TDS 3301 Data Mining

Trimester 1

GROUP ASSIGNMENT

Prepared by :

Name	Student ID	Phone number
Chang Kai Boon	1181101282	011-29946989
Soe Zhao Hong	1181101614	018-9434610

Question 1

Domain Selected : Healthcare

Table 1: Problems solved by researchers

Author	Identify Covid-19 infection	Predict the incident of Covid-19 infection	Predict the recovery of Covid-19 infected patients	Risk prediction of pandemic of the patient	Prediction on patients' adverse effects after covid vaccination	Study of Covid-19 vaccination development
Ahmad et al. (2021)					✔	
Ahouz et al. (2021)		✔				
Mohammad et al. (2020)		✔				
Radenliev et al. (2020)			✔	✔		✔
Abdulkareem et al. (2021)						✔
Subudhi et al. (2021)				✔		
Mohammad Abu et al. (2021)			✔			
Arpaci et al. (2021)		✔				
Lalmuanawma et al. (2020)	✔	✔				✔
Sharma et al. (2020)	✔					
Muhammad et al. (2020)		✔				
Kassania et al. (2021)	✔					
Wu et al. (2020)	✔					
Somasekar et al. (2020)				✔		
Al-Najjar et al.(2020)			✔	✔		

Discussion:

Table 1 shows the problems solved by different researchers when dealing with Covid19. From the table, most of the reports have been centering around identifying and predicting the incident of Covid-19 infection. This can be shown by the number of counts where 5 out of 15 papers focused on predicting the incidence of Covid-19 infection while 4 out of 15 papers focus on identifying Covid-19 infection using data mining techniques. For example, Ahouz et al.(2021) predict the Covid19 infected people in each geographic region as well as each continent in the coming 2 weeks to better manage the disease. On the other hand, Kassania et al. (2021) built a model for automatic Covid19 classification based on Chest X-Ray and CT scan as an assistant tool to help the doctors in identifying Covid-19 patients.

Besides, the risk prediction of the pandemic of the patient is the second most problem that the researchers are trying to solve using data mining techniques. There are 4 out of 15 papers focus on risk prediction. From the paper written by Somasekar et al. (2020), the team worked on patient risk prediction of pandemic based on risk factors such as patients characteristics, comorbidities, initial symptoms, vital signs for prognosis of disease and forecasting fatality rate. This can help the doctor to evaluate the risk of each patient and determine which patients should be sent to ICU based on the clinical data collected.

On top of that, prediction of recovery for Covid-19 infected patients and the study of Covid-19 vaccination development are the third and fourth problem as reported by the researchers. This can be shown by the number of counts where 3 out of 15 papers focus on these two problems respectively. A classifier has been built by Mohammad et al. (2021) to predict the status of recovery of Covid-19 patients. In order to fight against the Covid-19 pandemic, we definitely need an efficient vaccine that can be distributed equally and broadly so that everyone can get the vaccine. Thus, Abdulkareem et al. (2021) work on the Covid 19 world vaccination progress to find the best algorithms for vaccine distribution.

Last but not least, there is one of the papers in the table above that focuses on prediction of patients' adverse effects after covid vaccination. There is no doubt that vaccination is one of the most effective and well-accepted approaches to control pandemics, however, a small group of people will experience severe post-vaccination side effects. Therefore, this study is particularly useful as it can help to identify the most significant features of patients' past medical history that can give rise to adverse effects of covid vaccination and predict the need for hospitalization and treatment for patients after vaccination.

Table 2: Data mining techniques used by researchers

Author	Random Forest	Least-Square Boosting	Linear Regression	MCA	Decision Tree	KNN	Naive Bayes	SVM	Logistic Regression	XGBoost	ResNet	LSTM	ANN	DenseNet	VGG
Ahmad et al. (2021)	✔							✔		✔					
Ahouz et al. (2021)		✔													
Mohammad et al. (2020)			✔									✔			
Radenliev et al. (2020)				✔											
Abdulkareem et al. (2021)	✔				✔	✔	✔								
Subudhi et al. (2021)			✔			✔	✔	✔							
Mohammad Abu et al. (2021)					✔		✔								
Arpaci et al. (2021)					✔				✔						
Lalmuanawma et al. (2020)	✔							✔		✔	✔	✔			
Sharma et al. (2020)											✔				
Muhammad et al. (2020)					✔		✔	✔	✔				✔		
Kassania et al. (2021)	✔				✔					✔	✔			✔	✔
Wu et al. (2020)	✔														
Somasekar et al. (2020)	✔										✔	✔		✔	✔
Al-Najjar et al.(2020)													✔		

Discussion:

Table 2 depicts the information about different data mining techniques used in work related to Covid-19. From the above table, Random Forest has been widely accepted in most of the reported research work. Based on the 15 papers, 6 out of them employed Random Forest. The focus areas are Covid-19 prediction and identification. From the paper written by Wu et al. (2020), Random Forest was used to extracte the blood indices from blood test data to build the final assistant discrimination tool for Covid19 indentication with sensitivity, specificity and accuracy of 95%, 97% and 96%.

More recent work has also reported on using Decision Tree to predict Covid-19 infection and recovery of Covid-19 patients. This can be shown by the number of counts where 5 out of 15 papers using Decision Tree as the methodology. For example, in the paper written by Mohammad et al. (2021), the process of generating classification rules was based on the decision tree algorithm and was evaluated for prediction of the maximum and minimum of days for the recovery of Covid-19 patients. On the other hand, Abdulkareem et al.(2021) discovered that the Decision Tree outperforms other algorithms such as KNN and Naive Bayes in terms of time and accuracy for Covid-19 world vaccination progress prediction.

Moreover, SVM and Naive Bayes were also used to predict the Infection of Covid-19. This can be shown by the number of counts where 4 out of 15 papers focused on predicting Covid-19 infection using SVM and Naive Bayes respectively. From the paper written by Muhammad et al. (2020), the result of performance evaluation of model showed that SVM has the highest sensitivity of 93% and Naive Bayes has the highest specificity of 94% compared to other machine learning algorithm such as Logistic Regression and ANN. Furthermore, Linear regression and Logistic Regression have also been used by two of the papers above in predicting Covid-19 infection.

On top of that, deep learning models such as ResNet, DenseNet, VGG and LSTM were employed by many researchers in identifying and predicting risk and Covid-19 Infection. This is due to the fact that deep learning algorithms are found to have more potential, robustness and advance among the other algorithms. For example, Kassania et al. (2021) compared popular deep learning based feature extraction frameworks for automatic Covid-19 detection on Chest X-Ray and CT , DenseNet and ResNet achieved the performance with 99% and 98% of accuracy respectively.

In conclusion, data mining techniques have been applied widely by researchers when dealing with Covid-19 and significantly improved the identification, prediction, medication and vaccine development process for the Covid-19 pandemic. However, there are some challenges and limitations such as limited datasets and models are not deployed enough to show real-world operations. Therefore, future work needs to be carried out to focus on these two challenges so that the model's performance can be more robust and accurate.

Question 2

Table 3: Tools and Programming Languages for Data Mining

Software /Tools	Open Source	Free	Drag & Drop	Platform Independent	Array of Packages	Ease of Learning	Data Management
python	✓	✓		✓	✓	✓	✓
R	✓	✓		✓	✓		✓
PowerBI	✓	✓	✓	✓	✓	✓	✓
excel	✓			✓		✓	
SAS			✓	✓	✓	✓	✓
RapidMiner	✓		✓	✓	✓	✓	✓
IBM SPSS Modeler	✓		✓	✓	✓	✓	✓
Orange	✓	✓	✓	✓	✓	✓	✓
KNIME	✓	✓	✓	✓	✓	✓	✓
MATLAB				✓	✓	✓	✓

Discussion:

Based on table 3, three types of programming languages are presented which are Python, R, and MATLAB are normally used for data mining. These three programming languages are platform-independent, provide a lot of library packages, and provide a good data management service. Python and R are open sources and free programming languages but MATLAB is not open source and free. So, Python and R are usually used by a lot of programmers since both of these languages are open source and free. Although R language is free and open-source, R language is difficult to learn for the beginner compared with MATLAB and Python. Python can manage a lot of work because Python is a general-purpose programming language but its performance is the same as the R language which is very slow. There are 7 types of data mining tools such as Excel, SAS, RapidMiner, IBM SPSS Modeler, Orange, and KNIME are presented in table 3 which all of these tools are platform-independent. Besides, all of these tools are open source except SAS but only PowerBI, Orange, and KNIME are free to users. Others only provide some free trials for users. Drag and drop, a feature that is very user-friendly, and most of the tools will provide this feature to users, but only Excel does not provide the drag and drop. In addition, Excel also does not link with other libraries from other languages and does not provide good data management. Although Excel does not provide several features, it is easy to learn and view for beginners which is the same as other data mining tools. Based on the table, it shows that other data mining tools offer several libraries and good data management but most of them also have some weaknesses during the data mining process. For example, RapidMiner and KNIME will take too much memory and slow down the system. PowerBI has a very bulky user interface where some functions will block the important view and this makes it not user-friendly. IBM SPSS Modeler is very expensive and has limited functionality which is very similar to Excel. Then, Orange has not facilitated the analysis and prediction of live data.

References

[1] Ahamad, M. M., Aktar, S., Uddin, M. J., Rashed-Al-Mahfuz, M., Azad, A. K. M., Uddin, S., Alyami, S. A., Sarker, I. H., Liò, P., Quinn, J. M. W., & Moni, M. A. (2021, January 1). *Adverse effects of COVID-19 VACCINATION: Machine learning and statistical approach to identify and CLASSIFY incidences of morbidity AND Post-vaccination reactogenicity*. medRxiv. https://doi.org/10.1101/2021.04.16.21255618.

[2] Ahouz, F., & Golabpour, A. (2021). Predicting the incidence of COVID-19 using data mining. *BMC Public Health 2021 21:1*, *21*(1), 1–12. https://doi.org/10.1186/S12889-021-11058-3

[3] Al-Najjar, H., & Al-Rousan, N. (2020). A classifier prediction model to predict the status of Coronavirus CoVID-19 patients in South Korea. *European Review for Medical and Pharmacological Sciences*, *24*(6), 3400–3403. https://doi.org/10.26355/EURREV_202003_20709

[4] Abdulkareem, N. M., Abdulazeez, A. M., Zeebaree, D. Q., & Hasan, D. A. (2021). COVID-19 World Vaccination Progress Using Machine Learning Classification Algorithms. *Qubahan Academic Journal*, *1*(2), 100–105. https://doi.org/10.48161/QAJ.V1N2A53

[5] Arpaci, I., Huang, S., Al-Emran, M., Al-Kabi, M. N., & Peng, M. (2021). Predicting the COVID-19 infection with fourteen clinical features using machine learning classification algorithms. *Multimedia Tools and Applications 2021 80:8*, *80*(8), 11943–11957. https://doi.org/10.1007/S11042-020-10340-7

[6] Lalmuanawma, S., Hussain, J., & Chhakchhuak, L. (2020). Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. *Chaos, Solitons & Fractals*, *139*, 110059. https://doi.org/10.1016/J.CHAOS.2020.110059

[7] Kassania, S. H., Kassanib, P. H., Wesolowskic, M. J., Schneidera, K. A., & Detersa, R. (2021). Automatic Detection of Coronavirus Disease (COVID-19) in X-ray and CT Images: A Machine Learning Based Approach. *Biocybernetics and Biomedical Engineering*, *41*(3), 867–879. https://doi.org/10.1016/J.BBE.2021.05.013

[8] Mohammad Abu-dalbouh, H., & Abdullah Alateyah, S. (2021). PREDICTIVE DATA MINING RULE-BASED CLASSIFIERS MODEL FOR NOVEL CORONAVIRUS (COVID-19) INFECTED PATIENTS' RECOVERY IN THE KINGDOM OF SAUDI ARABIA. *Journal of Theoretical and Applied Information Technology*, *99*(8). www.jatit.org

[9] Mohammad, S., Ayyoubzadeh, S. M., Zahedi, H., Ahmadi, M., & Kalhori, S. R. N. (2020). Predicting COVID-19 Incidence Through Analysis of Google Trends Data in Iran: Data Mining and Deep Learning Pilot Study. *JMIR Public Health Surveill 2020;6(2):E18828 Https://Publichealth.Jmir.Org/2020/2/E18828*, *6*(2), e18828. https://doi.org/10.2196/18828

[10] Muhammad, L. J., Algehyne, E. A., Usman, S. S., Ahmad, A., Chakraborty, C., & Mohammed, I. A. (2020). Supervised Machine Learning Models for Prediction of COVID-19 Infection using Epidemiology Dataset. *SN Computer Science 2020 2:1*, *2*(1), 1–13. https://doi.org/10.1007/S42979-020-00394-7

[11] Radanliev, P., de Roure, D., & Walton, R. (2020). Data mining and analysis of scientific research data records on Covid-19 mortality, immunity, and vaccine development - In the first wave of the Covid-19 pandemic. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, *14*(5), 1121–1132. https://doi.org/10.1016/J.DSX.2020.06.063

[12] Sharma, S. (2020). Drawing insights from COVID-19-infected patients using CT scan images and machine learning techniques: a study on 200 patients. *Environmental Science and Pollution Research 2020 27:29*, *27*(29), 37155–37163. https://doi.org/10.1007/S11356-020-10133-3

[13] Somasekar, J., Pavan Kumar, P., Sharma, A., & Ramesh, G. (2020). Machine learning and image analysis applications in the fight against COVID-19 pandemic: Datasets, research directions, challenges and opportunities. *Materials Today: Proceedings*. https://doi.org/10.1016/J.MATPR.2020.09.352

[14] Subudhi, S., Verma, A., Patel, A. B., Hardin, C. C., Khandekar, M. J., Lee, H., McEvoy, D., Stylianopoulos, T., Munn, L. L., Dutta, S., & Jain, R. K. (2021). Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19. *Npj Digital Medicine 2021 4:1*, *4*(1), 1–7. https://doi.org/10.1038/s41746-021-00456-x

[15] *Top 10 Data Mining Tools*. (n.d.). Retrieved September 8, 2021, from https://www.jigsawacademy.com/blogs/data-science/data-mining-tools/

[16] *Top 21 Data Mining Tools*. (n.d.). Retrieved September 8, 2021, from https://www.imaginarycloud.com/blog/data-mining-tools/

[17] Wu, J., Zhang, P., Zhang, L., Meng, W., Li, J., Tong, C., Li, Y., Cai, J., Yang, Z., Zhu, J., Zhao, M., Huang, H., Xie, X., & Li, S. (2020). Rapid and accurate identification of COVID-19 infection through machine learning based on clinical available blood test results. *MedRxiv*, 2020.04.02.20051136. https://doi.org/10.1101/2020.04.02.20051136