Feature Selection based Artificial Intelligence Techniques for the Prediction of COVID like Diseases

Pijush Dutta¹, Shobhandeb Paul², Ahmed J. Obaid ³ Souvik Pal⁴, Koushik Mukhopadhyay⁵

¹Department of Electronics and Communication Engineering, Global Institute of Management and Technology, Krishnagar, India
²Department of Electronics and Communication Engineering, Guru Nanak institute of Technology, Sodepur, India.
³Faculty of Computer Science and Mathematics, University of Kufa, Iraq.
⁴Department of Computer Science and Engineering, Global Institute of Management and Technology, Krishnagar, India
⁵Department of Computer Science and Engineering, Durgapur Institute of Advanced Technology and Management, Durgapur, India

pijushdutta099@gmail.com, devshobhan4@gmail.com, ahmedj.aljanaby@uokufa.edu.iq, Souvikpal22@gmail.com, koushikkuma@gmail.com

Abstract:

Identification of disease from therapeutic statistical evidences area single confronted task which can make a point of importance in the field of medical science. But according to the literature survey, it has been seen that still there are some chances that this challenging task can be fulfilled. In this research a feature ranking algorithm Random Forest is used for ranked the features of the attributes & later on four machine learning algorithm has been used i.e. Random forest, decision Tree, support Vector Machine & XG Boost classification algorithm to classify similar disease datasets like Jaundice, Malaria, Covid, Common cold, Typhoid, Dengue & Pneumonia. Comparison between the classifier is done on the basis of with ranking with feature selection & ranking without feature selection with the help of parameters of confusion matrix, Matthews’s correlation coefficient (MCC), area under the curve (AUC), Receiver Operating Characteristics Curve (ROC) & computational time. The results of the simulations shows the effectiveness of Covid like disease prediction is done by the feature selection ranking & classification algorithm.

Keywords: Machine Learning, Medical data, Feature Selection, Disease prediction.

1. Introduction

In the peak of state-of-the-art research field for medical symptomatic era the relevance of artificial intelligence is quite prevalent (Kononenko, 2001). Artificial Intelligence field depicted outcomes are often received by the specialists for pinning point over a patient's syndrome. A physician regularly
gathers his insight dependent on quiet manifestations and the affirmed analysis. Consequently symptomatic precision is exceptionally subject to a doctor's insight. Recently it is accessible to annex and stockpile to digitally store huge amount of information, categorization of computerized medical deployed cornerstones have become a feasible speculation to reinforce the physicians in their work with hasty and full prove diagnosis for patients (Mohapatra et al., 2015). Through this AI technique new test can be predict with higher degree of accuracy depending upon the available records. Such diversification is really fruitful in medical informatics (Gunčar et al., 2018). Manifolded statistical provisions to know how methods are applied in data classification of a large amount of data sets in medical field depending upon the applicability of data(Zhou et al., 2009), (Nikovski, 2000), (Paliwal & Kumar, 2009). To be conversant in the factors to be worked with the dataset is sometimes not feasible to data analyzer. AI based approaches are less reliant on such field of proficiency.

Prediction of Sickness with the help of medical informatics has recently gained symbolic contemplation in the data science research field. The medical data which are collecting from either electronic health records or from administrative are stored into the computer based environment. These digitized evidences are being turned into account in a far reaching dimensions in the account of medicative ground works by its nature for spelling out in the pharmaceutics unit fulfillment(Metsemakers et al., 2017), calibrating the consummation function for the network which can be taken care of(Sumner & Lim, 2021), delving into the structures and expenditure of care(Mundal et al., 2021), flourishing contamination hazard prophecy model (Pylypchuk et al., 2018), (Wolff et al., 2019), deep-rooted disease surveillance (Allison et al., 2017) etc. Alam et al., 2019 developed a classification model with feature ranking concept for 10 different diseases. From the experimental result it is seen that classified algorithm is generalized for identify the actual disease with similar disease. Khanmohammadi & Rezaeiahari, 2014 Proposed an analytical hierarchy process based supervised machine learning algorithm on ten different medical datasets. Experiments reported that among six algorithms SVM is the most fascinating innovations for promoting the clinical decision support system.

The significant additions of this paper are as follows.

- At the preliminary stage, disease datasets are taking as an input of the process system. Here seven diseases Jaundice, Malaria, Covid, Common cold, Typhoid, Dengue & Pneumonia are taken for the investigation criteria; we can come up with a matter of course for the technique accustomed for therapeutic data distribution that bring to bears a feature regarded by a felicitous chalk-talk of a suitable classifier algorithm.
- Random forest-based feature rank selection is performed to identify the higher weightage feature from actual data and finally Random Forest, Decision tree, support Vector machine & XG Boost algorithm as the final classifier for our prognostications.
- In final stage we made a comparative study of the above-mentioned classifier for predicting the 7 benchmark disease with feature ranking selection & without feature ranking selection.
- The best classifier identified on the basis of three parameters MCC, AUC & computation time which are described in the last section.

2. Dataset
The pursuance’s of the proposed feature ranked based machine learning algorithm is shaking down on seven real yardstick categorized problem withholds from UCI Machine Learning Repository & other 10 review of the paper we prepare for the datasets and their characteristics shown in Table 1. The training, testing, and validation datasets are automatically distributed during the classifier algorithm. The Seven touchstone datasets on which evaluation results are consummated are: Jaundice, Malaria, Covid, Common cold, Typhoid, Dengue & Pneumonia, which are ready willing and able from the UCI Machine Learning Repository. Table 1 summarizes the number of features (28), instances (1588) and here we don’t use any classes for each dataset used in this study.

| Disease      | Number of Instances | Number of features |
|--------------|---------------------|--------------------|

Table 1. Description of Datasets
Jaundice: It contains 182 instances characterized by nine features (Cordero et al., 2020): (1) nausea and vomiting, (2) abdominal pain, (3) fever, (4) weakness, (5) loss of appetite, (6) headache, (7) confusion, (8) swelling of the legs and abdomen (9) yellow discoloration of the skin.

Malaria: It contains 203 number of datasets with ten features (Poostchi et al., 2018): (1) Headache, (2) Diarrhea, (3) Being very tired (fatigue), (4) Body aches, (5) Yellow skin (jaundice), (6) Kidney failure (7) Seizure (8) Confusion (9) Bloody stools (10)Convulsions

Covid 19: It contains 352 number of datasets with thirteen features (Randhawa et al., 2020), (Asnaoui et al., 2020): (1) fever, (2) dry cough (3) tiredness (4) throbbing and agonisation (5) inflammation in throat (6) diarrhoea (7) conjunctivitis (8) headache (9) deficiency of taste or smell (10) a rash on skin, or blemish of fingers or toes (11) complication in taking breath or shortness of breath (12) chest cramp or pressure (13) loss of speech or movement

Common fever: It contains 179 number of datasets with eight features (Lopez et al., 2017): (1) perspiring, (2) Chills and shuddering with vibration (3) Headache (4) Muscle aches (5) Loss of appetite (6) Irritability (7) Dehydration (8) General weakness.

Typhoid: It contains 224 number of datasets with nine features (Andrianto et al., 2019): (1) high fever (2) weakness (3) stomach pain (4) headache (5) poor appetite (6) rash (7) fatigue (8) confusion (9) constipation, Diarrhea.

Dengue: It contains 197 number of datasets with nine features (Hair et al., 2019): (1) high fever (2) Severe headaches (3) Pain behind the eyes (4) Severe joint and muscle pain (5) Fatigue (6) Nausea (7) Vomiting (8) Skin rash, (9) Mild bleeding.

Pneumonia: It contains 251 number of datasets with twenty five features (C. Liu et al., 2020), (Luján-García et al., 2020): (1) cough (2) coughed up from lungs (3) fever (4) fast breathing and shortness of breath (5) shaking chills (6) chest pain (7) fast heartbeat (8) fatigue and weakness (8) nausea and vomiting (9) diarrhoea (10) sweating (11) headache (12) muscle pain (13) confusion (14) dusky or purplish skin colour.

3. Proposed Methodology
We scrutinised a dataset of evidences of healthiness of 1588 victims having seven different diseases taken UCI machine learning. Of 1588 datasets, 1270 datasets are used to test the model & 318 datasets for validation. Each of the rows of the datasets contains 28 different attributes/ features observed in diseases represent by binary symbol 1 for yes & 0 for no. We compared Decision Tree (DT), Random forest(RF), Support Vector machine(SVM), & XG Boost classifiers to predict diagnosis of the patient records.

| Disease    | Instances | Features |
|------------|-----------|----------|
| Jaundice   | 182       | 9        |
| Malaria    | 203       | 10       |
| Covid 19   | 352       | 13       |
| Common fever | 179   | 8        |
| Typhoid    | 224       | 9        |
| Dengue     | 197       | 9        |
| Pneumonia  | 251       | 25       |
Figure 1. General methodology of the Covid like disease prediction

Figure 1 shows a diagram of model workflow. Initially features of each disease are identifying where same the features may or may not be match between two or diseases. As the feature selection is one of special criteria in machine learning which utilized to improve the accuracy, improvement of the data classification & remove the data redundancy during the classification (Ma et al., 2017). In this research we use Random forest as a feature selection tool to identify the higher priority of the features selections representing as a rank 1, rank 2, rank 3 & so on. A top ranked feature based subset of data generated. Finally, Four different machine learning algorithms (DT, RF, SVM & XG Boost ) are applied on the selected features to validate the final model.

3.1 Feature Selection:
Feature selection is one of the dimensionality reduction technique in the field of data mining & machine learning for the application of image classification, extended or shortened the features space & object based classification and so on (Suebsing & Hiransakolwong, 2009). This Feature selection model helps to removing the redundant, noisy & irrelevant features from the datasets & makes the model simpler to classify the data better. Feature Selection can be done by the two different approaches: filter and wrapper (Chou et al., 2008). In filter method performance don’t rely upon the any learning algorithm, it automatically eliminate the features which don’t provide sufficient score. Beside of these the main advantage of the filter method is it has low computational cost (Chou et al., 2008).In The wrapper feature selection model the performance depends upon the learning algorithm, offers high classification performance but when it is dealing with large datasets it has high computational cost(Lee et al., 2000). In this paper, we adapt RF as a rank selection algorithm to identify the higher weightage of features and prepare the subsets of data which can bring to a successful classification for the datasets for Covid like disease.

3.2 Random forest as an ensemble learning method
The main objective of the feature selections are a) to improve the model accuracy & b) generalized the data to gain more insights. In RF feature selection method two important variable indexes measured: Gini importance index & Permutation importance index. In RF samples which are not used to train consider as a leaves out of bag (OOB) samples are handled to allot the incisiveness. Ranking of the features can be measured by making a comparison between intact OOB and sample OOB with the rundown countenances.

Let we consider a bootstrap samples $b = 1, \ldots, B$, a single tree $t$ on this $OOB_b$ sample denoted by $err_{OOB_t}$. Haphazardly permute the values of $X_j$ in $OOB_t$ to get a flustered sample implied by $OOB^j_t$. Figure out $err_{OOB_t}$ and the difference of the error between the intact OOB samples and the OOB samples with the permuted feature $X_j$. Then average permutation importance measure (APIM) for the permuted feature $X_j$ is expressed in Eqn 1.

Average permutation index measure (APIM):

$$APIM(X_j) = \frac{1}{D} \sum_{k=1}^{D} PIM_k(X^j) \quad (1)$$

Where Permutation index measure, $PIM_k(X^j) = \frac{1}{n_{tree}} \sum (err_{OOB^j_t} - err_{OOB^j_t}) \quad (2)$

3.3 Forecast predictive analysis evaluation:
Effectiveness of the proposed model determine by five features which are expressed in Eqn 3 to Eqn 7 respectively.

$$MCC = \frac{TP.TN- FP.FN}{\sqrt{(TP+FP).(TP+FN).(TN+FP).(TN+FN)}} \quad (3)$$

$$Accuracy = \frac{TP+TN}{2TP} \quad (4)$$

$$F1 \text{ score} = \frac{2TP}{2TP+FN+FP} \quad (5)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (6)$$

$$Specificity= \frac{TN}{TN+FP} \quad (7)$$

The MCC metric first introduced by B.W. Matthews applied in different domain such as are prediction of protein secondary structure (Huang et al., 2010), performance in biomedical groundwork (Yang et al., 2013), (G. Liu et al., 2013), for development and validation of predictive models for personalized medicine (Consortium, 2010). MCC is mainly a binary classification, ranges between -1 to +1. For complete agreement the value of correlation coefficient represent by 1; for disagreement it shows binary -1 & for uncorrelated prediction its value is 0. In the process of Matthews’s correlation coefficient (MCC) computation, we can use the confusion matrix to get the relative value of recommended result classification [31–35].

4. Result Analysis on Multiple Disease Detection:
The proposed dataset is given below:

| Disease | Itching | Vomiting | Fatigue | weight loss | High fever | yellowish skin | dark urine | abdominal pain | chills | throat irritation | redness of eyes |
|---------|---------|----------|---------|-------------|------------|----------------|------------|----------------|--------|-----------------|----------------|
| 0        | 1       | 1        | 1       | 1           | 1          | 1              | 1          | 0              | 0      | 0               | 0              |
| 1        | 1       | 1        | 1       | 1           | 1          | 1              | 1          | 0              | 0      | 0               | 0              |
| 2        | 0       | 1        | 1       | 1           | 1          | 1              | 1          | 0              | 0      | 0               | 0              |
| 3        | 1       | 0        | 1       | 1           | 1          | 1              | 1          | 0              | 0      | 0               | 0              |
| 4        | 1       | 1        | 0       | 1           | 1          | 1              | 1          | 0              | 0      | 0               | 0              |
After analyzing the datasets, it was found that the dataset is a type of classification problem, thus main classification algorithms were applied to train the model and obtain the accuracy. To play out all these three algorithms we utilized the following features processor and platform: Intel i3, sixth era processor, OS: Ubuntu 20.04 and RAM +8Gb, python 3.7.6, and Jupyter journal 6.03. We assess the presentation of the models utilizing distinctive famous metrics from the literature. Subtleties of these measurements and important notions have been discussed about in the review of machine learning. All through the tests, the 10-crease cross-approval result is the normal of the outcomes.

**Figure 2.** Disease & their attributes

We generated prediction results through Random forest, decision Tree, Support Vector Machine & XG Boost classifier affixed to all the emphasized part and to all the data instances.

**Figure 3.** Barplot of the datasets for the proposed Diseases

| Methods                | MCC   | Accuracy | F1score | Sensitivity | Specificity |
|------------------------|-------|----------|---------|-------------|-------------|
| Random Forest          | +0.94 | 0.94     | 0.97    | 0.98        | 0.97        |
| Decision Tree          | +0.933| 0.933    | 0.90    | 0.93        | 0.90        |
| Support vector machine | +1    | 0.966    | 0.97    | 0.97        | 0.97        |
| XG Boost               | +0.863| 0.863    | 0.87    | 0.90        | 0.87        |
Figure 4. Results of RF for with & without ranking

Figure 5. Results of DT for with & without ranking
Figure 6. Results of SVM for with & without ranking
Prediction score depending upon the confusion matrix parameters & Matthews’s correlation coefficient (MCC) varied with maximum value of MCC, Accuracy, Specificity & least values of sensitivity. From the Table 2 it is clearly view that among this four criteria (MCC, accuracy, Sensitivity & Specificity) no one of the mentioned algorithm satisfied all these features. XG Boost carried through the under most prediction score among the techniques we give our endeavour most, obtaining a result with MCC (+0.863), sensitivity (0.90) & Specificity (0.87). Support vector machine (SVM) brought to pass perfect MCC of +1.00, indicates such classifier constrained adequacy in clinical settings. So Random Forest (RF) & Decision Tree (DT) achieved the better prediction.

Figure 7. Results of XG Boost for with & without ranking

Figure 8. ROC for Analysis for Random Forest, XGB, Decision Tree & SVM

Table 3. Comparative study based on Execution time & AUC

| Methods       | Training Time | Prediction Time | AUC   |
|---------------|---------------|----------------|-------|
| Random Forest | 0.174s        | 0.012s         | 0.9457|
| Decision Tree | 0.013s        | 0.001s         | 0.8936|
From the Table 3 it is seen that decision tree have less execution time by means of training & prediction datasets and better AUC (area under curve) indicates least misclassified datasets.

### Table 4. Comparative study based on overall score

| Algorithm     | Accuracy | MCC  | F1 Score | Sensitivity | Specificity | Training Time | Prediction Time | AUC     | Total |
|---------------|----------|------|----------|-------------|-------------|---------------|----------------|---------|-------|
| RF            | 0.33     | 0.33 | 0.66     | 0.66        | 0.66        | 0             | 0              | 0.66    | 5.28  |
| DT            | 0.33     | 0.33 | 0.33     | 0.33        | 0.33        | 0.33          | 0.66           | 0.33    | 2.97  |
| SVM           | 0.66     | 0.66 | 0.66     | 0.66        | 0.33        | 0.66          | 0.33           | 0       | 4.29  |
| XG Boost      | 0        | 0    | 0        | 0           | 0           | 0             | 0              | 0.33    | 0.33  |

5. Conclusion:
Data classification in health domain is one of the tangled & accosted grind stones. Due to the roundabout nature, various ways and means have been contemplated in the treatise survey. In this theory we have generalized the methodology for the classification task of a Covid 19 like disease. Overall process done into two steps, in first phase we have prospective a feature ranking based methodology using Random forest that employ the countenance excerption strategy & in second phase model was classification train based on highly ranked features.

In this exploration we have lead the broad investigations on 7 benchmark datasets. Classification of model performance has been done depending upon without feature ranking & with feature ranking model. To wrap up, we have done not only refined a high ultra precise model, but also generalized a method that should perform well for the all the 7 benchmark problem which have a collateral temperaments in the data motives.

It is worth making it known to top ranked feature have been erected to be more important from a pharmaceutical two cents worth the thoughts, that why we used feature ranking to lead the medical datasets better classified & predicted.

Although in the data classification section we have used four classifier algorithms such as: Random forest, decision Tree, Support vector machine & XGBoost. From the Table 4 it is clearly seen that Random forest classifier outperformed than the other classifier for a rank based features selection for medical data classification. On the other hand it is also concludes that the present general methodology expected to perform well for kind of medical data sets which is also open for future scope.

References:
[1] Alam, M. Z., Rahman, M. S., & Rahman, M. S. (2019). A Random Forest based predictor for medical data classification using feature ranking. *Informatics in Medicine Unlocked, 15*, 100180.
[2] Allison, F., Lix Lisa, M., & Kim, R. (2017). Estimating multimorbidity prevalence with the Canadian chronic disease surveillance system. *Health Promotion and Chronic Disease Prevention in Canada: Research, Policy and Practice, 37*(7), 215.
[3] Andrianto, B., Suprapto, Y. K., Pratomo, I., & Irawati, I. (2019). Clinical decision support system for typhoid fever disease using classification techniques. *2019 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, 248–252.
[4] Asnaoui, K. E., Chawki, Y., & Idri, A. (2020). Automated methods for detection and classification pneumonia based on x-ray images using deep learning. ArXiv Preprint ArXiv:2003.14363.

[5] Aydadenta, H., & Adiwijaya, A. (2018). A Clustering Approach for Feature Selection in Microarray Data Classification Using Random Forest. Journal of Information Processing Systems, 14(5), 1167–1175. https://doi.org/10.3745/JIPS.04.0087

[6] Chou, T.-S., Yen, K. K., & Luo, J. (2008). Network intrusion detection design using feature selection of soft computing paradigms. International Journal of Computational Intelligence, 4(3), 196–208.

[7] Consortium, M. (2010). The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. Nature Biotechnology, 28(8), 827.

[8] Cordero, C., Schieve, L. A., Croen, L. A., Engel, S. M., Siega-Riz, A. M., Herring, A. H., Vladutiu, C. J., Seashore, C. J., & Daniels, J. L. (2020). Neonatal jaundice in association with autism spectrum disorder and developmental disorder. Journal of Perinatology, 40(2), 219–225.

[9] Gunčar, G., Kukar, M., Notar, M., Brvar, M., Černelč, P., Notar, M., & Notar, M. (2018). An application of machine learning to haematological diagnosis. Scientific Reports, 8(1), 411. https://doi.org/10.1038/s41598-017-18564-8

[10] Hair, G. M., Nobre, F. F., & Brasil, P. (2019). Characterization of clinical patterns of dengue patients using an unsupervised machine learning approach. BMC Infectious Diseases, 19(1), 1–11.

[11] Huang, N., Lee, I., Marcotte, E. M., & Hurles, M. E. (2010). Characterising and predicting haplinsufficiency in the human genome. PLoS Genet, 6(10), e1001154.

[12] Khammohammadi, S., & Rezaiahari, M. (2014). AHP based Classification Algorithm Selection for Clinical Decision Support System Development. Procedia Computer Science, 36, 328–334. https://doi.org/10.1016/j.procs.2014.09.101

[13] Kononenko, I. (2001). Machine learning for medical diagnosis: History, state of the art and perspective. Artificial Intelligence in Medicine, 23(1), 89–109. https://doi.org/10.1016/S0933-3657(01)00077-X

[14] Lee, W., Stolfo, S. J., & Mok, K. W. (2000). Adaptive Intrusion Detection: A Data Mining Approach. Artificial Intelligence Review, 14, 533–567.

[15] Liu, C., Wang, X., Liu, C., Sun, Q., & Peng, W. (2020). Differentiating novel coronavirus pneumonia from general pneumonia based on machine learning. Biomedical Engineering Online, 19(1), 1–14.

[16] Liu, G., Mercer, T. R., Shearwood, A.-M. J., Siira, S. J., Hibbs, M. E., Mattick, J. S., Rackham, O., & Filipovska, A. (2013). Mapping of mitochondrial RNA-protein interactions by digital RNase footprinting. Cell Reports, 5(3), 839–848.

[17] Lopez, D. M., de Mello, F. L., Giordano Dias, C. M., Almeida, P., Araújo, M., Magalhães, M. A., Gazeta, G. S., & Brasil, R. P. (2017). Evaluating the surveillance system for spotted fever in Brazil using machine-learning techniques. Frontiers in Public Health, 5, 323.

[18] Luján-García, J. E., Yañez-Márquez, C., Villuendas-Rey, Y., & Camacho-Nieto, O. (2020). A transfer learning method for pneumonia classification and visualization. Applied Sciences, 10(8), 2908.

[19] Ma, L., Fu, T., Blaschke, T., Li, M., Tiede, D., Zhou, Z., Ma, X., & Chen, D. (2017). Evaluation of feature selection methods for object-based land cover mapping of unmanned aerial vehicle imagery using random forest and support vector machine classifiers. ISPRS International Journal of Geo-Information, 6(2), 51.

[20] Metsemakers, W.-J., Smets, B., Nijs, S., & Hoekstra, H. (2017). Infection after fracture fixation of the tibia: Analysis of healthcare utilization and related costs. Injury, 48(6), 1204–1210.

[21] Mohapatra, P., Chakravarty, S., & Dash, P. K. (2015). An improved cuckoo search based extreme learning machine for medical data classification. Swarm and Evolutionary Computation, 24, 25–49. https://doi.org/10.1016/j.swevo.2015.05.003

[22] Mundal, I., Lara-Cabrera, M. L., Betancort, M., & De las Cuevas, C. (2021). Exploring patterns in psychiatric outpatients’ preferences for involvement in decision-making: A latent class analysis approach. BMC Psychiatry, 21(1), 1–12.
[23] Nikovski, D. (2000). Constructing Bayesian networks for medical diagnosis from incomplete and partially correct statistics. *IEEE Transactions on Knowledge and Data Engineering, 12*(4), 509–516. https://doi.org/10.1109/69.868904

[24] Paliwal, M., & Kumar, U. A. (2009). Neural networks and statistical techniques: A review of applications. *Expert Systems with Applications, 36*(1), 2–17. https://doi.org/10.1016/j.eswa.2007.10.005

[25] Poostchi, M., Silamut, K., Maude, R. J., Jaeger, S., & Thoma, G. (2018). Image analysis and machine learning for detecting malaria. *Translational Research, 194*, 36–55.

[26] Pylypchuk, R., Wells, S., Kerr, A., Poppe, K., Riddell, T., Harwood, M., Exeter, D., Mehta, S., Grey, C., & Wu, B. P. (2018). Cardiovascular disease risk prediction equations in 400 000 primary care patients in New Zealand: A derivation and validation study. *The Lancet, 391*(10133), 1897–1907.

[27] Randhawa, G. S., Soltysiak, M. P., El Roz, H., de Souza, C. P., Hill, K. A., & Kari, L. (2020). Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study. *Plos One, 15*(4), e0232391.

[28] Suebsing, A., & Hiransakolwong, N. (2009). Euclidean-based feature selection for network intrusion detection. *Published on: 2009 International Conference on Machine Learning and Computing*.

[29] Sumner, J., & Lim, Y. W. (2021). Back to the Hospital for Chronic Care: A Hybrid Generalist Model. *Journal of General Internal Medicine, 36*(3), 782–785.

[30] Wolff, R. F., Moons, K. G., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., & Mallett, S. (2019). PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Annals of Internal Medicine, 170*(1), 51–58.

[31] Yang, J., Roy, A., & Zhang, Y. (2013). Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics, 29*(20), 2588–2595.

[32] Zhou, X.-H., McClish, D. K., & Obuchowski, N. A. (2009). *Statistical Methods in Diagnostic Medicine*. John Wiley & Sons.

[33] Yagyanath Rimal et al 2021 J. Phys.: Conf. Ser. 1804 012054

[34] Azmi Shawkat Abdulbaqi, Ahmed J. Obaid & Alyaa Hashem Mohammed (2021) ECG signals recruitment to implement a new technique for medical image encryption, Journal of Discrete Mathematical Sciences and Cryptography, DOI: 10.1080/09720529.2021.1884378.

[35] Kautish, S., Peng, S.-L., & Obaid, A. J. (2021). Computational Intelligence Techniques for Combating COVID-19. Springer International Publishing.