Prediction of Covid-19 Using Ensemble based Machine Learning Approach

Asma Akhtar, Nasir Abbas
[1]Department of Computer Science, Virtual University of Pakistan, Lahore, Pakistan
[2]WHO PEI, Lahore
asmaakhtarjanjua@gmail.com

Abstract- The viral infection of COVID-19 has been wreaking havoc on earth since December 2019. It has affected every aspect of human life. This dangerous disease took millions of lives and made an even larger number of people sick with dreadful symptoms. To stop the spread of this fatal infection reliable COVID-19 screening is crucial at an early stage. Some studies have been conducted which depict the importance of using a routine blood test for the initial screening of COVID-19-positive patients. In this study, we employed several machine learning techniques to predict COVID-19 using a complete blood count. Variants of the different classifiers are generated by tuning their parameters to get better accuracy. The algorithms that are tuned and optimized include “Support Vector Machine, K Nearest Neighbors, Multi-Layer Perceptron, Random Forest, Decision Tree, and Naive Bayes". These variants are used to detect COVID-19. Variants that depict higher accuracy are chosen from each classifier family. Ensemble learning is used to integrate the highest-performing variants. The decision Tree variant is chosen for the ensemble as it performed better than the other classifier variants. “Accuracy, Recall, Precision, and F-Measure" are the performance measures used.

Keywords: - Machine learning, COVID-19 Prediction, Ensemble learning, Classifier Variants, CBC Test.

1 INTRODUCTION:

Coronavirus is dangerous for animals and humans as well. There are several kinds of coronaviruses that can be transmitted from animals to human beings [1, 2]. COVID-19 was a species found to be unknown to our civilization as it was not discovered before 2019 hence making this situation extremely alarming. In January 2020 World Health Organization (WHO) and the Chinese authority declared it as a global pandemic [3]. So, to stop this infection from spreading several attempts were made for early diagnosis and treatment. IgM/IgG antibodies tests have surfaced for the detection of COVID-19 but they come bearing disadvantages with sensitivity and specificity being very low for initial diagnosis. Due to the high dosage of radiation, tests like CT scans and chest X-rays are not much of a help even though they showed good results. Another test developed for the detection of this disease is known as “Reverse Transcription Polymerase Chain Reaction (RT-PCR)” which is also considered the gold standard [4] but like the other tests it has some drawbacks. It is slow, expensive, and requires special laboratory apparatus so many countries cannot afford to have such labs and equipment. Machine learning is a part of artificial intelligence that allows the software to learn to perform a task or solve an issue on its own [5, 6]. Recently some experiments have shown that the blood attributes of COVID-19 infected individual get affected. So, working with these alterations of blood parameters can prove to be effective for the initial screening of the disease. The proposed framework will be faster, simple, and readily available even in low-income countries hence slowing down the spread of this fatal disease.

2 RELATED WORK

Several machine learning techniques are being explored by researchers for the improvement of study and experimentation carried out across various fields like engineering, health sciences, etc. Some of the research disciplines which employed machine learning algorithms include software defect prediction [13, 14, 15, 16, 17, 18, 19, 20, 21], rainfall prediction [22, 23, 48, 49, 50], sentiment analysis [24, 25, 26, 27, 28], disease prediction and medical data mining [29, 30, 31, 32, 33, 34, 35, 36, 37, 38], energy demand forecasting [39]. ML methods have previously been used to predict COVID-19 by utilizing scientific and clinical data. This organized a few comparative projects. [7] Employed an ML approach known as Random Forest [40] to diagnose COVID-19 using regular blood tests. Several healthcare facilities in Lanzhou, China, provided a dataset containing 105 COVID-19-positive patients. The suggested methodology utilized 11 out of 49 parameters and attained an accuracy of 96.95%, a sensitivity of 95.12, and a spa

Specificity of 96.97%. Based on full blood count tests, experts in [8] developed a methodology for the first diagnosis of COVID-19 using Random Forest [41] and Support Vector Machine [42]. 294 blood test samples were obtained from Kunshan healthcare institutions. This dataset consisted of 208 COVID-19-positive cases. Three different predictions were made i.e., moderate vs viral, severe vs. viral, and severe vs. moderate. 15 key features were chosen. The support Vector Machine based model performed better with 88% sensitivity, 80% specificity, and 84% accuracy. Researchers in [9] employed extreme gradient boosting (XGBoost) [43] in their framework. University Medical Center Ljubljana located in Slovenia provided a dataset comprising 5333 blood samples along with other clinical data. This dataset contains 160 COVID-19-infected individuals. 35 key attributes were chosen out of a total of 117. The architecture has an 81.9% sensitivity, 97.9% specificity, and an Accuracy of 97%. The researchers [10] developed a machine-learning system for the prediction of COVID-19 using demographic variables and 27 blood features of people with the disease. Logistic Regression [44], Decision Tree [45], Random Forest, and Gradient Boosting Decision Tree [46] were among the algorithms employed in the architecture. A dataset of 3346 individuals was collected with 1394 positive cases but for validation, a dataset comprising 1822 individuals with 549 COVID-19 confirmed cases were acquired from another hospital. Gradient Boosting Decision Tree performed better than the other algorithms with 85.3% AUC, 75.8% sensitivity, and 80.2% specificity. For COVID-19 forecasting, scholars in [11] used five ML algorithms: Svms, Logistic Regression, DTs, Random Forest, and a deep neural network [47]. The dataset included 912 sample data, including 361COVID-19-confirmed cases gathered from 18 local hospitals in Zhejiang. Only ten of the 31 characteristics were chosen for use. Logistic Regression performed better than other predictors, with 91% efficiency, 87% sensitivity, and 95% validity.

MATERIALS AND METHOD

The complete dataset which comprises 5644 patient records was collected between March 28, 2020, and April 3, 2020. It was acquired from Albert Einstein Israelite Hospital located in Sao Paulo, Brazil [12]. This dataset was made publicly available by Kaggle. The data was anonymized as well as normalized to have a zero average and a single deviation. Basophils, red blood cells (RBC), leukocytes, immune cells, mean corpuscular hemoglobin concentration (MCHC), hemoglobin, hematocrit, white cells, mean platelet volume (MPV), red blood cell distribution width (RBCDW), mean corpuscular hemoglobin (MCH), mean corpuscular volume (MCV), monocytes, and eosinophils were chosen as characteristics for the conceptual methodology. Age and neutrophils were removed from the architecture since many patients had incomplete information. For early diagnosis of COVID-19, patients in the semi-intensive unit and critical care unit were eliminated. We divided the data into two categories: training (70%) and testing (30%). For data analysis, we used "Weka," a technology developed by Waikato University in New Zealand. The suggested framework is depicted in Figure 1.

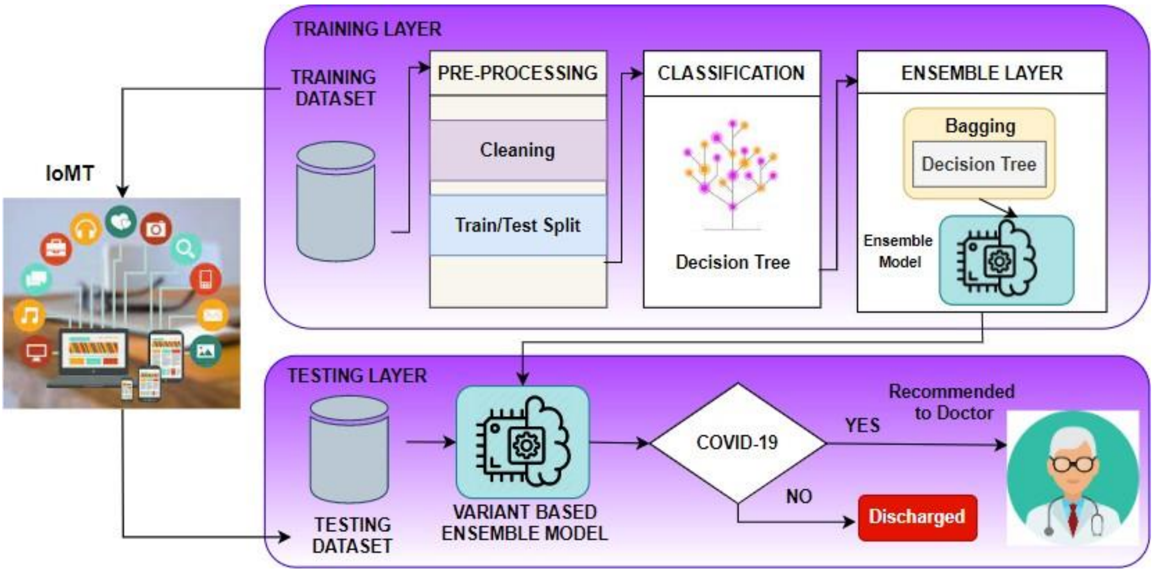


Figure 1: Proposed Classification Model

Variants of different classification algorithms are created by optimizing their parameters. The base classifier with its default parameters is considered the first variant. Three variants of “Support Vector Machine” are generated by tuning its value of complexity parameter ‘C’. The base classifier has C=1 (default) which is our first variant, the rest of the three variants have the value of C = 10, 25, 50. For “Naïve Bayes” two variants are created by tuning the parameters ‘Use Kernel Estimator’ and ‘Use Supervised Discretization’. Four variants are created for “Multi-Layer Perceptron” by tuning the parameter ‘h’ which refers to the number of neurons. This parameter ‘h’ has values 3, 5, 7, and 9 for four variants. Similarly, by optimizing the parameters ‘Max depth’ two variants of “Random Forest” are generated with max depth values 10 and 15. “K Nearest Neighbors” produced three variants by optimizing the parameter ‘no of neighbors. The values selected for this parameter are 1 (default), 3, 5, and 7. Lastly for the “Decision Tree,” four variants are created by tuning its parameter known as the ‘confidence factor’. The values of this parameter selected for the variants include; 0.25 (default), 0.20, 0.15, 0.10, and 0.05. After a careful analysis of the accuracy, three variants were chosen which will be discussed in the next section.

3 RESULTS AND DISCUSSIONS

Now we will observe and analyze the performance of our proposed framework for detecting COVID-19. The measures that were used to evaluate the machine learning classifiers employed in this research are using a confusion matrix.
“True Positive (TP): Instances which are actually positive and also classified as positive”.
“False Positive (FP): Instances which are actually negative but classified as positive”.
“False Negative (FN): Instances which are actually positive but classified as negative”.
“True Negative (TN): Instances which are actually negative and also classified as negative”.
The details of the performance measures used for evaluation are described below:

"Precision gives a ratio between instances that are predicted as positive and True Positive instances"

$$Precision = \frac{TP}{(TP + FP)}$$

"Recall is defined as the ratio of True Positive (TP) modules with respect to the total number of modules that are actually positive"

$$Recall = \frac{TP}{(TP + FN)}$$

“F-measure provides the average of Precision & Recall”

$$F - measure = \frac{Precision \times Recall \times 2}{(Precision + Recall)}$$

“Accuracy depicts how much the prediction is accurate”

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)}$$

These performance measures are made available by "Weka". Various variants of multiple classifiers were created and then these variants were used to predict COVID-19. The variants that performed better than the base classifier and other variants are shortlisted. On the basis of accuracy measure three variants, MLP-4, RF-2, and DT-5, were chosen for the proposed framework as they performed better than the base classifier (first variant) and the rest of the classifiers of their families.

Table 1: Accuracy of Selected Variants

Variants	Optimized Parameter	Training Accuracy		Testing Accuracy	
		Default Classifier	Variant	Default Classifier	Variant
MLP-4	No of Neurons = $(attributes + Classes)$/2= 7	97.3747	95.9427	85.4749	88.2682
RF-2	Max depth = 10	100	100	87.1508	89.3855
DT-5	Confidence factor= 0.05	97.136	92.6014	84.9162	89.3855

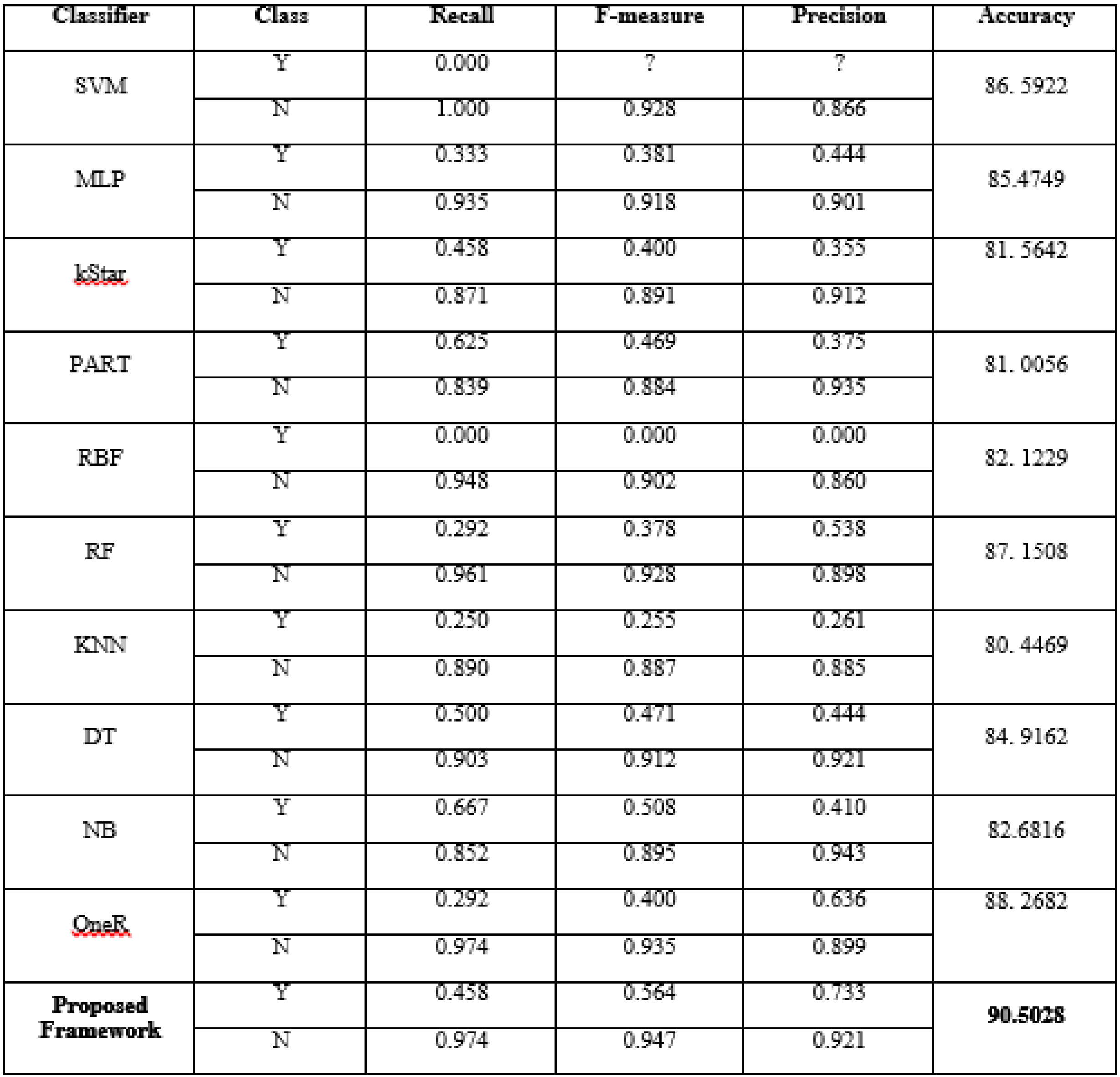
For an initial comparison with the proposed model, 10 supervised machine learning algorithms are used with their default parameters. Out of the three variants chosen, DT-5 was selected for the ensemble. This variant was used with the Bagging classifier and it performed better than all the base classifiers as well as all the other variants. Table 8 depicts Training results and Table 9 gives Testing results for each class Y and N. Since the accuracy measures are sensitive to class imbalance problems so a question mark ‘?’ is shown in this case.

Table 2: Training Results

Classifier	Class	Recall	F-measure	Precision	Accuracy
SVM	Y	0.000	?	?	86.3962
	N	1.000	0.927	0.864	
MLP	Y	0.842	0.897	0.960	97.3747
	N	0.994	0.985	0.976	
kStar	Y	1.000	1.000	1.000	100
	N	1.000	1.000	1.000	
PART	Y	0.930	0.835	0.757	94.9881
	N	0.953	0.970	0.989	
RBF	Y	0.018	0.034	0.500	86.3962
	N	0.997	0.927	0.866	
RF	Y	1.000	1.000	1.000	100
	N	1.000	1.000	1.000	
KNN	Y	1.000	1.000	1.000	100
	N	1.000	1.000	1.000	
DT	Y	0.825	0.887	0.959	97.136
	N	0.994	0.984	0.973	

NB	Y	0.667	0.535	0.447	84.2482
	N	0.870	0.905	0.943	
OneR	Y	0.298	0.405	0.630	88.0668
	N	0.972	0.934	0.898	
Proposed Framework	Y	0.684	0.804	0.975	95.4654
	N	0.997	0.974	0.953	

Table 3: Testing Results



Classifier	Class	Recall	F-measure	Precision	Accuracy
SVM	Y	0.000	?	?	86. 5922
	N	1.000	0.928	0.866	
MLP	Y	0.333	0.381	0.444	85.4749
	N	0.935	0.918	0.901	
kStar	Y	0.458	0.400	0.355	81. 5642
	N	0.871	0.891	0.912	
PART	Y	0.625	0.469	0.375	81. 0056
	N	0.839	0.884	0.935	
RBF	Y	0.000	0.000	0.000	82. 1229
	N	0.948	0.902	0.860	
RF	Y	0.292	0.378	0.538	87. 1508
	N	0.961	0.928	0.898	
KNN	Y	0.250	0.255	0.261	80. 4469
	N	0.890	0.887	0.885	
DT	Y	0.500	0.471	0.444	84. 9162
	N	0.903	0.912	0.921	
NB	Y	0.667	0.508	0.410	82.6816
	N	0.852	0.895	0.943	
OneR	Y	0.292	0.400	0.636	88. 2682
	N	0.974	0.935	0.899	
Proposed Framework	Y	0.458	0.564	0.733	**90.5028**
	N	0.974	0.947	0.921	

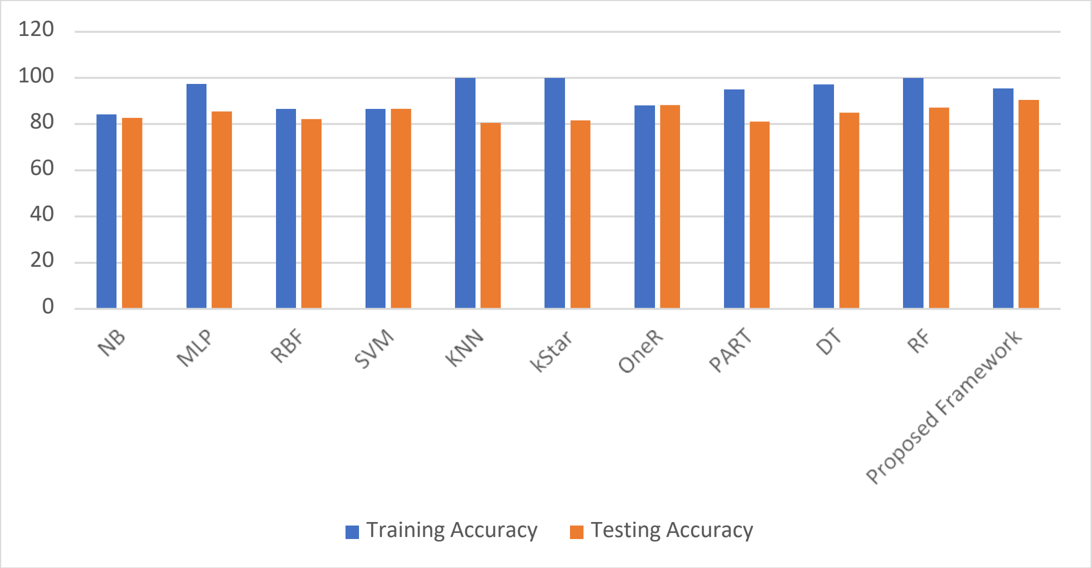


Figure 2: Accuracy results

4 CONCLUSIONS

As COVID-19 is a highly contagious viral infection so we need to diagnose it as early as possible to prevent it from spreading any further. Complete Blood count results proved effective and useful for the initial screening of COVID-19. In this paper, we created variants of six algorithms " Support Vector Machine, K Nearest Neighbors, Multi-Layer Perceptron, Random Forest, Decision Tree, and Naive Bayes" by optimizing them. Then Variant of the "Decision Tree" is selected and integrated by using the "Bagging" ensemble technique. “Accuracy, Recall, Precision, and F-Measure" results are discussed.

REFERENCES:

[1] Shaban, W. M., Rabie, A. H., Saleh, A. I., & Abo-Elsoud, M. A. (2020). A new COVID-19 Patients Detection Strategy (CPDS) based on hybrid feature selection and enhanced KNN classifier. *Knowledge-Based Systems*, *205*, 106270.

[2] Barstugan, M., Ozkaya, U., & Ozturk, S. (2020). Coronavirus (covid-19) classification using ct images by machine learning methods. *arXiv preprint arXiv:2003.09424*.

[3] Kang, H., Xia, L., Yan, F., Wan, Z., Shi, F., Yuan, H., ... & Shen, D. (2020). Diagnosis of coronavirus disease 2019 (COVID-19) with structured latent multi-view representation learning. *IEEE transactions on medical imaging*, *39*(8), 2606-2614.

[4] Zu, Z. Y., Jiang, M. D., Xu, P. P., Chen, W., Ni, Q. Q., Lu, G. M., & Zhang, L. J. (2020). Coronavirus disease 2019 (COVID-19): a perspective from China. *Radiology*, *296*(2), E15-E25.

[5] Rabie, A. H., Saleh, A., & Abo-Al-Ez, K. (2015). A new strategy of load forecasting technique for smart grids. *Int. J. Modern Trends Eng. Res.(IJMTER)*, *2*(12), 332-341.

[6] Rabie, A. H., Ali, S. H., Ali, H. A., & Saleh, A. I. (2019). A fog based load forecasting strategy for smart grids using big electrical data. *Cluster Computing*, *22*(1), 241-270.

[7] Wu, J., Zhang, P., Zhang, L., Meng, W., Li, J., Tong, C., ... & Li, S. (2020). Rapid and accurate identification of COVID-19 infection through machine learning based on clinical available blood test results. *MedRxiv*.

[8] Bao, F. S., He, Y., Liu, J., Chen, Y., Li, Q., Zhang, C. R., ... & Ouyang, L. (2020). Triaging moderate covid-19 and other viral pneumonias from routine blood tests. *arXiv preprint arXiv:2005.06546*.

[9] Kukar, M., Gunčar, G., Vovko, T., Podnar, S., Černelč, P., Brvar, M., ... & Notar, M. (2021). COVID-19 diagnosis by routine blood tests using machine learning. *Scientific reports*, *11*(1), 1-9.

[10] Yang, H. S., Hou, Y., Vasovic, L. V., Steel, P. A., Chadburn, A., Racine-Brzostek, S. E., ... & Wang, F. (2020). Routine laboratory blood tests predict SARS-CoV-2 infection using machine learning. *Clinical chemistry*, *66*(11), 1396-1404.

[11] Nan, S. N., Ya, Y., Ling, T. L., Nv, G. H., Ying, P. H., & Bin, J. (2020). A prediction model based on machine learning for diagnosing the early COVID-19 patients. *medRxiv*.

[12] Data4u, E. (2020). Diagnosis of covid-19 and its clinical spectrum.

[13] Iqbal, A., Aftab, S., Ullah, I., Bashir, M. S., & Saeed, M. A. (2019). A feature selection based ensemble classification framework for software defect prediction. *International Journal of Modern Education and Computer Science*, *11*(9), 54.

[14] Iqbal, A., & Aftab, S. (2020). Prediction of defect prone software modules using MLP based ensemble techniques. *International Journal of Information Technology and Computer Science*, *12*(3), 26-31.

[15] Matloob, F., Aftab, S., & Iqbal, A. (2019). A Framework for Software Defect Prediction Using Feature Selection and Ensemble Learning Techniques. *International Journal of Modern Education & Computer Science*, *11*(12).

[16] Iqbal, A., & Aftab, S. (2020). A Classification Framework for Software Defect Prediction Using Multi-filter Feature Selection Technique and MLP. *International Journal of Modern Education & Computer Science*, *12*(1).

[17] Ali, U., Aftab, S., Iqbal, A., Nawaz, Z., Bashir, M. S., & Saeed, M. A. (2020). Software defect prediction using variant based ensemble learning and feature selection techniques. *International Journal of Modern Education & Computer Science*, *12*(5).

[18] Matloob, F., Aftab, S., Ahmad, M., Khan, M. A., Fatima, A., Iqbal, M., ... & Elmitwally, N. S. (2021). Software defect prediction using supervised machine learning techniques: a systematic literature review. *Intelligent Automation & Soft Computing*, *29*(2), 403-421.

[19] Matloob, F., Ghazal, T. M., Taleb, N., Aftab, S., Ahmad, M., Khan, M. A., ... & Soomro, T. R. (2021). Software defect prediction using ensemble learning: A systematic literature review. *IEEE Access*

[20] Khan, M. A., Elmitwally, N. S., Abbas, S., Aftab, S., Ahmad, M., Fayaz, M., & Khan, F. (2022). Software Defect Prediction Using Artificial Neural Networks: A Systematic Literature Review. *Scientific Programming*, *2022*.

[21] Daoud, M. S., Aftab, S., Ahmad, M., Khan, M. A., Iqbal, A., Abbas, S., ... & Ihnaini, B. (2022). Machine Learning Empowered Software Defect Prediction System

[22] Aftab, S., Ahmad, M., Hameed, N., Bashir, M. S., Ali, I., & Nawaz, Z. (2018). Rainfall prediction in Lahore City using data mining techniques. *International journal of advanced computer science and applications*, *9*(4).

[23] Rahman, A. U., Abbas, S., Gollapalli, M., Ahmed, R., Aftab, S., Ahmad, M., ... & Mosavi, A. (2022). Rainfall Prediction System Using Machine Learning Fusion for Smart Cities. *Sensors*, *22*(9), 3504.

[24] Ahmad, M., Aftab, S., Muhammad, S. S., & Ahmad, S. (2017). Machine learning techniques for sentiment analysis: A review. *Int. J. Multidiscip. Sci. Eng*, *8*(3), 27.

[25] M. Ahmad, S. Aftab, and I. Ali, “Sentiment Analysis of Tweets using SVM,” Int. J. Comput. Appl., vol. 177, no. 5, pp. 25–29, 2017.

[26] M. Ahmad and S. Aftab, “Analyzing the Performance of SVM for Polarity Detection with Different Datasets,” Int. J. Mod. Educ. Comput. Sci., vol. 9, no. 10, pp. 29–36, 2017.

[27] M. Ahmad, S. Aftab, M. S. Bashir, N. Hameed, I. Ali, and Z. Nawaz, “SVM Optimization for Sentiment Analysis,” Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 4, 2018.

[28] M. Ahmad, S. Aftab, M. S. Bashir, and N. Hameed, “Sentiment Analysis using SVM: A Systematic Literature Review,” Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 2, 2018.

[29] Ahmad, M., Alfayad, M., Aftab, S., Khan, M. A., Fatima, A., Shoaib, B., ... & Elmitwally, N. S. (2021). Data and machine learning fusion architecture for cardiovascular disease prediction. CMC-COMPUTERS MATERIALS & CONTINUA, 69(2), 2717-2731.

[30] Ahmed, U., Issa, G. F., Khan, M. A., Aftab, S., Khan, M. F., Said, R. A., ... & Ahmad, M. (2022). Prediction of diabetes empowered with fused machine learning. IEEE Access, 10, 8529-8538.

[31] Aftab, S., Alanazi, S., Ahmad, M., Khan, M. A., Fatima, A., & Elmitwally, N. S. (2021). Cloud-based diabetes decision support system using machine learning fusion.

[32] Krishnamoorthi, R., Joshi, S., Almarzouki, H. Z., Shukla, P. K., Rizwan, A., Kalpana, C., & Tiwari, B. (2022). A novel diabetes healthcare disease prediction framework using machine learning techniques. Journal of Healthcare Engineering, 2022.

[33] Riyaz, L., Butt, M. A., Zaman, M., & Ayob, O. (2022). Heart Disease Prediction Using Machine Learning Techniques: A Quantitative Review. In International Conference on Innovative Computing and Communications (pp. 81-94). Springer, Singapore.

[34] Lu, H., Uddin, S., Hajati, F., Moni, M. A., & Khushi, M. (2022). A patient network-based machine learning model for disease prediction: The case of type 2 diabetes mellitus. Applied Intelligence, 52(3), 2411-2422.

[35] Kavitha, C., Mani, V., Srividhya, S. R., Khalaf, O. I., & Tavera Romero, C. A. (2022). Early-Stage Alzheimer's Disease Prediction Using Machine Learning Models. Frontiers in public health, 240.

[36] Ahsan, M. M., & Siddique, Z. (2022). Machine learning-based heart disease diagnosis: A systematic literature review. Artificial Intelligence in Medicine, 102289.

[37] Alanazi, R. (2022). Identification and prediction of chronic diseases using machine learning approach. Journal of Healthcare Engineering, 2022.

[38] Nadakinamani, R. G., Reyana, A., Kautish, S., Vibith, A. S., Gupta, Y., Abdelwahab, S. F., & Mohamed, A. W. (2022). Clinical Data Analysis for Prediction of Cardiovascular Disease Using Machine Learning Techniques. Computational Intelligence and Neuroscience, 2022.

[39] Ghazal, T. M. (2022). Energy demand forecasting using fused machine learning approaches. Intelligent Automation & Soft Computing, 31(1), 539-553.

[40] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[41] Ho, T. K. (1995, August). Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition (Vol. 1, pp. 278-282). IEEE.

[42] Schölkopf, B., Smola, A. J., & Bach, F. (2002). Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press.

[43] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

[44] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley & Sons.

[45] Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. IEEE transactions on systems, man, and cybernetics, 21(3), 660-674.

[46] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

[47] Haykin, S. (2001). Neural networks: principles and practice. Bookman, 11(900).

[48] Barrera-Animas, A. Y., Oyedele, L. O., Bilal, M., Akinosho, T. D., Delgado, J. M. D., & Akanbi, L. A. (2022). Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting. Machine Learning with Applications, 7, 100204.

[49] Hussein, E. A., Ghaziasgar, M., Thron, C., Vaccari, M., & Jafta, Y. (2022). Rainfall Prediction Using Machine Learning Models: Literature Survey. Artificial Intelligence for Data Science in Theory and Practice, 75-108.

[50] Thamilselvan, R., Prabhu, M., Selvi, K. T., Karthik, V., & Bhuvaneshwaran, T. (2022, February). An Extensive Review of Rainfall Prediction using Machine Learning and Deep Learning Techniques. In 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS) (pp. 198-205). IEEE.