Høgskolen i Østfold

IDEA PAPER

Benchmarking of machine learning methods that aid clinical tools based on data from human immune system

Sowgandh Nallanichakravartula
Department of Computer Science, Østfold University College, Halden, Norway
E-mail: sowgandn@hiof.no

Introduction

A benchmark is a standard way used to evaluate the effectiveness of an activity. Various benchmarks can be done to evaluate how an activity is performing when compared with other comparable activities [1]. In machine learning (ML), the term benchmarking refers to the assessment and comparison of ML methods with regard to their ability to learn patterns. The approach of benchmarking is important to know the ability of new methods and helps in deciding the appropriate ML method for a given problem [2]. The availability of numerous different machine-learning models makes it difficult to identify the most suitable machine-learning model for the evaluation of any given dataset. This issue can be solved by addressing the benchmarking of different ml models [3].

The recent pandemic caused by a virus has shown the world how important it is to be prepared with vaccines, drugs and diagnostics for future diseases that can compromise human immune system. Machine learning methods have been suggested to offer promising solutions in expediting vaccine design, drug discovery and immune-based diagnostics [4]. There have been numerous machine learning methods that were proposed recently that can learn the patterns associated with various immune conditions [4], [5]. However, a systematic benchmarking of all the proposed machine learning methods is needed to make use of robust methods in clinical settings and other settings that are vital for human health. There can be many aspects of machine learning methods and problem formulations that impact the performance and generalizability of machine learning methods such as (a) the assumptions of the data generating process, (b) non-linear patterns that exist in the datasets, (c) sparisty of signals, (d) distributional shifts and need for domain adaptation, (e) imbalance datasets, (f) need for careful choices of performance evaluation and optimization metrics and so on. We argue that a systematic benchmarking of ML methods in any domain should address many of the points mentioned above and for that we provide empirical evidence through simulations. We propose to benchmark ML methods that can aid clinical tools based on data from human immune system.

Research Questions

Although there is a clear need for benchmarking ML methods that can aid the expedition of development of clinical tools for immune diseases, much is not known on what can constitute suitable benchmarking datasets, which performance metrics to use, how the comparison of performance between ML methods should be carried out and so on. Therefore, in this project, we aim to address the following research questions:

(RQ1). What constitutes suitable benchmarking datasets for evaluating ML methods tailored for immune disease domain?

(RQ2). Which ML methods remain robust to a diverse set of ML challenges and problem scenarios.

Literature Review

The studies over the past two decades have shown that many immune conditions like vaccinations, immune diseases, infections, and so on leave signatures in human bodies that can be detected through laboratory procedures. By taking a blood sample and using sophisticated laboratory technologies called

"sequencing", all the immune-related molecules in the human body can be scanned and studied. Using this approach, all the past and current immune-related conditions of individuals can be detected very accurately. This allows the scientific field to speculate that studying the immune components through blood is the future of clinical diagnostics [6].

The use of sequencing to catalogue all the immune-related molecules has become a common practice to find the signatures of specific immune conditions. The development and application of ML methods have become crucial tool in finding the signatures associated with immune conditions [7], [8]. In spite of the numerous ML methods being developed for this application [5], there exists no knowledge of a robust ML method that can overcome many of the common challenges of ML as described above.

Methodology

The suitable benchmarking datasets with various machine learning challenges like linear, non-linear, sparsed, and distribution shifts datasets described in the literature review section need to be simulated. Later, Immune datasets will be generated by using simulation tools that are specifically designed to simulate human immune datasets by consulting the relevant literature [9]. In addition to baseline methods, benchmarking of ML methods that were developed for vaccine design, drug discovery, and immune-based diagnostics needed to be performed [4]. Depending upon the type of problem and the application domain, the choice of evaluation metrics can vary. Careful choice of performance metrics is important as some metrics can mislead practitioners [10]. It is also important to state the interpretability criteria while selecting any model. Also, the benchmarking process should be performed in a reproducible manner.

In contrary to a traditional systematic literature review on the topic, we here aim to chart out the essential characteristics of both benchmarking datasets and ML model evaluation using suitable simulations. We particularly consider the following aspects: (a) the assumptions of the data generating process, (b) non-linear patterns that exist in the datasets, (c) sparisty of signals, (d) distributional shifts and need for domain adaptation, (e) imbalance datasets, (f) need for careful choices of performance evaluation and optimization metrics.

The sub-sections in this section will have the following structure: in each sub-section, we describe the problem under "problem statement", simulation details under the section "simulations", machine learning methods' training and evaluation details under "ML method evaluation" and the main conclusion under the title "Key observations".

1. Assumptions of the data generating process

Problem statement: Often, the choice of a machine learning method has to depend on the suitability of the methods for the assumptions of data generating process. If the relationship between the predictor variables and the outcome variable is linear, then simple linear models will perform adequately or even outperform complex models. In other words, sophisticated modern ML methods that are less interpretable (often referred to as black-box models) are not needed [11].

Simulations: For demonstrating the problem described above, we simulated four independent predictor variables each with 100 observations that follow a normal distribution, where the mean of normal

distribution varied somewhere between 5 and 10, while the standard deviation was kept constant at 1. The outcome variable was simulated in such a way that there exists a near-linear relation between the predictor variables and outcome variable as follows: $y = 3x_1+6x_2+4x_3+x_4+\epsilon$

ML method evaluation: The generated data was split into training and test splits with 80% and 20% observations respectively in each split. Linear regression and Random forest were trained on training data and tested on testing data. Root mean square error and R squared were the two metrics used here. The data simulations and ML evaluation process was repeated three times independently and the mean and standard deviation (SD) of performance metrics were computed.

Key observations: The findings showed that linear regression outperformed Random Forests when there exists linearity between the predictor variables and the outcome variable (**Table 1, Figure 1**). This suggests that complex models are not needed when the data generating process suits the assumptions of simple linear models.

Table:1 Performance metrics on datasets that have linearity between the predictor variables and the outcome variable.

	Linear Regression	**Random Forest**
RMSE	1.62e-14	2.69
R2	1	0.82

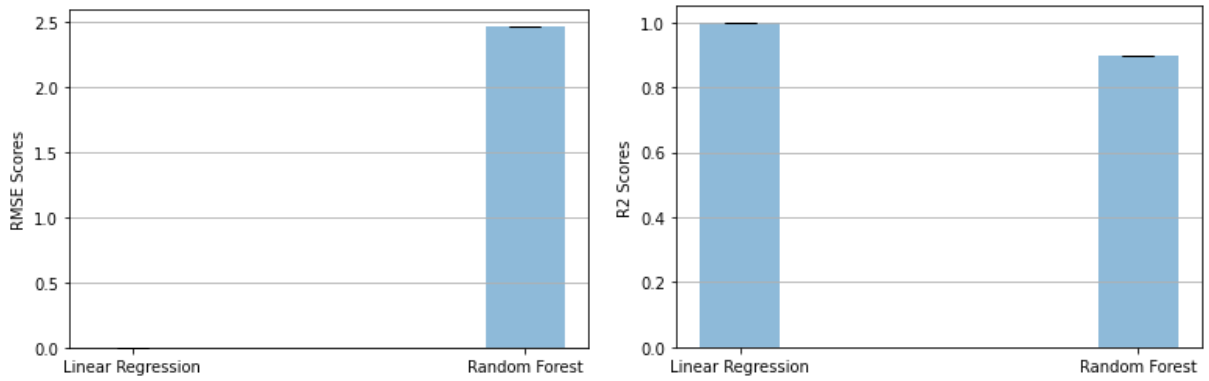

Figure: 1. Performance metrics of two different ML methods: linear regression and random forests trained on datasets that have linearity between the predictor variables and the outcome variable. RMSE is lower the better. R2 higher the better. The findings show that linear regression outperforms random forests when there exists linearity in the data-generating process.

2. Non-linear patterns in datasets

Problem statement: If the relationship between the predictor variables and the outcome variable is non-linear, then simple linear models will not perform well. Methods that account for non-linearity will be needed when non-linear patterns exist in datasets [12].

Simulations: For demonstrating the problem described above, we simulated four independent predictor variables each with 100 observations that follow a normal distribution, where the mean of normal distribution varied somewhere between 5 and 10, while the standard deviation was kept constant at 1. The outcome variable was simulated in such a way that there exists a non-linear relation between the predictor variables and outcome variable as follows: $y = 2x_1+3(x_2)^3-5(x_3)^2+x_4+\epsilon$

We also simulated another version of the datasets, where higher-degree polynomial features were added to the dataset as predictor variables. Specifically, two more predictor variables with high-degree polynomial features were added as follows: $x_5 = (x_2)^3$ and $x_6 = (x_3)^2$.

ML method evaluation: The generated data was split into training data and test data splits with 80% and 20% observations respectively. Linear regression and Random forest were trained on training data and tested on testing data. Root mean square error and R squared were the two metrics used here. The data simulations and ML evaluation process was repeated three times independently and the mean and standard deviation (SD) of performance metrics were computed.

Key observations: The findings showed that Random Forests and linear regression did not perform well when there exists non-linearity between the predictor variables and the outcome variable **(Table 2, Figure 2)**. However, when we added higher degree polynomial features to the predictor variables, we noticed a substantial improvement in the model performance **(Table 3, Figure 3)**. This suggests that sophisticated models that somehow learn the non-linear patterns are needed when non-linearity exists in datasets.

Table:2 Performance metrics on datasets that have non-linearity between the predictor variables and the outcome variable.

	Linear Regression	**Random Forest**
RMSE	103.89	76.13
R2	0.98	0.98

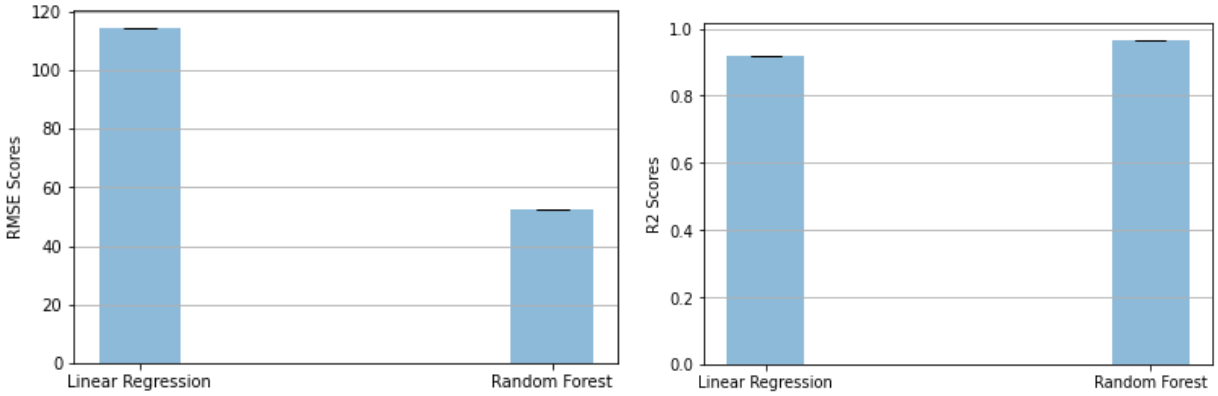

Figure: 2. Performance metrics of two different ML methods: linear regression and random forests trained on datasets that have non-linearity between the predictor variables and the outcome variable. RMSE is lower the better. R2 higher the better. The findings show that linear regression and random forests don't perform well when there exists non-linearity in the data-generating process.

Table:3 Performance metrics on datasets that have non-linearity with added polynomial features between the predictor variables and the outcome variable.

	Linear Regression	**Random Forest**
RMSE	1.716e-13	5.71
R2	1	0.99

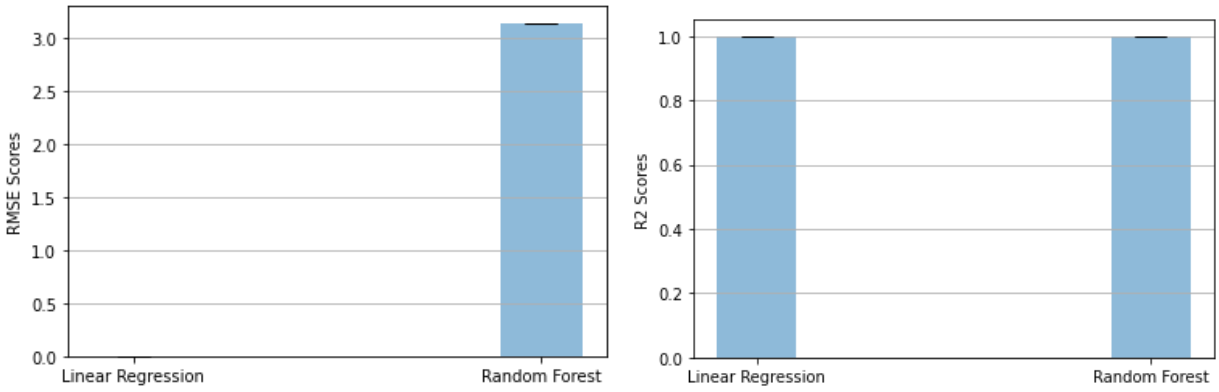

Figure: 3. Performance metrics of two different ML methods: linear regression and random forests trained on datasets that have non-linearity between the predictor variables and the outcome variable. RMSE is lower the better. R2 higher the better. Higher degree polynomial features are added to the dataset. The findings show that linear regression outperforms random forests when there exists non-linearity with added higher degree polynomial features in the data-generating process.

3. Sparse signals

Problem statement: Sparsity of signals is not an uncommon phenomenon. In fact, in the real-world many ML problem formulations can be explained by sparse signals. If only a few predictor variables among a large number of predictor variables help to explain the outcome variable, there exists a sparsity in signal. In such cases, any methods that do not handle the sparsity, for instance through some form of regularization/penalization will not generalize well for unseen data [13].

Simulations: For demonstrating the problem described above, we simulated 10,000 independent predictor variables each with 100 observations that follow a normal distribution, where the mean of normal distribution varied somewhere between 5 and 10, while the standard deviation was kept constant at 1. The outcome variable was simulated in such a way that there exists sparsity - where only a couple of variables were made predictive. The relation between the predictor variables and outcome variable was simulated as follows: $\boldsymbol{y = 2x_1+3x_2+\epsilon}$.

ML method evaluation: The generated data was split into training data and test data splits with 80% and 20% observations respectively. Linear regression and Random forest were trained on training data and tested on testing data. In addition, we also trained another version of the linear regression model but with lasso regularization (L1) with a penalization constant of 0.01 (alpha). Root mean square error and R squared were the two metrics used here. The data simulations and ML evaluation process was repeated three times independently and the mean and standard deviation (SD) of performance metrics were computed.

Key observations: The findings showed that both Random Forest and linear regression did not perform well when there exists sparsity in the signal **(Table 4, Figure 4)**. However, when we added regularization to the predictor variables, we noticed a substantial improvement in the model performance **(Table 5, Figure 5)**.

Table:4 Performance metrics on datasets that have sparse-signal between the predictor variables and the outcome variable.

	Linear Regression	**Random Forest**
RMSE	3.76	1.09
R2	-0.48	-0.05

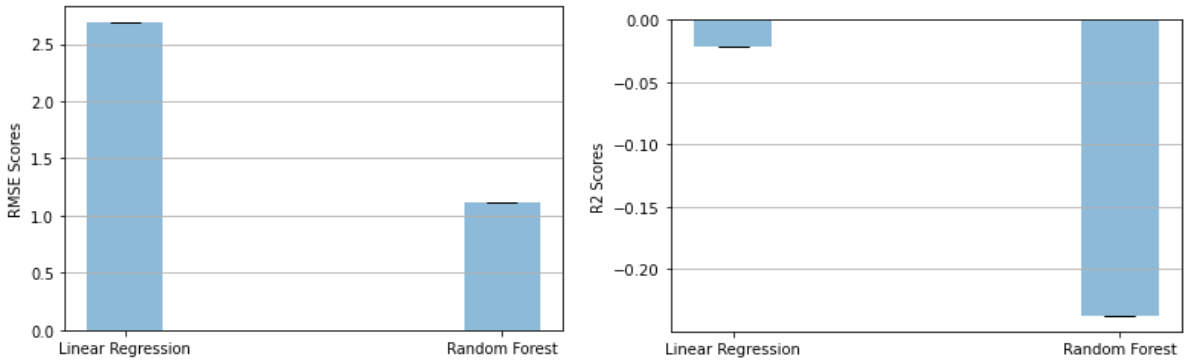

Figure: 4. Performance metrics of two different ML methods: linear regression and random forests trained on datasets that have a sparse signal between the predictor variables and the outcome variable. RMSE is lower the better. R2 higher the better. The findings show that linear regression and random forests don't perform well when there exists a sparse signal in the data-generating process.

Table:5 Performance metrics on datasets that have sparse-signal with added regularization between the predictor variables and the outcome variable.

	Linear Regression
RMSE	0.014
R2	0.99

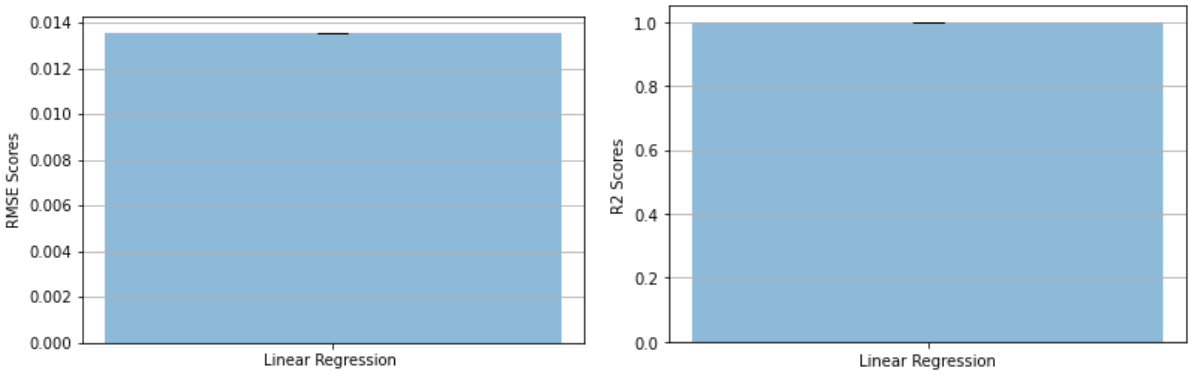

Figure: 5. Performance metrics of ML method: linear regression trained on datasets that have a sparse signal between the predictor variables and the outcome variable. RMSE is lower the better. R2 higher the better. Regularization is performed on the data. The findings show that linear regression performs well when there exists a sparse signal with added regularization in the data-generating process.

4. Distributional shifts

Problem statement: The main goal of ML methods is to generalize well on unseen test data. If the training data comes from one population distribution and the test data comes from another population distribution then the learnt model will not generalize well on new test data. This is often referred to as distributional shits in ML literature. Models that remain robust to distributional shifts are needed [14].

Simulations: For demonstrating the problem described above, we simulated four independent predictor variables each with 100 observations that follow a normal distribution, where the mean of normal distribution varied somewhere between 5 and 9, while the standard deviation was kept constant at 1. The outcome variable was simulated as a linear function between the predictor variables and the outcome variable as follows: $\boldsymbol{y = 5x_1 - 3x_2 + 1.5x_3 + 50x_4 + \epsilon}$. Another set of new test data was generated in the same way as described above, except for introducing a distributional shift in the linear relationship between the predictor variables and the outcome variable. Specifically, in the new test data the relation between $\boldsymbol{X}$ and $\boldsymbol{y}$ was simulated as follows: $\boldsymbol{y = 3x_1 - 1x_2 + 3.5x_3 + 30x_4 + \epsilon}$.

ML method evaluation: The generated data was split into training data and test data splits with 80% and 20% observations respectively. Linear regression and Random forest were trained on training data and tested on testing data. The data simulations and ML evaluation process was repeated three times independently and the mean and standard deviation (SD) of performance metrics were computed. Root mean square error and R squared were the two metrics used here. The model is tested again on the data which was generated by introducing a distributional shift in the linear relationship between the predictor variables and the outcome variable.

Key observations: The findings showed that both Random Forest and linear regression did not perform well when there exists distributional shifts between the predictor variables and the outcome variable. **(Table 6, Figure 6)**.

Table: 6 Performance metrics on datasets that have distributional shifts between the predictor variables and the outcome variable.

	Linear Regression	**Random Forest**
RMSE	145.37	147.05
R2	-16.06	-19.53

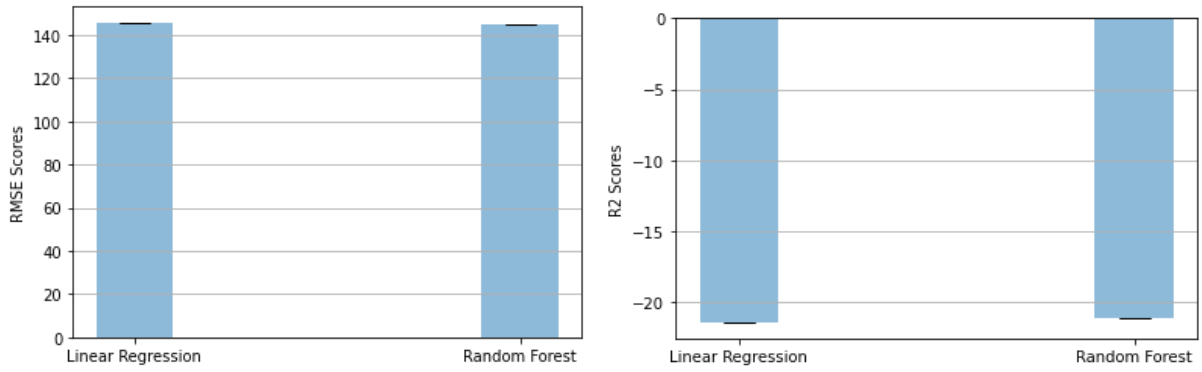

Figure: 6. Performance metrics of two different ML methods: linear regression and random forest trained on datasets that have distribution shifts between the predictor variables and the outcome variable. RMSE is lower the better. R2 higher the better. The findings show that either linear regression or random forest don't perform well when there exist distributional shifts in the data-generating process.

Performance metrics

Machine learning datasets are grouped into numerous types, such as numerical datasets, classification datasets, and so on. Using different performance metrics at different times is inaccurate since it is dependent on the type of dataset utilized, and the user draws wrong conclusions [15]. It is necessary to understand which performance measures are ideal for a certain dataset type [12]. However, it is critical to understand how a performance metric may be used to evaluate a model and which metrics pair well with models [13]. Furthermore, it is critical to understand what the performance metric indicates about the model, as well as the limits that prevent the use of sub-optimal performance metrics and this may

compensate for the limitations of employing individual metrics in isolation for a specific problem [18]. Depending on the kind of problem, whether regression, classification, or clustering, the SLR contains a list of metrics that are often employed by practitioners [19]. One of the primary results drawn from SLR is the need of evaluating the performance of ML models using multi-fitness criteria. In other words, a combination of numerous measurements targeted for a specific problem may compensate for the shortcomings of employing individual metrics alone.

Summary

It has been suggested that using machine learning techniques could speed up the development of vaccines, pharmaceuticals, and immune-based diagnostics. Recently, many machine learning techniques that can discover the patterns connected to different immune conditions have been proposed. The performance and generalizability of machine learning methods covered in the literature review section can be affected by a variety of machine learning methodologies. The main aim of this paper is to use appropriate simulations to identify the key traits of both benchmarking datasets and ML model evaluation. We considered the following : (a) the assumptions of the data generating process, (b) non-linear patterns that exist in the datasets, (c) sparisty of signals, (d) distributional shifts and need for domain adaptation, (e) the imbalance datasets, (f) need for careful choices of performance evaluation and optimization metrics problems and simulated data accordingly and these were discussed in details in the literature review section.

Linear models performed well with the data where there exists linearity between the predictor variables and the outcome variable. It is discovered that when there exists non-linearity between the predictor variables and the outcome variable, linear models haven't performed well. But when higher degree polynomial features were added to the data, linear models started to perform well. We also understood that performing regularization is important when there exists a sparse signal in the data. It is found that some metrics will not work with imbalanced data and also a careful choice of performance metrics is very important.

References

[1] R. Dattakumar and R. Jagadeesh, "A review of literature on benchmarking," *Benchmarking Int. J.*, vol. 10, no. 3, pp. 176–209, Jan. 2003, doi: 10.1108/14635770310477744.

[2] J. Thiyagalingam, M. Shankar, G. Fox, and T. Hey, "Scientific machine learning benchmarks," *Nat. Rev. Phys.*, vol. 4, no. 6, pp. 413–420, Jun. 2022, doi: 10.1038/s42254-022-00441-7.

[3] "[1707.04131] Foolbox: A Python toolbox to benchmark the robustness of machine learning models." https://arxiv.org/abs/1707.04131 (accessed Nov. 14, 2022).

[4] H. Lv *et al.*, "Application of artificial intelligence and machine learning for COVID-19 drug discovery and vaccine design," *Brief. Bioinform.*, p. bbab320, Aug. 2021, doi: 10.1093/bib/bbab320.

[5] "Mining adaptive immune receptor repertoires for biological and clinical information using machine learning - ScienceDirect." https://www.sciencedirect.com/science/article/pii/S2452310020300524 (accessed Nov. 13, 2022).

[6] R. A. Arnaout *et al.*, “The Future of Blood Testing Is the Immunome,” *Front. Immunol.*, vol. 12, 2021, Accessed: Nov. 18, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fimmu.2021.626793
[7] R. O. Emerson *et al.*, “Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire,” *Nat. Genet.*, vol. 49, no. 5, Art. no. 5, May 2017, doi: 10.1038/ng.3822.
[8] J. Glanville *et al.*, “Identifying specificity groups in the T cell receptor repertoire,” *Nature*, vol. 547, no. 7661, Art. no. 7661, Jul. 2017, doi: 10.1038/nature22976.
[9] “Artificial intelligence predicts the immunogenic landscape of SARS-CoV-2 leading to universal blueprints for vaccine designs - PubMed.” https://pubmed.ncbi.nlm.nih.gov/33361777/ (accessed Nov. 14, 2022).
[10] S. Faghani *et al.*, “Mitigating bias in radiology machine learning: 3. Performance metrics,” *Radiol. Artif. Intell.*, p. e220061, 2022.
[11] M. Du, N. Liu, and X. Hu, “Techniques for interpretable machine learning,” *Commun. ACM*, vol. 63, no. 1, pp. 68–77, Dec. 2019, doi: 10.1145/3359786.
[12] “Highlighting nonlinear patterns in population genetics datasets | Scientific Reports.” https://www.nature.com/articles/srep08140 (accessed Nov. 14, 2022).
[13] X. Lv, G. Bi, and C. Wan, “The Group Lasso for Stable Recovery of Block-Sparse Signal Representations,” *IEEE Trans. Signal Process.*, vol. 59, no. 4, pp. 1371–1382, Apr. 2011, doi: 10.1109/TSP.2011.2105478.
[14] “[2107.07455] Shifts: A Dataset of Real Distributional Shift Across Multiple Large-Scale Tasks.” https://arxiv.org/abs/2107.07455 (accessed Nov. 14, 2022).
[15] G. S. Handelman *et al.*, “Peering Into the Black Box of Artificial Intelligence: Evaluation Metrics of Machine Learning Methods,” *Am. J. Roentgenol.*, vol. 212, no. 1, pp. 38–43, Jan. 2019, doi: 10.2214/AJR.18.20224.
[16] A. Singh, N. Thakur, and A. Sharma, “A review of supervised machine learning algorithms,” in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, Mar. 2016, pp. 1310–1315.
[17] R. Gentleman and V. J. Carey, “Unsupervised Machine Learning,” in *Bioconductor Case Studies*, F. Hahne, W. Huber, R. Gentleman, and S. Falcon, Eds. New York, NY: Springer, 2008, pp. 137–157. doi: 10.1007/978-0-387-77240-0_10.
[18] B. J. Erickson and F. Kitamura, “Magician’s corner: 9. Performance metrics for machine learning models,” *Radiol. Artif. Intell.*, vol. 3, no. 3, 2021.
[19] M. Z. Naser and A. H. Alavi, “Error metrics and performance fitness indicators for artificial intelligence and machine learning in engineering and sciences,” *Archit. Struct. Constr.*, pp. 1–19, 2021.
[20] M. I. Jordan and T. M. Mitchell, “Machine learning: Trends, perspectives, and prospects,” *Science*, vol. 349, no. 6245, pp. 255–260, Jul. 2015, doi: 10.1126/science.aaa8415.
[21] A. Bajaj, “Performance Metrics in Machine Learning [Complete Guide].” 2022.
[22] S. Orozco-Arias, J. S. Piña, R. Tabares-Soto, L. F. Castillo-Ossa, R. Guyot, and G. Isaza, “Measuring performance metrics of machine learning algorithms for detecting and classifying transposable elements,” *Processes*, vol. 8, no. 6, p. 638, 2020.
[23] P. Brereton, B. A. Kitchenham, D. Budgen, M. Turner, and M. Khalil, “Lessons from applying the systematic literature review process within the software engineering domain,” *J. Syst. Softw.*, vol. 80, no. 4, pp. 571–583, Apr. 2007, doi: 10.1016/j.jss.2006.07.009.
[24] D. Wallach and B. Goffinet, “Mean squared error of prediction as a criterion for evaluating and comparing system models,” *Ecol. Model.*, vol. 44, no. 3, pp. 299–306, Jan. 1989, doi: 10.1016/0304-3800(89)90035-5.
[25] T. Chai and R. R. Draxler, “Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature,” *Geosci. Model Dev.*, vol. 7, no. 3, pp. 1247–1250, Jun. 2014, doi: 10.5194/gmd-7-1247-2014.

[26] A. de Myttenaere, B. Golden, B. Le Grand, and F. Rossi, “Mean Absolute Percentage Error for regression models,” *Neurocomputing*, vol. 192, pp. 38–48, Jun. 2016, doi: 10.1016/j.neucom.2015.12.114.
[27] T. O. Kvålseth, “Cautionary Note about R 2,” *Am. Stat.*, vol. 39, no. 4, pp. 279–285, Nov. 1985, doi: 10.1080/00031305.1985.10479448.
[28] J. Leach and U. Thayasivam, “Optimizing Data Evaluation Metrics for Fraud Detection Using Machine Learning,” *Int. J. Math. Comput. Sci.*, vol. 16, no. 8, pp. 52–59, 2022.
[29] R. R. Sanni and H. S. Guruprasad, “Analysis of performance metrics of heart failured patients using python and machine learning algorithms,” *Glob. Transit. Proc.*, vol. 2, no. 2, pp. 233–237, 2021.
[30] V. PLEVRIS, G. SOLORZANO, N. P. BAKAS, and M. E. A. B. SEGHIER, “INVESTIGATION OF PERFORMANCE METRICS IN REGRESSION ANALYSIS AND MACHINE LEARNING-BASED PREDICTION MODELS”.
[31] K. R. Shahapure and C. Nicholas, “Cluster Quality Analysis Using Silhouette Score,” in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, Oct. 2020, pp. 747–748. doi: 10.1109/DSAA49011.2020.00096.
[32] U. Maulik and S. Bandyopadhyay, “Performance evaluation of some clustering algorithms and validity indices,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 12, pp. 1650–1654, Dec. 2002, doi: 10.1109/TPAMI.2002.1114856.
[33] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, “Machine learning: a review of classification and combining techniques,” *Artif. Intell. Rev.*, vol. 26, no. 3, pp. 159–190, Nov. 2006, doi: 10.1007/s10462-007-9052-3.
[34] V. Petrov, A. Gennadinik, and E. Avksentieva, “Metrics for machine learning evaluation methods in cloud monitoring systems,” in *Proceedings of the 2022 8th International Conference on Computer Technology Applications*, 2022, pp. 168–175.
[35] E. W. Steyerberg, T. van der Ploeg, and B. Van Calster, “Risk prediction with machine learning and regression methods,” *Biom. J.*, vol. 56, no. 4, pp. 601–606, 2014, doi: 10.1002/bimj.201300297.
[36] A. Botchkarev, “Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology,” *ArXiv Prepr. ArXiv180903006*, 2018.
[37] T. Gupta and S. P. Panda, “Clustering Validation of CLARA and K-Means Using Silhouette & DUNN Measures on Iris Dataset,” in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, Feb. 2019, pp. 10–13. doi: 10.1109/COMITCon.2019.8862199.