Contribution made by the team

1. Raman Sharma took over the duty of designing the base code and following algorithms,testing out their performance and combining the algorithms so as to increase the accuracy of the output.
2. Bharat Sharma came in charge of formatting the research paper,procuring various learning materials and previous research in order to make the project as presentable as possible.
3. Rohan Bhalla, in collaboration with Raman, helped procure and mine the dataset required for the training and testing of the algorithm while providing insights on how each step taken changes the output and required results,as well as providing support to Bharat for content writing in the research paper.

The team started this project on 30th of January,2023 on account of Rohan Bhalla being a late entrant in the subject and hence,the team.

Introduction

Influenza is an acute respiratory disease caused by the influenza A, B or C virus.Although type D virus does exist,it mainly affects cattle with frequent spillover to other species. It often emerges as outbreaks and epidemics worldwide, mainly during the winter season. Significant numbers of influenza virus particles are present in the respiratory secretions of infected persons, so infection can be transmitted by sneezing and coughing via large particle droplets. The mean duration of influenza virus shedding in immunocompetent adult patients is ~5 days but may continue for up to 10 days or more, particularly in children, elderly adults, patients with chronic illnesses, and immunocompromised hosts. Influenza typically begins with the abrupt onset of high-grade fever, myalgia, headache, and malaise. These rapid manifestations are accompanied by symptoms of respiratory tract illnesses such as nonproductive cough, sore throat, and nasal discharge. After a certain affliction and period of time, influenza can affect other organs such as the lungs, brain, and heart more than it can affect the respiratory tract and cause hospitalisation.

The virus spreads as yearly outbreaks and in the past has affected at least 2 million people every outbreak and caused >200,000 deaths. Notable records of the virus exist such as “Spanish flu”, “Asian Influenza”,and ”Hong-Kong Influenza”.

Machine learning researches have reached to a point where such impactful diseases can be detected early on,playing a significant role in efficient classification and medication. This project uses various methods such as Machine Learning techniques alongwith Deep Learning networks on a dataset through which the model is trained and validated. By providing a sufficient accuracy rate, this project is helping to further integrate Machine Learning models in medication systems and improvising them. The dataset has been procured from a trustable source <>,which among many other great sources out there,provides a well-defined dataset with all the necessary information required for the model along with a good sample size, helping us to further give precision to our model.

KNN

The k-nearest neighbours algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifying technique, which uses proximity to make classifications or predictions about the grouping of individual data points. It is a classification algorithm,though it can be used for regression problems as well. For a new data point to an existing dataset or a data point within the dataset,the algorithm classifies the data point by finding the most similar training examples in the feature space.It is a lazy-learning model,meaning that it only stores a training dataset rather than undergoing a training stage. This also implies that all the computation occurs when a classification or prediction is being made. Since it heavily relies on memory to store all its training data, it is also referred to as an instance-based or memory-based learning method.

ANN

ANN is a type of machine learning algorithm that is modelled after the structure and function of the human brain. It is used for classification tasks. The basic building block of an ANN is a neuron, which takes input from other neurons or external sources, performs some mathematical operations on the inputs, and generates an output. Neurons are connected to each other through weighted connections, which are adjusted during training to improve the network's performance.ANNs are trained using a process called backpropagation, where the error between the network's output and the desired output is propagated backwards through the network, and the weights are adjusted to reduce the error. This process is repeated iteratively until the network's performance reaches a satisfactory level, although they can be computationally expensive to train and require a large amount of data to achieve good performance.

SVM

Support vector machines are a set of supervised learning methods used for classification, regression, and outlier detection. There are specific types of SVMs you can use for particular machine learning problems, like support vector regression (SVR) which is an extension of support vector classification (SVC). SVMs are different from other classification algorithms because of the way they choose the decision boundary that maximises the distance from the nearest data points of all the features. What makes the linear SVM algorithm better than some of the other algorithms, like k-nearest neighbours, is that it chooses the best line to classify your data points. It chooses the line that separates the data and is the furthest away from the closest data points as possible. It can find complex relationships between your data without you needing to do a lot of transformations on your own.

RANDOM FOREST

A random forest is a machine learning technique that’s used to solve regression and classification problems. It utilises ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems. The (random forest) algorithm establishes the outcome based on the predictions of the decision trees. It predicts by taking the average or mean of the output from various trees. Increasing the number of trees increases the precision of the outcome. A decision tree consists of three components: decision nodes, leaf nodes, and a root node. The algorithm divides a training dataset into branches, which further segregate into other branches. This sequence continues until a leaf node is attained. The leaf node cannot be segregated further. The nodes in the decision tree represent attributes that are used for predicting the outcome. Decision nodes provide a link to the leaves.

SENSITIVITY

Sensitivity,also known as Recall, is the ability of the learning algorithm which takes into account its proportion to correctly identify “True Positives” in the test dataset. It considers the amount of true positives taken against true positive and false negative results obtained from the output.

TP/(TP+FN)

SPECIFICITY

Specificity,in line with sensitivity,is the ability of the learning algorithm which takes into account its proportion to correctly identify “True Negatives” in the test dataset. It considers the amount of true negatives taken against all negative results obtained from the output.

TN/(TN+FP)

TRUE POSITIVE

These are the values in the output dataset that are predicted to be correct by the learning algorithm while being actually correct.

TRUE NEGATIVE

These are the values in the output dataset that are predicted to be incorrect by the learning algorithm while being actually incorrect.

FALSE POSITIVE

These are the values in the output dataset that are predicted to be correct by the learning algorithm while being actually incorrect.

FALSE NEGATIVE

These are the values in the output dataset that are predicted to be incorrect by the learning algorithm while being actually correct.

F1 SCORE

Defined as the harmonic mean of the model’s precision and sensitivity, it is used to find the accuracy of the learning algorithm.

TP/[TP+(½)(FP+FN)]

AUC-ROC CURVE

A metric used to depict the performance of the model graphically at different threshold values. Initially,ROC,or Record Operating Characteristic curve takes into account the rates of true positive and false positive values. AUC-ROC,or Area Under ROC Curve,then considers the area present under said curve. This value ranges from 0 to 1,taking the area present in between (0,0) and (1,1) in the curve.

PRECISION

It is the measure of how close the output results are to each other,i.e, the quality of predictions made by the algorithm. It is calculated by dividing the number of true positives by the total number of positives(true positive and false positive taken together).

TP/(TP+FP)

ACCURACY

It is the measure of how close the output results are to the desired results,i.e.,the number of correct predictions made by the algorithm. It is calculated by dividing the total number of true results by the total number of all results.

(TP+FP)/[TP+FP+TN+FN)

RESULTS

After the algorithms were applied,metrics such as recall,F1 score and confusion matrix were used in order to analyse and present the data in a meaningful and orderly manner. Based on the dataset procured and various learning techniques applied in order to obtain desired outputs,the result of this project comes as follows:

1. Applying the Linear Support Vector Machine algorithm,the True Positive Rate turns out to be 63.70, while Specificity turns out to be 78.95. The F1 score for training and test data is 70.06 and 70.03,respectively. The accuracy for this algorithm is 70.66. On its own,it is a pretty low performing algorithm when compared to subsequently applied techniques.
2. Applying the K-Nearest Neighbour algorithm,the True Positive Rate turns out to be 76.42, while Specificity turns out to be 79.90. The F1 score is 78.00,so is its accuracy. It is a well performing algorithm.
3. Applying the Associated Neural Network algorithm, the True Positive Rate turns out to be 52.50, while the Specificity turns out to be 70.80. The F1 score is 59.37. The accuracy for this algorithm is 60.83. Of all the techniques used, this is the lowest performing algorithm.
4. Applying the Random Forest algorithm algorithm, the True Positive Rate turns out to be 84.09,while Specificity turns out to be 90.95. The F1 score for training and test data is 92.91 and 87.78,respectively. The accuracy for this algorithm is 87.20. This is the highest performing algorithm of all algorithms used.

In order to obtain much higher accuracy and more significant results,these techniques were combined with each other,providing these results:

1. Applying SVM and KNN algorithms together, the True Positive Rate turns out to be 63.67, while Specificity turns out to be 78.95. The F1 score is 70.10. The accuracy of this algorithm is 70.72. In accordance with combined techniques,this is the lowest performing algorithm.
2. Applying KNN and ANN algorithms together, the True Positive Rate turns out to be 75.49, while Specificity turns out to be 76.22. The F1 score is 77.10. The accuracy of this algorithm is 75.83. The same results come out when Random Forest and KNN algorithms are combined. It is a decent performing algorithm.
3. Applying Random Forest and ANN together, the True Positive Rate turns out to be 85.68, while Specificity turns out to be 88.60. The F1 score is 87.68. The accuracy of this algorithm is 87.02. It is a high performing algorithm.
4. Applying Random Forest and SVM algorithms together, the True Positive Rate turns out to be 85.94, while Specificity turns out to be 88.99. The F1 score is 87.98. The accuracy of this algorithm is 87.35. Of all the techniques used,this is the highest performing algorithm.

Numerous viable steps have been taken in order to achieve desired results and for the project to maintain its finesse.

1. Obtaining the dataset:

The dataset containing appropriate structure and variables was procured from Bacterial and Viral Bioinformatics Resource Center. For this project, all obtained data remains constrained to humans only, in consideration of feasible and ethical reasons. The data was cleaned of necessary outliers and impurities. All blank columns were firstly taken out. Next, since the “symptoms” column contained data points in the form of key-value pairs, they were separated into rows and columns,respectively. Furthermore, all the Yes/No/Blank data points were converted into binary data,with Blank and No being considered as 0, and Yes being considered as 1. For the purpose of the algorithms, columns such as country name,patient name,email id and similar were removed. The data points in the gender column were converted to binary data,with Males getting initialised as 0 and Females getting initialised as 1. Then, only rows having all instances filled were taken into consideration,the rest were removed. Similarly, all columns having more than 95% data as null values were removed.

1. Figuring out the algorithms:

On the dataset, to obtain desired and viable results, algorithms suitable for such large comprehensible data were applied. The techniques include:

i) SVM: The data was first split into two sets of testing and training data,with 30% of the data taken up for testing. After setting up the standard scaler, parts of training and testing data were passed into it. These transformed parts, along with the other parts of the training set were used to train the model, which was then fitted upon the test set of the data.

ii) KNN: The data was split into training and testing data, with 20% of the data being used for testing. The model was trained on the training set, the value of k being taken as 5.

iii) ANN: The data was split into training and testing data, with 20% of the data being used for testing. A Dense neural network was set up with 24 nodes having rectified linear unit activation and 1 having sigmoid activation,in order to get output in the form of 0s and 1s. For compiling the model, ‘Adam’ optimizer was used which has a learning rate of 0.001. Then, the training set divided into batches of 32 and epochs taken as 100 was used to train the model.

iv) Random Forest: The data was split into training and testing data, with 20% of the data being used for testing. Then, the classifier object was set up taking 100 trees into consideration. Then, this model was trained on the training set.

Further, to obtain better results and expand on our research, two techniques were combined and taken at a time.

REFERENCE

1. Early Detection Of Influenza using machine learning techniques by Sajal Maheshwari, Anushka Sharma, Ranjan Kumar and Pratyush.
2. Spreeuwenberg P, Kroneman M, Paget J. Reassessing the Global Mortality Burden of the 1918 Influenza Pandemic. Am J Epidemiol. 2018 Dec 1;187(12):2561-2567. doi: 10.1093/aje/kwy191. PMID: 30202996; PMCID: PMC7314216.
3. W. Paul Glezen, Emerging Infections: Pandemic Influenza, Epidemiologic Reviews, Volume 18, Issue 1, 1996, Pages 64–76, <https://doi.org/10.1093/oxfordjournals.epirev.a017917>
4. J. S. Malik Peiris, Menno D. de Jong, Yi Guan.: Avian Influenza Virus (H5N1): a Threat to Human Health. Clinical Microbiology Reviews 20 (2), 243-267, (2007). DOI: 10.1128/CMR.00037-06
5. Sullivan SJ, Jacobson RM, Dowdle WR, Poland GA. 2009 H1N1 influenza. Mayo Clin Proc. 2010 Jan;85(1):64-76. doi: 10.4065/mcp.2009.0588. Epub 2009 Dec 10. PMID: 20007905; PMCID: PMC2800287.
6. Grohskopf LA, Alyanak E, Ferdinands JM, et al. Prevention and Control of Seasonal Influenza with Vaccines: Recommendations of the Advisory Committee on Immunization Practices, United States, 2021–22 Influenza Season. MMWR Recomm Rep 2021;70(No. RR-5):1–28. DOI: [http://dx.doi.org/10.15585/mmwr.rr7005a1.](http://dx.doi.org/10.15585/mmwr.rr7005a1)
7. Squires RB, Noronha J, Hunt V, García-Sastre A, Macken C, Baumgarth N, Suarez D, Pickett BE, Zhang Y, Larsen CN, Ramsey A, Zhou L, Zaremba S, Kumar S, Deitrich J, Klem E, Scheuermann RH. Influenza research database: an integrated bioinformatics resource for influenza research and surveillance. Influenza Other Respir Viruses. 2012 Nov;6(6):404-16. doi: 10.1111/j.1750-2659.2011.00331.x. Epub 2012 Jan 20. PMID: 22260278; PMCID: PMC3345175.
8. Lavrac N. Selected techniques for data mining in medicine. Artif Intell Med. 1999 May;16(1):3-23. doi: 10.1016/s0933-3657(98)00062-1. PMID: 10225344.
9. Tan AC, Gilbert D. Ensemble machine learning on gene expression data for cancer classification. Appl Bioinformatics. 2003;2(3 Suppl):S75-83. PMID: 15130820.
10. Battineni G, Sagaro GG, Chinatalapudi N, Amenta F. Applications of Machine Learning Predictive Models in the Chronic Disease Diagnosis. J Pers Med. 2020 Mar 31;10(2):21. doi: 10.3390/jpm10020021. PMID: 32244292; PMCID: PMC7354442.
11. Rajalakshmi V., Sasikala D., Kala A., A Predictive Analysis for Heart Disease Using Machine Learning. In: Intelligent Computing and Applications. Advances in Intelligent Systems and Computing, vol 1172. Springer (2021), DOI: 10.1007/978-981-15-5566-4\_42
12. Meyfroidt G, Güiza F, Ramon J, Bruynooghe M. Machine learning techniques to examine large patient databases. Best Pract Res Clin Anaesthesiol. 2009 Mar;23(1):127-43. doi: 10.1016/j.bpa.2008.09.003. PMID: 19449621.
13. The data used for this study was obtained from <https://www.bv-brc.org>.
14. Jane V. Carter, Jianmin Pan, Shesh N. Rai, Susan Galandiuk.: ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves,Surgery 159(6), 1638-1645, (2016). DOI: 10.1016/j.surg.2015.12.029
15. Zhu, W., Zeng, N. and Wang, N. (2010) Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS Implementations. NESUG Proceedings: Health Care and Life Sciences, Baltimore, Maryland. <http://www.nesug.org/Proceedings/nesug10/hl/hl07.pdf>
16. Hastie T., Tibshirani R., Friedman J.: The Elements of Statistical Learning. Springer Series in Statistics. Springer, New York, NY (2009). DOI: 10.1007/978- 0-387-84858-7\_15
17. Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). <https://doi.org/10.1023/A:1010933404324>
18. Venables, W.N. and Ripley, B.D. (2002) Modern Applied Statistics with S. Springer, New York, 271-300.<https://doi.org/10.1007/978-0-387-21706-2>
19. Ripley, B.D. (1996) Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge. <https://doi.org/10.1017/CBO9780511812651>
20. JOUR Veloso, Bruno, Leal, Fátima, Gonzalez-Velez, Horacio, Malheiro, Benedita, Burguillo, Juan, 2018/06/01, Scalable data analytics using crowdsourced repositories and streams, 122, 10.1016/j.jpdc.2018.06.013, Journal of Parallel and Distributed Computing.
21. Arturo López Pineda, Ye Ye, Shyam Visweswaran, Gregory F. Cooper, Michael M. Wagner, Fuchiang (Rich) Tsui, Comparison of machine learning classifiers for influenza detection from emergency department free-text reports,,Journal of Biomedical Informatics, Volume 58, 2015, Pages 60-69, ISSN 1532-0464, <https://doi.org/10.1016/j.jbi.2015.08.019>. (https://www.sciencedirect.com/science/article/pii/S1532046415001872)
22. BOOK, Marquez, Edna, Barron, Valeria, 2019/11/01, 1, 5, Artificial Intelligence system to support the clinical decision for influenza, 10.1109/ROPEC48299.2019.9057056
23. Monto AS, Gravenstein S, Elliott M, Colopy M, Schweinle J. Clinical signs and symptoms predicting influenza infection. Arch Intern Med. 2000 Nov 27;160(21):3243-7. doi: 10.1001/archinte.160.21.3243. PMID: 11088084.

FUTURE WORK

In conformity with the objective of this paper and the methodology used, we have gone to our furthest extent to use different techniques up-to-date with the standards of usage at the time of writing. For future practises to surpass us, the processes for data mining should be improved and used in such a manner that processed data is understandable to humans as well as the model, without much use of binary and/or arithmetic values. Further, given how combining two learning models produced better results than previous researches, using three or more models at the same time and use of different and/or more efficient models which have a better learning rate and require less preprocessing could be emphasised. Models could be programmed in such a way that they adapt quickly to newly comprehended data, without deviating too much from the desired results even if the data consists of some minor looked-over impurities.

Literature Review

1. In the article “Reassessing the global mortality burden of 1918 Influenza pandemic” by Peter Spreeuwenberg, Madelon Kroneman, and John Paget, the estimation has been done through two methods: direct estimation method which involves study and calculation of cause-specific deaths, while the second one is a more indirect way of approximation, and enumerates all-cause mortality. They highlight that a predefined death rate was needed to identify differences arising from calculation of excess mortality.
2. In the report “Prevention and Control of seasonal influenza with vaccines” by Lisa A. Grohskopf, MD, Elif Alyanak, MPH, Jill M. Ferdinands, PhD, Karen R. Broder, MD, Lenee H. Blanton, MPH, H. Keipp Talbot, MD, and Alicia M. Fry, MD, they observe the implications of the vaccines administered by the United States in lieu of the influenza season which coincides with the spread of SARS-COV2 virus, also known as Covid-19. They mention the arrangement and protocols needed to be followed by the authorities for controlled handout of the vaccines.
3. In the article “Influenza Research Database: an integrated bioinformatics resource for influenza research and surveillance” by R. Burke Squires, Jyothi Noronha, Victoria Hunt, Adolfo García‐Sastre, Catherine Macken, Nicole Baumgarth, David Suarez, Brett E. Pickett, Yun Zhang, Christopher N. Larsen, Alvin Ramsey, Liwei Zhou, Sam Zaremba, Sanjeev Kumar, Jon Deitrich, Edward Klem, and Richard H. Scheuermann, they have created an intricate and information-comprehensive database for the 2009 pandemic influenza A/H1N1 describing the genome-sequence data and characteristics of the virus that are potentially much use for research
4. In references 15,16,17,18 and 19, information about statistical learning, pattern recognition and neural networks, use and application of various machine learning models is presented.In references 21,22 and 23, use of artificial intelligence with medicinal research is presented, along with comparison between major classifier models of machine learning.

STUDY AREA

For the objective of this research, the data chosen comprises the symptoms the patient exhibits along with the gender, age and collection year of the patient. The data pre processing done is to escalate and expand upon such features. Furthermore, the techniques used in this regard such as encoding will help the model to understand the data better. The data is obtained from the databases of BV-BRC, and contains data of 30,531 patients identified over the years and countries.

In previous researches done by various authors, the use of singular algorithms was appropriated to carry-out the research and calculate the results and conclusions. In this research paper, we have emphasised on the use of combined learning algorithms instead of using the learning models one at a time in order to obtain more accurate results. We have shown the shortcomings of such previously-used methods and models of KNN, ANN, Random Forest and SVM taken one at a time initially in our paper. As we progress, the models were combined in regards to their better metrical scores.

Within the data used in this paper, the patient’s information such as collection year,age of patient and symptoms shown by the patient were studied and passed into the models which were then evaluated and concluded upon. The dataset also contains the type of pathogen the patient is infected with, the demographic information of the patient, the strain the patient is infected with, and many other ineffective information.

CONCLUSION

Based on the results obtained, the following conclusions can be made:

1. Combined models perform highly better than their singular counterparts.
2. The ANN algorithm is not suitable for such labelled and varied data given its inaccuracy and metrical scores.
3. The Random Forest algorithm works very well with such a large amount of data and produces high scores and accuracy, irrespective of how it is implemented.
4. Combining K-Nearest Neighbour algorithm with either Associated Neural Network algorithm or Random Forest Algorithm yields the same result.

Data used

Within the data used in this paper, the patient’s information such as collection year,age of patient and symptoms shown by the patient were studied and passed into the models which were then evaluated and concluded upon. The dataset also contains the type of pathogen the patient is infected with, the demographic information of the patient, the strain the patient is infected with, and many other ineffective information.

The dataset containing appropriate structure and variables was procured from Bacterial and Viral Bioinformatics Resource Center. For this project, all obtained data remains constrained to humans only, in consideration of feasible and ethical reasons.

Some attributes such as patient’s name, email-id and similar were removed in the beginning since they contain irrelevant information to the purpose of the project and the models used.

Data Preprocessing

The data was cleaned of necessary outliers and impurities. All blank columns were firstly taken out. Next, since the “symptoms” column contained data points in the form of key-value pairs, they were separated into rows and columns,respectively. Furthermore, all the Yes/No/Blank data points were converted into binary data,with Blank and No being considered as 0, and Yes being considered as 1.The data points in the gender column were converted to binary data,with Males getting initialised as 0 and Females getting initialised as 1. Then, only rows having all instances filled were taken into consideration,the rest were removed. Any row or column value having its instance as null was removed and disregarded. Similarly, all columns having more than 95% data as null values were removed. A certain number was assigned to the string values in the “chronic condition” column which were then aggregated so as to compile some conditions and form a model-understandable arithmetic value.

Training and testing of data

The data was divided into sets of 70% or 80% training and rest for testing,to be comprehended by the learning models. The data was divided using the train\_test\_split command available in Python through the use of the Scikit-Learn package. Hence, the models were trained with either 13,007 training samples or 14,865 training samples.

Pathogen Test Result ~

| Symptoms Indicated | Values used |
| --- | --- |
| Fever | 0,1 |
| Chills | 0,1 |
| Conjunctivitis | 0,1 |
| Cough | 0,1 |
| Diarrhoea | 0,1 |
| Headache | 0,1 |
| Loss Of Appetite | 0,1 |
| Malaise | 0,1 |
| Myalgia | 0,1 |
| Nausea | 0,1 |
| Running Nose | 0,1 |
| Short Breath | 0,1 |
| Throat | 0,1 |
| Vomiting | 0,1 |
| Wheezing | 0,1 |
| Other Symptoms | 0,1 |

| Chronic Conditions Combined | Values Used |
| --- | --- |
| None | 0 |
| HPT | 1 |
| RSP | 2 |
| Chronic Lung Disease | 3 |
| END | 4 |
| CDV | 5 |
| ASM | 6 |
| HEM | 7 |
| NRL | 8 |
| OBS | 9 |
| IMS | 10 |
| CAN | 11 |
| URO | 12 |
| FLU | 13 |
| Immuno | 14 |
| Diabetes | 15 |
| Obesity | 16 |
| END;CAN | 17 |
| HEM;CAN | 18 |
| HPT;END | 19 |
| CDV;HPT | 20 |
| END;NRL | 21 |
| CDV;NRL | 22 |
| URO;END | 23 |
| CDV;URO | 24 |
| RSP;ASM | 25 |
| OBS;OTH-Chronic Lung Disease;ASM | 26 |
| CDV;END | 27 |
| CDV;END;NRL | 28 |
| OTH;Chronic Lung Disease;ASM | 29 |