported on WebView

HTML5 Features	Android WebView	iPhone WebView
Elements of media: video and audio	Yes	Yes
Input types: search, TEL, number and so on	Yes	Yes
Semantic elements: footer, header, articles, summary, meter etc.	Mostly	Mostly
Web Storage	Yes	Yes
Application catches	Yes	Yes
SVG and Canvas Graphics	Yes	Yes
Attributes of CSS3	Mostly	Mostly
Math ML	No	Mostly
Drag and drop	No	No

The issues discussed are the main reasons why software companies and business organizations struggle to decide when choosing cross-platform development approach. Latest frameworks that support cross-platform development claim their capabilities are strong enough to compete with native development. This research investigates the main frameworks that support cross-platform development and attempts to identify whether the frameworks are capable to minimize or solve the issues faced by developers when developing cross-platforms.

Table 3 contains some of the HTML5 features that are supported by Android and iOS WebView. Although many HTML5 features can be implemented, programmers cannot rely on HTML5 development, because limitations affect the overall development of cross-platform applications.

2.2 Cross-Platform Development Frameworks

Cross-platform frameworks have made the development process way easier by providing APIs and Libraries which support the implementation of native functionalities. Those frameworks support the development of applications a 'native look and feel'. There are several open source frameworks available for use. Large companies such as Facebook and Microsoft are investing on this technology to provide better approach for cross-platform development. Due to the large number of frameworks, this section is dedicated to evaluation of three well known cross-platform frameworks: Apache Cordova, Xamarin and React Native. React Native is relatively new open source framework. It was launched in 2016 and is widely used in many software development companies.

2.2.1 Apache Cordova Framework (PhoneGap)

Apache Cordova is a well-known open source framework developed for supporting hybrid mobile applications using web technologies such as HTML5, CSS3 and JavaScript. Development with this approach requires WebView embedded, structure API to access functions of native applications, plugins and file storage (Bosnic, Papp & Novak, 2016).

This framework is widely adapted by many software companies because it overcomes the limitations of HTML5 and has a native feel and look (Pardeshi & Shirvas 2013). Applications developed with this framework are supported by Android, iPhone, Palm, Windows Phone, Symbian and Blackberry (Allen, Graupera & Lundrigan, 2010). Apache Cordova provides access to camera, GPS, file system and accelerometers with the embedded HTML5 code which is then displayed in mobile view just like native applications. Apache Cordova also supports multiple plugins and libraries which allow extending functionalities with JavaScript codebase. Under this framework many tools for development are supported including Ionic SDK, Adobe PhoneGap, Visual Studio allowing developers to have multiple selection of tools for developments (Francese et al. 2013; Natili, 2013).

Despite many benefits, this framework tends to have a slower performance compare to native especially when many plugins are used. The graphic user interface (GUI) and back end functionalities can be implemented similar to native. However, extra effort is required to implement the graphic user interface for different mobile screen sizes. A study from Heitkötter, Hanschke & Majchrzak (2012) was conducted to evaluate Cordova, Titanium, and Web applications with the aim to compare with native development using several criteria. The result from this study indicate that Cordova is similar to native approach in terms of the look and feel. In addition, users may not be able to differentiate between hybrid applications developed with Cordova and native applications.

A survey by Dalmasso at al. (2013) shows that Apache Cordova have less consumption of memory, power, and CPU, because it does not have a dedicated user interface component. In addition, Rieger & Majchrzak (2016) state that hybrid mobile can have native appearance with the APIs which support the development of cross-platform application to access majority features similar to native applications. However, performance may still be an issue depending on the browser capabilities.

2.2.2 Xamarin Framework

Using C# codebase and Microsoft tools, Xamarin allows creating cross-platform applications for Android, iOS and Window with native user interface. This framework is used by millions of developers because it provides good environment for development of cross-platform applications. The application development for native Android, iOS and Windows is simplified by using a single codebase, existing teams with same development skills for all platforms. After completing development, Xamarin enables applications to be compiled into native platform such as iOS, Android or Window, meaning become purely native. The advantage of using Xamarin is code reuse. Xamarin.foms are powerful for mapping the native code and the user interface allowing to reuse the same forms. The performance is better compare to Apache Cordova since the application is compiled to native after development. However, the use of many plugins results in a slower performance. The decision lies on the development team, who should carefully select plugins and libraries when developing with Xamarin framework. As this framework is relatively new, many bugs have been reported from developers and testers. Boushehrinejadmoradi et al. (2015), built a testing tool called X-Checker to test the process of Xamarin when compiling applications to native Windows and Android. They detected 47 bugs related to inconsistency of translation between the Android and Window Phone APIs. Despite the reported bugs, Xamarin new version updates gradually fixes majority of bugs detected. However, in certain cases Xamarin framework updates may affect the development of previous applications. This requires consideration whether the application that has been developed previously should be upgraded as new releases are available.

2.2.3 React Native Framework

React native is an open source framework that was developed under Facebook. React Native currently supports only iOS and Android platforms. This framework is soon expected to be compatible with other platforms as well. In 2013, Facebook that working with HTML5 is not good solution to create compatible cross-platform applications, instead Facebook focused on creating a framework that was developed with HTML5, JavaScript based, and includes native programming languages such as object C and Java to support development. Apart from HTML5, Css and JavaScript, this framework requires understanding of the react.js as well as additional scripting languages including JSX, XML, Xcode and ECMA Script (Eisenman, 2016). Although this framework is new, it is widely adapted by many organizations. The issues of mobile cross-platform developments can be minimized using React Native framework because the application development itself allows the native programming languages to overtake at any time when complex requirements need to be implemented. The benefit of using this framework is the reuse of components which are compiled to native within the framework itself without requiring WebView components. This will directly increase the performance including speed, features, native feel and look. In React Native framework the Document Object Model (DOM) components are rendered using in-memory DOM instead of WebView allowing to render the minimal use of DOM only when necessary (Eisenman, 2016).

React Native allows reuse of UI components without rewriting, which allows to switch an application that already exist without rending it (Eisenman, 2016). For instance, an application developed with Cordova can be reused in React Native. Its entire focus is to build solid UI unlike MeteorJS or AngularJs whose main concern is whether the application works. Also, the IU has native approach with JavaScript interaction that is asynchronous with the native development environment. Efficiency in terms of development with rapid product development and native results is another advantage of this framework. React Native has simplified development of cross-platform applications and performance has been increase with DOM abstraction which takes full control over the application without relying on WebView. This approach allows linking of plugins with native, empowering the application to have full access to zoom, rotation and compass, while less memory load is used.

3. Selecting the Right Framework

Deciding between native and hybrid development depends on the complexity of application as well as target platform. For large companies, choosing cross-platform development on in stage is risky since this approach is constantly upgrading and it is not stable. However, companies that need to compete in a digital market or want to explore a digital market within short term results, the cross-platform development approach is better solution. Mercado, Muniana & Meneely (2016) evaluated the difference between native and hybrid application using 787,228 user reviews data from play store and app

store by evaluating users rating for particular applications. The focus on the research was on reliability, security, performance and the usability of the applications. The finding from their study indicates that hybrid approach tends to have more user complaints. However, certain applications that have been developed with the latest frameworks, are nearly with native with feel and look and it is impossible for users to know whether application is hybrid or native.

Cross-platform developments are supported by many frameworks which are built on top of HTML5, CSS and JavaScript. The best approach for cross-platform development depends on the application requirements and complexity. Table 4 compares Apache Cordova, React Native and Xamarin frameworks as well as their potential to develop cross-platform applications. Each framework has different approach for creating cross-platform applications.

Table 4: Comparison between three Frameworks

Features	**Apache Cordova**	**Xamarin**	**React Native**
Use of Native Programming Languages	-	C, C++	Object C and Java
Target Platforms	Android, iOS, Windows, Blackberry, web browser etc.	Android, iOS, Windows, Blackberry etc.	Only Android and iOS
User Interface Development	HTML5, CSS3, JavaScript, (AngularJs, Ionic)	HTML5, CSS3, JavaScript	HTML5, CSS3, JavaScript, React Native Components
Reusable code	Yes	Yes	Yes
Native Feel and Look	Moderate – Supported by Ionic SDK	High – Compiled to native app	High – Is built as Native App
Build as Hybrid	Yes	Yes	No
Data Object Model	WebView	WebView	Own Model
Support API's	Yes	Yes	Yes
Able to access data on web	Yes	Yes	Yes

Table 4 indicates that the frameworks have similar objectives, but they use different approach during development. While Apache relay on WebView container and is developed completely hybrid with HTML5, CSS and JavaScript, React Native uses native approach for exchanging data on the application. Also, Xamarin relay mostly on C, C++ programming language and then compile the application to cross-platforms. During compiling applications Xamarin may have several errors and developers need extra work to solve them. React Native allows native programming languages such as Java and C++ to take over at any time when HTML5 and JavaScript do not support certain functions. Also, React Native has better performance compare to Apache Cordova an Xamarin.

However, the limitation of React Native is that it only supports Android and iOS platforms.

The comparison above shows that all frameworks offer good solutions to improve most of the issues when developing cross-platform applications. Most issues are manageable when the development team are professional and they are able to select the right approach for developing cross-platform applications. In addition, the right approach to develop cross-platforms depends on the type of applications that are expected to be developed.

4. Conclusion and Future Research

This study has revealed that cross platform development is the direction for future mobile application development due to its benefits over native development. Cross-platform developments are currently going through a process of redefining the development process in order to be compatible and to overcome the weaknesses of native developments. Presently, selecting cross-platform developments requires consideration of the target platforms, the type of application and availability of the resources in order to develop a reliable application. HTML5 has few limitations that affect the overall performance and user experience. However, much improvements have been made recently with the frameworks introduced by large companies such as Microsoft (Xamarin) and Facebook (React Native). These frameworks are promising due to their ability to create application with a single code base and then compile into native applications.

The overall benefits of cross-platform development are for end users and organizations. Using this approach brings users updated applications within a short time, while organizations spend less money, and are able to bring innovation solution into the market by targeting multiple platforms.

Further research is needed to test the cross-platform development frameworks since there are many, but none of them appears to be completely stable. Therefore, being able to test these frameworks would contribute better to the improvement of cross-platform development and would establish standardisation for consistent cross-platform application development.

References

Allen, S., Graupera, V. & Lundrigan, L. (2010). Pro smartphone cross-platform development. 1st Ed. New York: Apress.

Bosnic, S., Papp, I. & Novak, S. (2016). The development of hybrid mobile applications with Apache Cordova. *2016 24th Telecommunications Forum* (TELFOR). p.1-4.

Boushehrinejadmoradi, N., Ganapathy, V., Nagarakatte, S. & Iftode, L. (2015). Testing Cross-Platform Mobile App Development Frameworks (T). *2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. p.441-451.

Charkaoui, S., Adraoui, Z. & Benlahmar, E. (2014). Cross-platform mobile development approaches. *2014 Third IEEE International Colloquium in Information Science and Technology (CIST)*. p.188-191.

Dalmasso, I. S. K. Datta, K., Bonnet C. & Nikaein, N. (2013). Survey, comparison and evaluation of cross platform mobile application development tools. *2013 9th International Wireless Communications and Mobile Computing Conference (IWCMC),* Sardinia. p. 323-328.

Eisenman, B. (2016). *Learning React Native*. 1st ed. United States of America: Rally Media Inc. p. 10-50.

Francese, R., Risi. M., Tortora, G. & Scanniello, G. (2013). Supporting the development of multi-platform mobile applications. Web Systems Evolution (WSE), 2013 15th IEEE International Symposium. p. 87,90, 27-27.

Halidovic, R. & Karli, G. (2014). Cross-Platform Mobile App Development using HTML5 and JavaScript while leveraging the Cloud*. IOSR Journal of Engineering.* 4(2). p.06-11.

Heitkötter, H., Hanschke, S. & Majchrzak, T.A. (2012) Comparing cross-platform development approaches for mobile applications. *In: 8th International Conference on Web Information Systems and Technologies, WEBIST*. p. 299–311

IBM Corporation (2012). Native, web or hybrid mobile-app development. Somers, NY 10589 [Online] Available at: ftp://public.dhe.ibm.com/software/pdf/mobile-enterprise/WSW14182USEN.pdf [Accessed 20 March 2017].

Korf, M. and Oksman, E. (2016). Native, HTML5, or Hybrid: Understanding Your Mobile Application Development Options. [online] Salesforce Developers. Available at: https://developer.salesforce.com/page/Native,_HTML5,_or_Hybrid:_Understanding_Your_Mobile_Application_Development_Options [Accessed 25 March 2017].

Latif, M., Lakhrissi, Y., Nfaoui, E. & Es-sbai, N. (2016). Cross platform approach for mobile application development: a survey. *2016 International Conference on Information Technology for Organizations Development (IT4OD)*. Fez. p. 1-5.

Mercado, I., Muniana M. & Meneely A. (2016). The Impact of Cross-Platform Development Approaches for Mobile Applications from the User's Perspective. *Proceedings of the International Workshop on App Market Analytics - WAMA 2016*. p.43.

Natili, G. (2013). PhoneGap Beginner's Guide. 1st ed. Birmingham: Packt Publishing.

Pardeshi, A. (2013). To Study and Design a Cross-Platform Mobile Application for Student Information System using PhoneGap Framework. *International Journal of Emerging Technology and Advanced Engineering.* 3(9), p.390.

Rieger, Ch. & Majchrzak, T. A. (2016). Weighted Evaluation Framework for Cross-Platform App Development Approaches. *Springer International Publishing AG 2016 S. Wrycza (Ed.):* SIGSAND/PLAIS 2016. LNBIP 264, pp. 18–39.

Sui, L (2016). 44% of World Population will Own Smartphones in 2017. [Online] Available at: https://www.strategyanalytics.com/strategy-analytics/blogs/smart-phones/2016/12/21/44-of-world-population-will-own-smartphones-in-2017#.WgESRGiCzIU [Accessed 20 March 2017]. Strategyanalytics.com [Accessed 28 March 2017].

Journal of Applied Technology and Innovation (eISSN: 2600-7304)
vol. 1, no. 2, (2017), pp. 79-93

Cloud Testing: Requirements, Tools and Challenges

Ahmad Dahari Bin Jarno
CyberSecurity Malaysia
Mines Resort City, 43300, Seri Kembangan, Selangor, Malaysia
Email: dahari@cybersecurity.my

Shahrin Bin Baharom
CyberSecurity Malaysia
Mines Resort City, 43300, Seri Kembangan, Selangor, Malaysia
Email: shahrin.baharom@cybersecurity.my

Maryam Shahpasand
FSec research centre
Faculty of Computing, Engineering & Technology
Asia Pacific University of Technology & Innovation
Technology Park Malaysia, 57000 Kuala Lumpur, Malaysia
Email: dr.maryam.shahpasand@apu.edu.my

Abstract - Cloud Computing is a technology which gives computation, software, data and storage services over the network. The software industry is laying increased emphasis on Quality Assurance (QA) and Testing requirements for successful product development today. To ensure a high level of security of cloud services and applications, testing is an appropriate approach to detect possible vulnerabilities before real case scenarios occur. Therefore, many public cloud providers revealed the growing number of Testing Centre of Excellence (COE) and software test automation being encouraged by companies.This paper presents a critical review of cloud security testing. Moreover, gaps in recent related publications are revealed, testing tools and offers of software test automation. The prospective research implications are pointed out to foster the understanding and relations of current research fields.

Index Terms - Cloud Computing, Cloud Testing, Security, Vulnerability

1. Introduction

Cloud computing is a trendy and state-of-the-art solution in the information technology sector. Especially, organisations benefit from particular advantages of cloud computing like increased scalability and portability resulting in enhanced efficiency and cost reduction (Singh et al., 2016). Research on the identification of challenges and benefits of cloud computing has been done since 2008 (Buyya et al., 2008; Armbrust et al., 2010;

Riungu-Kalliosaari et al., 2016). In cloud computing, applications are hosted, deployed, and delivered as services over the Internet. New cloud application services can be developed by tailoring existing ones, while hiding the complexity of the underlying implementation. Cloud applications may be able to adapt to changes in their environment, which should be highly secure and reliable. The infrastructure on which cloud applications are built is characterized by power, storage and virtualization. Cloud Computing Security Lab focus on the software testing of the cloud and seeks to formulate new approaches, tools and techniques for improving the testability of cloud based applications.

Cloud testing refers to testing of resources such as hardware, software and etc. that are available on demand basis. Even the testing here can be viewed “as a service”. For cloud services offerings, it is essential to make sure that the services (also defined as product) not only meet its functional requirements but also non-functional requirements. Functional requirements describe the features, functioning, and usage of a product, system and software from the perspective of the product and its user. Although referred to as "requirements," they really are a form of design, albeit high-level. Functional requirements also often are called "functional specifications," and "specification" is a synonym for design. Non-functional requirements are not non-functional at all. Rather, they describe various quality factors, or attributes, which affect the functionality's effectiveness. They do not exist in the abstract but only with respect to relevant functionality. They are often called "ilities," because many end in "ility," such as, usability, reliability, and maintainability.

Cloud testing is one of the newest forms of testing in IT industry. Applications designed for cloud usage run remotely from the Internet. Many cloud applications will have large numbers of users, so performance testing and load testing are particularly important for cloud application. However, cloud applications are also subject to quality problems, security problems, and usability problems. As a general rule the applications intended to operate in the cloud will have a normal set of "ground" tests prior to "cloud" testing, which usually occur at about the same time as integration or system testing.

Security testing is defined as the process of testing specialized towards security, where testing is the process of exercising the system to verify that it satisfies specified requirements and to detect errors. SaaS testing comprises of validating SaaS applications with respect to business workflows, multi-tenancy, integrity, reliability, ease of deployment, scalability, availability, accuracy, deploy ability, ease of use, testability, portability live updating. All these applications are tested with cloud based resources and among the testing criteria mentioned above the focus will be on three key components they are performance, compatibility and security. Security testing is a great resource for identifying and rectifying vulnerabilities or flaws in applications so that they are less susceptible to compromise in the event of cyber-attacks.

2. Cloud Testing Requirements

2.1 Basic Requirements

The following is the 5 major things to consider for cloud testing.

a. Functional Testing

Functional testing of both internet and non-internet applications can be performed using cloud testing. The process of verification against specifications or system requirements is carried out in the cloud instead of on-site software testing.

b. Browser Performance Testing

Finding out thresholds, bottlenecks & limitations is a part of testing. For this, testing performance under a particular workload is necessary. By using cloud testing, it is easy to create such environment and vary the nature of traffic on-demand. This effectively reduces cost and time by simulating thousands of geographically targeted users.

c. Load & Performance Testing

Load testing of an application involves creation of heavy user traffic, and measuring its response. There is also a need to tune the performance of any application to meet certain standards. However, a number of tools are available for that purpose.

d. Latency & Bandwidth Testing

Cloud testing is utilized to measure the latency between the action and the corresponding response for any application after deploying it on cloud and measure the bandwidth capability.

e. Stress Testing

Stress Test is used to determine ability of application to maintain a certain level of effectiveness beyond breaking point. It is essential for any application to work even under excessive stress and maintain stability. Stress testing assures this by creating peak loads using simulators. But the cost of creating such scenarios is enormous. Instead of investing capital in building on-premises testing environments, cloud testing offers an affordable and scalable alternative.

f. Compatibility Testing

Using cloud environment, instances of different Operating Systems can be created on demand, making compatibility testing effortless.

2.2 Requirements in IT Security

Another consideration is the need to decide between white box or black box testing. In black box testing, the penetration tester knows as little about the system as a real-world

hacker would know. This is advantageous because, as we discover and exploit vulnerabilities, no one can challenge our report by claiming "an attacker wouldn't know to do that." On the other hand, white box testing is advantageous in that it is much faster. Not only is reconnaissance and server discovery accelerated, it's easier to prioritize test efforts.

A big challenge to cloud security testing can be the lack of application logging to aid in focusing and enhancing your test efforts. Performing security testing in an isolated development environment means we will be able to tail logs and see evidence of your attacks' outcomes. In a cloud environment, we will rarely be grated this level of access. Therefore, we will only be able to gauge attack success by the application's behaviour. Some tests are such that providing input into control "A" on screen "Z" will result in invalid data on page "P". Be familiar with the data flow within our app and expect to have to poke all around the app to complete our testing. In conclusion, security testing in the cloud does change things, but it's not impossible. It's important to plan ahead, to communicate the changes in our test strategy, and to set appropriate expectations with our management. Above all, it is critical to communicate before and during our testing—primarily with our cloud provider, but also with our IT and security organizations.

2.3 Advance Requirements

Testing in a cloud has to not only ensure that the functional requirements are met, but a strong emphasis needs to be laid on non-functional testing as well.

a. Functional Testing

Goes without saying, that functional testing has to be performed to make sure that offering provides the services that the user is paying for. Functional tests ensure that the business requirements are being met. Some of the functional tests are described below.

- **Data Migration Testing:** This makes sure that the several of data modules function correctly with one another, thus making sure that their data in place.

- **Integration Testing:** Here the cloud based solution is to integrate all software tools to ensure the integration between each component and clouds working properly.

- **Interoperability Testing:** Any application must have the flexibility to work without any issues not only in different platforms, but also must work seamlessly when moving from cloud infrastructure to another.

b. Non-Functional Testing

Non-functional tests mainly focus on the web application based tests ensuring that they meet the desired requirements. Here are few forms of non-functional tests discussed below:

- **Availability Testing:** The cloud supervisor/ vendor has to make sure that the cloud is available round the clock. As there could be many mission critical activities going on, the administrator has to make sure that there is no adverse impact to the consumers

- **Multi Tenancy Testing:** Here, multiple users use a cloud offering. Testing must be performed to ensure that there is sufficient security and access control of data when multiple users are using a single instance.

- **Performance Testing:** Verification of the response time needs to be done to ensure that everything is intact even when there is a large number of a request to be satisfied. The network latency is also one of the critical factors to evaluate performance. Also, workload balancing needs to be done when there is a reduction in load, by decommissioning resources. Thus, load and stress testing are done in the cloud offering to make sure applications are performing optimally with increase/decrease in load and stress.

- **Security Testing:** Since with the cloud everything is available anytime, it's essential to make sure that all user sensitive information has no unauthorized access and the privacy of users remains intact. When maintaining the applications in cloud, user data integrity must also be verified.

- **Disaster Recovery Testing:** As already stated in availability testing, the cloud has to be available at all times and if there are any kind of failures like network outages, breakdown due to extreme load, system failures and etc., whilst measure how fast the failure is indicated and any data loss during this period.

3. Cloud Testing Tools

In this section, some of the different tools used in various kinds testing performed in a cloud are mentioned. The details of the tools are out of the scope of this article. Many of the tools are basically used for performance, load, stress testing. Some of these tools below can also be used for:

3.1 Web Functional / Regression Testing Tools:

- **SOASTA CloudTest:** CloudTest makes it easy to test to any level of expected usage – and beyond. A single interface allows to control the ramp and scale of user traffic from locations around the world and measure the effects in real time. Network

emulation lets a model to test with multiple connection types. Whether it testing 100 users or a million, the CloudTest platform helps to get the most out of every test by seamlessly spanning dozens of global cloud providers, and simulating web and mobile user traffic more accurately.

- **LoadStorm:** A web-based load testing tool/service from CustomerCentrix, LLC, as a distributed application that leverages Amazon Web Services to scale on demand with processing power and bandwidth as needed. Tests for web and mobile can be built using the tool in such a way as to simulate a large number of different users with unique logins and different tasks.

- **CloudTestGo:** An on-demand Performance Testing solution using JLTT, CSS' home-grown Performance Testing tool. It enables load testing of all products including web and non-web applications and services, in the cloud. It can setup quickly and cost-effectively create a real-world Load Testing environment without having to invest in complex infrastructures, new hardware or expensive software licensing.

- **AppPerfect:** It is a fully Automated Load test, Stress test and Performance Test solution that is easy to use and cost effective. Most application performance and stability issues arise only when the server is stressed with a high user load. AppPerfect Web Load Test helps to design and simulate thousands of users in a realistic manner which can be used to load test at application infrastructure for performance, reliability and scalability.

- **Jmeter:** Java desktop application from the Apache Software Foundation designed to load test functional behaviour and measure performance. Originally designed for testing Web Applications but has since expanded to other test functions; may be used to test performance both on static and dynamic resources (files, Servlets, Perl scripts, Java Objects, Data Bases and Queries, FTP Servers and more). Can be used to simulate a heavy load on a server, network or object to test its strength or to analyze overall performance under different load types; can make a graphical analysis of performance or test server/script/object behaviour under heavy concurrent load.

- **Cloudslueth:** CloudSleuth, a free cloud monitoring service from Compuware, does the detective work on public cloud performance. With identical sample application hosted on several public cloud service provider networks, response time and availability of the application is continuously monitored from over 30 Internet backbone nodes across North America, South America, Europe, and Asia Pacific. The performance data can give a snapshot of user experience of cloud services across the globe.

- **WebdriverIO:** WebdriverIO control a browser or a mobile application with just a few lines of code. It can be test code will look simple, concise and easy to read. The integrated test runner let will write asynchronous commands in a synchronous way so don't need to care about how to handle a Promise to avoid racing conditions.

- **Selenium:** Selenium is a portable software testing framework for web applications. Selenium provides a record/playback tool for authoring tests without learning a test scripting language (Selenium IDE). It also provides a test domain-specific language (Selenese) to write tests in a number of popular programming languages, including C#, Groovy, Java, Perl, PHP, Python, Ruby and Scale.

3.2 Cloud Security Testing Tools

- **Nessus:** Nessus is a network vulnerability scanner that uses the Common Vulnerabilities and Exposures architecture for easy cross-linking between compliant security tools. Nessus employs the Nessus Attack Scripting Language (NASL), a simple language that describes individual threats and potential attacks. Nessus has a modular architecture consisting of centralized servers that conduct scanning, and remote clients that allow for administrator interaction. Administrators can include NASL descriptions of all suspected vulnerabilities to develop customized scans.

- **Wireshark:** Wireshark is a free and open source packet analyzer. It is used for network troubleshooting, analysis, software and communications protocol development, and education. Wireshark is cross-platform, using the Qt widget toolkit in current releases to implement its user interface, and using pcap to capture packets and is very similar to tcpdump, but has a graphical front-end, plus some integrated sorting and filtering options.

- **Nmap:** Nmap ("Network Mapper") is a free and open source utility for network discovery and security auditing. It useful for tasks such as network inventory, managing service upgrade schedules, and monitoring host or service uptime. It uses raw IP packets in novel ways to determine what hosts are available on the network, what services (application name and version) those hosts are offering, what operating systems (and OS versions) they are running, what type of packet filters/firewalls are in use, and dozens of other characteristics. It was designed to rapidly scan large networks, but works fine against single hosts.

3.3 Load Test and Performance Monitoring Tools

- **Perfecto Mobiles, Keynote (Test center enterprise):** Perfecto Mobile is a global provider of cloud-based testing, automation and monitoring solutions for mobile applications and websites utilizing a wide selection of REAL and emulated mobile devices. Mobile Cloud Platform enables developers and testers to access and control a comprehensive range of the latest mobile smart phones and tablets connected to live networks around the world. Keynote Mobile Testing TCE provides enterprise platform for mobile (manual and automated) testing and service monitoring of enterprise mobile apps and websites. It gives easy remote access to all popular smart phones and tablets. Test developer can create hundreds of automation scripts and leverage them across multiple devices. It also enhances the tests scripts by using Java API to access much functionality of the remote devices. It also integrates with leading test tools like QTP and IBM ALM.* It cans plug-and-play local devices into their desktop computers for manual and automated testing. Devices can also be set-up in an enterprise lab environment (on-premise or in the Keynote Mobile Testing enterprise cloud) for sharing across local and remote teams

- **Monitis:** Monitis is a specialist provider of web and Cloud monitoring services that include website monitoring, site load testing, transaction monitoring, application and database monitoring, Cloud resource monitoring, and server and internal network monitoring within one easy-to-use dashboard. Monitis can provide of choice to increase uptime and user experience of their services and products. Monitis are fast to deploy, feature-rich in technology and provide a comprehensive single-pane view of on-premise and off-premise infrastructure and

- **BrowserMob:** BrowserMob Proxy is a simple utility that makes it easy to capture performance data from browsers, typically written using automation toolkits such as Selenium and Watir. BrowserMob Proxy can capture performance data for web apps (via the HAR format). Used to collect the performance data from the client side.

- **GFI:** GFI Software develops easier, smarter and affordable enterprise-class IT solutions for businesses. Their solutions enable IT administrators to easily and efficiently discover, manage and secure their business networks, systems, applications and communications wherever they exist. GFI is committed to its thousands of customers worldwide to deliver the trusted expertise, right-sized and smartly engineered IT solutions with a strong focus on security excellence. It also gives a complete picture of installed applications; hardware on your network; mobile devices that connect to the Exchange servers; the state of security applications (antivirus, anti-spam, firewalls, etc.); open ports; and any existing shares and services running on machines.

- **CloudHarmony:** CloudHarmony provides objective, impartial and reliable performance analysis to compare cloud services. It to be an impartial and reliable source for objective cloud performance analysis. It not affiliated with or owned by any cloud provider. It publish report on CDN and DNS services, the standard is now 100 percent uptime, while for other services, brief interruptions remain common.

- **InterMapper**: InterMapper is a cross-platform, network monitoring, and network mapping program. It comes with a variety of network probes based on ping, SNMP, http and other network protocols used to monitor the state of networked devices and servers. It displays the status of the devices it monitors in maps or lists. It also supports alarms for devices that have disappeared from the network or which are in a warning state, and can send alerts via email, pager, console alerts, or script execution

- **Blaze Meter**: Self-service, on-demand, cloud-based load testing. Simulate any user scenario for webapps, websites, mobile apps or web services. Launch a single dedicated server or a cluster of 100. Apache JMeter compatible - pre-configured JMeter environments with up to 144 CPU cores and 500 GB of memory. Set geo locations from among choices worldwide. Set up tests, access test results, view test reports, compare past test reports and more, all on a unitary console. Generate traffic using public cloud providers or install the on-premise load generator software on your own machines and test behind the firewall on your internal network. Free tools and resources for tips and tricks to optimize website and app performance.

4. Challenges

As exciting as cloud sounds, all is not hunky-dory here. There are some challenges with relying and using cloud as an infrastructure as well. Let's take a look at some of the primary concerns using the cloud.

Challenge #1: With everything available on demand to any user, security is a primary issue for the businesses as currently there is still a lot of discussion and research going on in the industry to set up security standards. User privacy protection, security standards on cloud, security of applications running within the cloud, security testing techniques are some of the primary issues that need to be addressed in the cloud infrastructure

Challenge #2: Another big challenge is the performance of an application in a cloud: specifically, in private clouds. It will be shared across many users and hence could lead

to delays. Also in case of some maintenance or outage related activities, the bandwidth may seem insufficient.

Challenge #3: Sometimes for testing purposes, we require certain configurations: with respect to servers, storage or networking which may not be supported by the cloud provider. This sometimes makes it difficult to emulate customer environments.

Challenge #4: Another commonly faced challenge is with respect to integration testing whereby the testers test the network, database, servers, etc. In such situations, the tester will not have control on the underlying environment. Secondly, the challenge is doubled when there has to be an interaction between these components because the tester will have to anticipate risks like crashes, network breakdown or servers going kaput. Cloud computing has today become one of those "big bangs" in the industry. Most organizations are now leaning to adopting the cloud because of its flexibility, scalability and reduced costs. The following table lists the main challenges in cloud computing environment.

Challenge #5: Having an automatically set testing environment according to user preferences is essential for cloud based software testing but is a challenge to engineers because of the lack of cloud enablement in several existing testing tools. Moreover, several cloud providers offer restricted. Configuration capabilities for their cloud service which leads to constraints in emulating dynamic testing environments.

Challenge #6: Performance testers usually execute their test plans on pre-fixed environments with agreed upon metrics, evaluation parameters, etc. But for cloud testing, the effect of dynamic scalability plays the role of a villain, if you were to ask us. On-demand scalability poses extra loads on testing tools to execute new test cases in real time without prefixed metrics or evaluation statistics.

Challenge #7: With lack of proper control of the underlying cloud environment, testers face a huge challenge there is interaction between these components.

Challenge #8: Due to the lack of universal standards in integrating public resources on the cloud with internal data architectures of organizations, there is a serious challenge to testing teams to maintain dynamic testing environments to offer testing as a service. Changing vendors would lead to the requirement of new solution architectures and platform modifications for test tools to operate smoothly.

Challenge #9: Testing teams often prepare test cases and scenarios with pre-determined data sets but when testing is offered as a dynamic cloud service, then the costs involved in encrypting test data on cloud systems have to be considered since they need to evaluate the security aspects of cloud testing as well.

5. Limitations and challenges on Security Cloud Testing

In the early state of the newly emerged cloud computing paradigm, most researchers focused on a broader and coherent understanding, including definitions, challenges and benefits (Armbrust et al., 2010; Riungu-Kalliosaari et al., 2016). Subsequently, security of cloud environments became one of the most crucial concerns in adopting and using the new technology (Ali et al., 2015; Singh et al., 2016). The recent survey of RightScale Inc. (2016) revealed, that security challenges are the second highest concerns in cloud computing. Distributed systems are possible targets for attacks causing radical extra charges such as data modifications or downtimes. Data loss or leakage represents 24.6% and cloud-related malware 3.4% of threats causing cloud outages (Ko & Lee, 2013). Most of the software security incidents are exploited vulnerabilities. Hence, Akhgar (2016) recommends developing security metrics to identify vulnerabilities. To ensure application security, security testing techniques are important and effective countermeasures for improvement (Felderer et al., 2016). Thus, implemented systems should be tested by the use of analytical techniques and engineering principles to detect security issues as early as possible (Bos et al., 2014). However, according to Shrivastva et al. (2014), is security testing one of the major challenges in cloud testing environments. Besides, Nachiyappan & Justus (2015) indicated that present cloud security testing has many open queries, such as quality assurance and security validation. The authors also stated the challenge of testing security measures in cloud environments. Kumar & Singh (2014) revealed the research issue of performing quality checks within cloud environments. Beyond, Madan et al. (2016) pointed out the need to develop an approach for cloud privacy testing. Although the body of knowledge on cloud testing is growing, the literature review reveals an enormous gap of sophisticated security testing approaches for testing the cloud. Researchers mostly focused on Test as a Service (TaaS) rather than on testing the cloud.

Security of the data is the biggest disadvantage. Storing data in a cloud means the data is, in theory, accessible to anyone and data and code are mostly stored in a remote location beyond an organization's legal and regulatory jurisdiction. Yet another challenge is that some cloud providers offer only limited types of configurations, technology, servers and storage, networking and bandwidth, making it difficult to create real-time test environments. Improper choice of cloud-based use and pricing options is another risk. While some vendors offer pay-as-you-go services, they are only cost-effective when the right plan and service provider are chosen for the anticipated needs (e.g. space vs. RAM vs. bandwidth). Costs can quickly spin out of control if resource estimates differ wildly from actual usage.

Integration testing in clouds - Although we have seen numerous published research papers addressing software integration testing issues and strategies, not much research results have been applied in the real engineering practice. One of the major reasons is the

existing software and components are developed without enabling technology and solution to support and facilitate systematic software integration. In a cloud infrastructure, engineers must deal with integration of different SaaS and applications in/over clouds in a black-box view based on their provided APIs and connectivity protocols. This could cause a lot of extra integration costs and difficulties due to the following issues:

- There is a lack of well-defined validation methods and quality assurance standards to address the connectivity protocols, interaction interfaces, and service APIs provided by SaaS and clouds APIs; and
- There is a lack of cost-effective integration solutions and framework to facilitate software application integration and assembly inside clouds and over clouds.

Infrastructure requirements: It is vital that Infrastructure requirements are rigorously set, because the very flexibility that the cloud offers for testing environments can itself be a risk if the requirements for those environments are inappropriate. Results will then be poor and negative perceptions of the cloud as a test environment will result from what was really an inattention to requirements.

6. Summary

Using the cloud for testing is immensely helping organizations to acquire the required tools, software licenses, infrastructures at a very low cost without having to set it up themselves and then worry about its maximum utilization.

For Cloud Security Lab/Test Lab: NeXpose is a tool for auditing cloud infrastructure. It also provides Vulnerability Management for cloud assessment. NeXpose can scan the entire infrastructure, application, database or Virtual Machine to detect vulnerability across them. Vulnerability management tools include the ability to detect and identify assets in an IT infrastructure, detect vulnerabilities, provide descriptions of vulnerabilities as well as links to patches and other forms of remediation, and generate a host of reports -- all from a central console. The other tool will be proposed using SAINT, Burpsuite, NetSkope, Qualys, Retina, VIM and GFI LanGuard.

Though offering testing as a service through cloud based automation throws up a lot of challenges down the line, organizations continue to focus on this trend, thanks to its numerous benefits. This is evident from another World Quality Report insight which shows that the share of testing budgets for transformational projects has increased from 41% in 2011 to 43 % in 2013.

The core philosophy behind a majority of the challenges stated is because of the underlying characteristics of the cloud platform. A closely-knit approach with the cloud

service provider can help to minimize this risk considerably as testing teams can built a solid cloud platform to deploy their tools.

Partnering with a reliable testing service provider is the perfect answer for organizations to eliminate testing related challenges and overheads especially with regards to automated test services. With decades of experience in offering world class testing services to a plethora of global giants, Cigniti is your road to eliminating testing overheads once and for all. Moreover, functional testing may take specific forms such as:

i. User acceptance testing: Also, known as beta testing, this category of testing evaluates an application's performance in the real world among its intended audience. It has the added benefits of helping to minimize change requests down the road and keeping overall project costs to a minimum. User acceptance testing can also build goodwill with end users and improve their satisfaction with the software in question.
ii. Interoperability testing: With interoperability testing, testers are looking to see that programs work with others on a variety of platforms. Many applications are now cross-platform and must meet mission-critical requirements such as exchanging data between different medical records systems. With the emergence of cloud computing, this type of testing may also check whether software can hand off workloads across both cloud and on-premises infrastructure.
iii. System verification testing: Here we get into slightly more technical testing. System verification may include code audits, revisions to any documentation, and testing of hardware and software components under normal as well as adverse environmental conditions. Voting machines are a prime example of appliances that require thorough system verification.
iv. How has the emergence of cloud computing affected the general practice of functional testing? We can already see signs of its influence on the applications testing that have been "hybridized," i.e. designed to leverage both internal and external (public cloud) IT resources. More specifically, these programs require careful attention to the interoperability of all the systems involved, as we noted above.

The cloud computing paradigm is a nascent technology with many benefits for organisations. On the other side, security of clouds is still one of the major concerns of clients to adopt and use the new computing paradigm. The review of the recent academic literature revealed that cloud security in general is still a major concern in the industry and academics alike. To make it clearer, a data breach revealed vulnerabilities at Yahoo! Inc., whereby 32 million user accounts were accessed by forged cookies to log in without a password (Yahoo! Inc., 2017). Another example is the Distributed Denial-of-Service (DDoS) attack against Dyn, which was just recently acquired by Oracle, causing a major breakdown of its Domain Name System (DNS) servers also affecting enterprises relying on SaaS (York, 2016). Besides, the survey unfolded that security and related testing activities are current research fields with a lot of open queries. Thus, many academic papers have been published to identify and address challenges in cloud security, vulnerabilities and threats. However, most of the researchers focused on TaaS rather

than on testing the cloud. Hence, this survey reveals a current gap in academic research in terms of testing the cloud security using an appropriate approach for SaaS applications. The authors imply to conduct further research on cloud security testing approaches, especially in SaaS and public environments, whereby internal and external factors need to be differentially considered.

References

Zafar, F., Khan, A., Ur, S., Malik, R., Ahmed, M., Anjum, A., Khan, M.I., Javed, N., Alam, M. & Jamil, F. (2017). A survey of cloud computing data integrity schemes: Design challenges, taxonomy and future trends. Computers & Security. 65. p.pp. 29–49.

Yahoo! Inc. (2017). Annual Report. [Online]. Delaware. Available from: https://investor.yahoo.net/secfiling.cfm?filingID=1193125-17-65791&CIK=1011006&soc_src=mail&soc_trk=ma. [Accessed: 27 March 2017].

Vohradsky, D. (n.d.). Cloud Risk - 10 Principles and a Framework for Assessment. Retrieved from http://www.isaca.org/Journal/archives/2012/Volume-5/Pages/Cloud-Risk-10-Principles-and-a-Framework-for-Assessment.aspx

The 10 Worst Cloud Outages (and What We Can Learn From Them). (2011, June 27). Retrieved from http://www.infoworld.com

Security of OpenStack Cloud. (n.d.). Retrieved from http://www.stratoscale.com

Security Authorization Process Guide. (n.d.).

Roadmap NIST, cloud application security and operations policy. (2015, July).

Red-hat OpenStack platform. (n.d.).

Private Cloud Providers Comparison. (n.d.). Retrieved from http://www.tomsitpro.com

National standard for cloud NIST. (n.d.).

Mirantis Reference Architecture VHC for Cloud Native Apps. (2016, March 3).

ISACA. (2011). IT Control Objectives for Cloud Computing: Controls and Assurance in the Cloud. Retrieved from http://www.isaca.org/cloud

ISACA. (2010). Business Model for Information Security. Retrieved from http://www.isaca.org/bmis

Goldsmith, R. F. (n.d.). Search software quality. Retrieved from http://searchsoftwarequality.techtarget.com/answer/Functional-vs-non-functional-requirements-what-is-the-difference

Enisa. (2009). Cloud Computing: Benefits, Risks and Recommendations for Information Security. Retrieved from http://www.enisa.europa.eu

Cloud Pentest. (n.d.).

Boucher. (2013). Boucher Testing Challenge.

Boucher, P. f. (n.d.). Security Authorization Process Guide Security Software.

Hofmann, D. W. (2010, November). Cloud Computing: The Limits of Public Clouds for Business Applications', IEEE Internet Computing.

Blake, M. B. (2010, November). Service-Oriented Computing and Cloud Computing: Challenges and Opportunities', IEEE Internet Computing.

Architecture, N. S. (2013, May 15). Architecture, NIST Security.

Boucher. (2013, June 18). Standard Roadmap NIST.

Penetration Testing Guidance. (2015, March).

Kirsch, B. (2015, April 7). Private cloud provider comparison. Retrieved from http://www.tomsitpro.com/articles/private-cloud-providers-comparison,2-899.html

Roadmap NIST. (July, 2015).

Julie Mathew, L. K. (2016, August 22). Best practice configure IBM cloud manager with OpenStack. Retrieved from http://www.ibm.com/developerworks/cloud/library/cl-best-practices-configure-ibm-cloud-manager-with-openstack-trs/index.html

York, K. (2016). Dyn Statement on 10/21/2016 DDoS Attack. [Online]. 2016. Available from: http://dyn.com/blog/dyn-statement-on-10212016-ddos-attack/. [Accessed: 27 March 2017].

A·P·U
PRESS
ASIA PACIFIC UNIVERSITY
OF TECHNOLOGY & INNOVATION

eISSN 2600-7304