Automated machine learning selection system for general prediction

Yashodhan Ketkar[a], Sushopti Gawade[b]

[a]Department of Information Technology Engineering, Pillai College of Engineering, New Panvel, 410206, India
[b]Department of Computer Engineering, Pillai College of Engineering, New Panvel, 410206, India

ARTICLE INFO

Keywords:
Automated Selection System
General Prediction
Machine Learning
Machine Learning in Medical Field
Supervised Learning Algorithm

ABSTRACT

The use of machine learning in various fields is still limited. The driving reason behind this is the lack of ease-to-use systems for non-technical people. The objective of this paper is to provide the general population access to a machine learning system. We propose an automated machine learning system for non-technical users. The proposed system automates the selection of the best model as per user requirements. We employed the proposed system on 3 different datasets in this study. The proposed system showed high accuracy in training the machine learning models and the selection of appropriate models as per user demand. The system showed improvement in performance with a larger data size. As per tests conducted the proposed system showed satisfactory performance.

1. Introduction

In recent years, machine learning has become very popular. Earlier, machine learning was limited to the research field. But quite recently, it has been used in various fields. This is partially due to the increasing availability of computers and growth in computing power.

The researchers are using machine learning in various fields. One of such fields is solar energy production. In this field, the prediction of solar radiation available per day can be very significant. Ağbulut, Gürel and Biçen (2021) employed a few machine learning algorithms for the prediction of the amount of solar radiation received in a day. Ağbulut et al. used natural properties such as weather, time, etc as features for the machine learning algorithms. Ağbulut et al. concluded that supervised learning models gave satisfactory results in predicting the amount of solar radiation. Ağbulut et al. concluded that the ANN model showed great potential for such specialized tasks.

In the field of chemistry, there is a wide range of factors that need to be considered. Zuranski, Martinez Alvarado, Shields and Doyle (2021) used machine learning in the field of chemistry. Zuranski et al. suggest that the models performed better than conventional statistical methods. Zuranski et al. concluded that more research is necessary before implementation.

Machine learning can be applied in network security for intrusion detection. Maseer, Yusof, Bahaman, Mostafa and Foozy (2021) state that the supervised learning methods perform satisfactorily for intrusion detection. Maseer et al. also remark that the supervised learning methods are limited to conventional problems.

Sarker (2021) in thier study used machine learning models to solve real-life problems. Sarker concluded that machine learning plays an important role in good decision-making. Sarker also remark that machine learning can be used in various fields.

ketkaryapr19me@student.mes.ac.in (Y. Ketkar); sgawade@mes.ac.in (S. Gawade)

Deo (2015) note a lack of information and meaningful research in the medical field. Deo suggest that more support, subject-specific scope, and accuracy are important factors for the higher impact of machine learning in the medical field. Deo also suggest that the ideal system should be able to take multiple data types for training.

Ghazal, Hasan, Alshurideh, Alzoubi, Ahmad, Akbar, Al Kurdi and Akour (2021) used machine learning with a combination of IoT for smart cities. Ghazal et al. concludes that machine learning showed promising results for smart cities. Ghazal et al. remark that machine learning was able to handle the high volume of data generated by sensors. Liu, Zhang, Hou, Mian, Wang, Zhang and Tang (2021) in thier paper suggests that machine learning can handle downstream tasks efficiently.

This suggests that even with the immense popularity and accessibility of machine learning, it is still underutilized by the general population. This can be attributed to lack of knowledge and skills. To solve this problem we are proposing a system with minimum user interaction. The proposed system can be operated by users without prior knowledge of machine learning. This system allows them to train their machine learning algorithms.

The proposed system can be deployed on a local network. The system can feed data manually or automatically according to user needs and policies. The local deployment also limits external access and reduces the influence of external factors on data.

2. Literature Review

Supervised learning algorithms are widely used in various fields. Burkart and Huber (2021) conducted a study on the application of machine learning in various industries. Burkart and Huber suggests that various industries already use supervised learning algorithms to solve various problems. Flah, Nunez, Ben Chaabene and Nehdi (2021) used machine learning in Civil Structural Health Monitoring (SHM). Flah et al. used supervised learning methods due to the abundance of well-labeled data. Flah et al. concluded that machine learning showed satisfactory results.

Supervised learning algorithms are very efficient in conventional problems. Pande, Khamparia, Gupta and Thanh (2021) used machine learning to detect the DDOS attack. Pande et al. used five main security factors as features. Pande et al. conclude that the random forest algorithm was able to detect intrusion with high accuracy.

Lee and Lin (2000) conducted a study on the automated selection of SVM algorithms. The goal of the study was to select the most optimal SVM model for a given task. Lee and Lin suggested that hyperparameters and used data have an extreme impact on prediction time.

Shimpi, Shah, Shroff and Godbole (2017) used machine learning to detect arrhythmia. Shimpi et al. concluded that SVM achieved up to 91.2% accuracy. Guvenir, Acar, Demiroz and Cekin (1997) also used machine learning to detect arrhythmia. The neural network provided good results.

Recently machine learning was used in the diagnosis of COVID-19 disease and study related to its structure. Chadaga, Prabhu, Vivekananda, Niranjana and Umakanth (2021) surveyed the research conducted on the machine learning approach to diagnose the COVID-19. Chadaga et al. suggested that the supervised learning algorithms showed promising results in diagnosis. Chadaga et al. suggested that various types of datasets were used in these studies.

Machine learning is used in the diagnosis process. Mandal and Sairam (2014) machine learning is used for the detection of Parkinson's disease. The machine learning algorithm achieved 100% accuracy with a 95-99% confidence level. The RF and SVM showed better performance compared to other algorithms. In their paper Ghaderzadeh, Asadi, Hosseini, Bashash, Abolghasemi and Roshanpour (2021) suggest that the inclusion of machine learning in diagnosis can lead to better results and higher accuracy. Ghaderzadeh et al. mentions that machine learning can handle a high volume of data.

Mei, Desrosiers and Frasnelli (2021) reviewed 200 studies published about the use of machine learning in Parkinson's disease diagnosis. Mei et al. found out that the use of machine learning improved clinical decisions. In thier study Prashanth, Roy, Mandal and Ghosh (2016) used machine learning for the detection of Parkinson's disease. The SVM model outperformed other models. Prashanth et al. concludes that machine learning provides a better detection method.

Arrhythmias are a symptom of cardiological disorders. Alfaras, Soriano and Ortín (2019) used a supervised learning method for the detection of arrhythmia. The machine learning system provided satisfactory results. Results showed high accuracy and sensitivity. Alfaras et al. remark on the need for good feature selection and selection guidelines for better-performing algorithms.

The machine learning system needs to be robust, extremely accurate, and easy to use. In the field of healthcare, according to Soman and Bobbie (2005) these are essential requirements. In this study, Soman and Bobbie used machine learning to classify arrhythmia. Soman and Bobbie remark that early detection of arrhythmia is critical for better treatment.

Luo (2016) conducted a study on the automatic selection of machine learning algorithms and their hyperparameters. Luo showed limitations in the biomedical industry with the help of machine learning systems. Qayyum, Qadir, Bilal and Al-Fuqaha (2020) studied the security and privacy aspect of machine learning solutions. Qayyum et al. suggest that while machine learning has great potential in the healthcare system, more research about security and privacy aspects is necessary.

Machine learning utilizes various approaches for the same problem. In thier study Kim, Lee, Park, Baek and Lee (2021) used two approaches to solve a diagnosis problem. In a direct approach, data was fed to a machine learning model. In an indirect approach, data was equalized before the machine learning process. The indirect method showed better results compared to the direct method. While the experiment was successful, Kim et al. suggests that machine learning is still unstable for the medical field.

The supervised learning methods are very effective in the medical field. Ibrahim and Abdulazeez (2021) suggest that ensembled supervised learning systems can further improve effectiveness. Ibrahim and Abdulazeez further remark that machine learning will reduce the number of errors caused by humans.

In the review about the use of machine learning in the medical field, Greener, Kandathil, Moffat and Jones (2022) suggested that big companies are already using machine learning for various tasks. Greener et al. suggested that a machine learning system needs to be handled by non-technical people and it should support various ranges and types of data.

The automated system will allow non-technical people to use machine learning. Ayat, Cheriet and Suen (2005) studied the automatic model selection for optimal SVM kernels. Ayat et al. used different automation approaches for optimal hyperparameter calculations. Ayat et al. state that the system can calculate up to two parameters without human interaction.

The automatic selection process can be extremely beneficial in dynamic environments. The excavation of soil or tunneling is one such environment. To predict the displacement induced by excavation, Zhang, Shen, Huang and Xie (2022) used machine learning. Zhang et al. used properties of soli and imputed them as features. Zhang et al. concluded that the unsupervised machine learning algorithm GA-MLP showed good potential, while AutoML is found to be the most optimal algorithm in these dynamic conditions. Maschler and Weyrich (2021) also suggested that machine learning can be used in dynamic environments successfully.

There are multiple ways to select the best-suited model, the important part of such a system is the selection system. The system uses various approaches for selection, ranking is one of those approaches. In thier study Brazdil, Soares and Da Costa (2003) suggest that the meta ranking is better

than the baseline ranking system. Brazdil et al. used a multi-criteria ranking system for the selection of ideal classifiers. Brazdil et al. suggest that this study laid some groundwork for future automated machine learning selection systems.

3. Design and Implmentation

3.1. System Architecture

The goal of the system is to be used by non-technical users. The system consists of three processes or modules. These processes are the training process, selection process, and prediction process. The first two processes interact with each other. Their task is to produce the best-suited model for the user's needs. The third process interacts with the selected models to generate the predictions. Figure 1, shows the architecture of the system. The user is indirectly allowed to access the training and prediction process. The user data is stored on a local drive for easier and faster access.

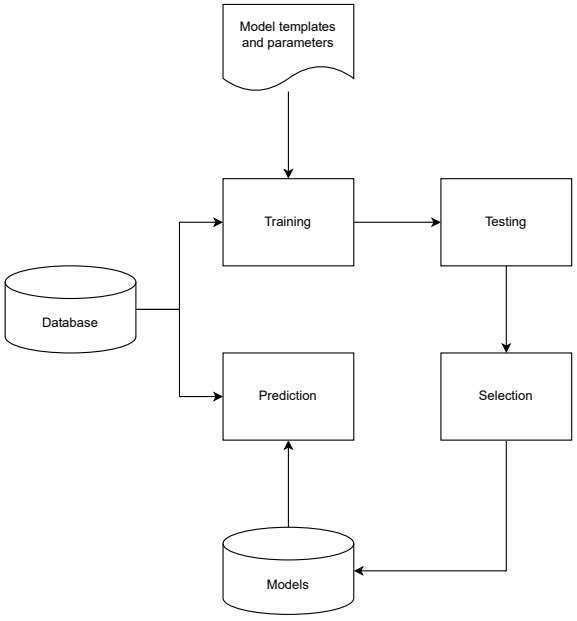

Figure 1: System Architecture

3.2. System Processes

As stated in previous section, the system consists of three primary processes. These processes are the training process, selection process, and prediction process. These processes are described in the following sections.

3.2.1. Training Process

The training process is the first process in the system. Figure 2, shows the structure of the training process. The training process has many functions. The first function is gathering data from the user and pre-processing it for training purposes. The training process also generates models with the template, which contains model structure and parameters. These models are trained with processed data and stored for future use. The performance of models is also calculated during this phase and store for selection purpose.

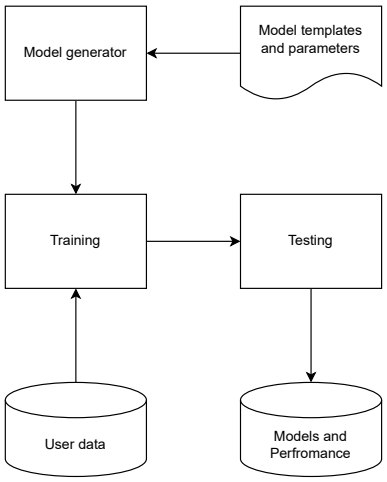

Figure 2: Training Process

3.2.2. Selection Process

The selection process is second in the system. Figure 3, shows the structure of the training process. This process evaluates the performance of models based on metrics and weightage. The metrics of models are generated during the training process. The performance weightage is defined by the user depending on requirements. The final performance score is calculated and used for selection of the best model for users task. The selected model is stored in a separate directory with a label for easier access. This model will be used for prediction problems.

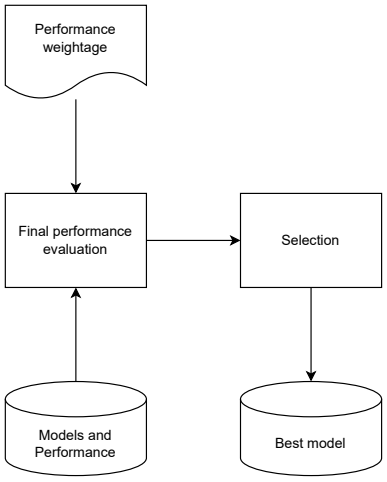

Figure 3: Selection Process

3.2.3. Prediction Process

The prediction process is the final process in the system. Figure 4, shows the structure of the prediction process. This process unpacks the best model and loads it for prediction. The model generates predictions with user-provided data. The output is displayed to the user. This output is also stored for future reference.

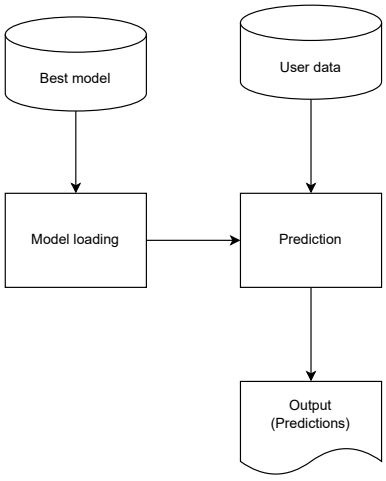

Figure 4: Prediction Process

3.3. Algorithms

The system uses two algorithms to run. These algorithms are training and selection algorithms and prediction algorithms.

Training and Selection Algorithm

1. Collect/Receive dataset.
2. Split data into 80:20 ratio for training and testing.
3. Build a model from presets.
4. Train models with training dataset and store models.
5. Evaluate the performance of models with the testing dataset.
6. Rank models with help of performance and premade tuning parameters.
7. Selects the best model and stores it for future use.

Prediction Algorithm

1. Collect/Receive dataset.
2. Load best-suited model data from the storage.
3. Unpack model for predictions.
4. Make predictions with the provided dataset and loaded model.
5. Return predictions to user.

3.4. Implmentation

3.4.1. Web Architecture

The system provides service with a web application. The users aren't allowed to interact with the system directly. This provision is to provide security and reduce the outside influence on results. The interface layer is used for a user to interact with the system indirectly. Figure 5, shows the web architecture of the system. The system is connected to the database directly. A direct connection is provided to access live data. User-provided data is processed by the system and stored in the database.

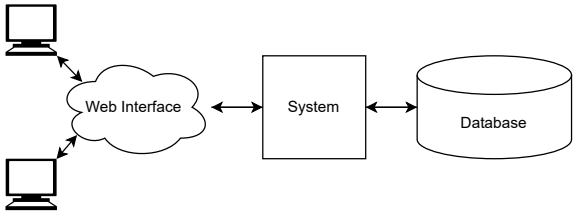

Figure 5: Web Architecture

Web Interface

The web interface provides users with a way to interact with the system. The primary function of the web interface is to provide secure access to the system process. Approved users can interact with various modules on the system. This interface allows users to access selection and prediction processes. Users can upload data to the system.

The interface presents prediction results as well as performance evaluation results. Prediction results are provided in tabular format and stored in CSV files for future reference. Performance results are presented in graphical format and also stored in CSV files for future reference.

4. Result and Analysis

4.1. Dataset

To test the selection system, we used three datasets. All three datasets have very different structures. Dataset 1 contains 149 records with 23 features [4]. Dataset 2 contains 201 records with 754 features [8]. Dataset 3 contains 65661 records with 188 features [10].

Dataset 1 is the smallest dataset with a lower number of features compared to other two datasets. Dataset 2 also contains few records, but contains the highest amount of features compared to other two datasets. While, dataset 3 has a large number of records and moderate amount of features.

4.2. Results

After providing the datasets mentioned in the previous section, the system trained a few models on those datasets. The application selected the best-suited model based on datasets.

Figure 6, shows the performance of various models trained on dataset 1. Due to the small number of records and few features model is overtrained in the case of few models. The values in Table 1, suggest that KNN and RF models were overtrained. The KNN model is selected as the best model for this dataset.

Figure 7, shows the performance of various models trained on dataset 2. This dataset also had a small number of records, but the number of features is extremely high. The values from Table 2, shows the random forest performed better across all parameters but prediction time. The system selected the Random Forest model as best suited mode.

Figure 8, shows the performance of various models trained on dataset 3. This dataset had a very high number of records and a moderate amount of features. The values

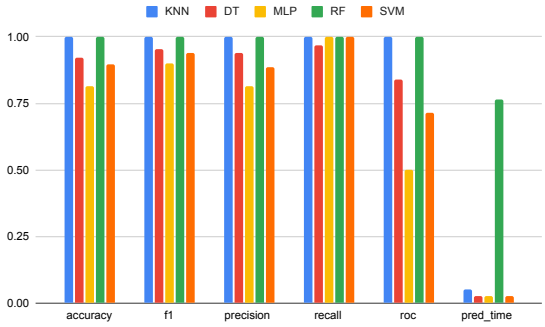

Figure 6: Performance Analysis Dataset 1

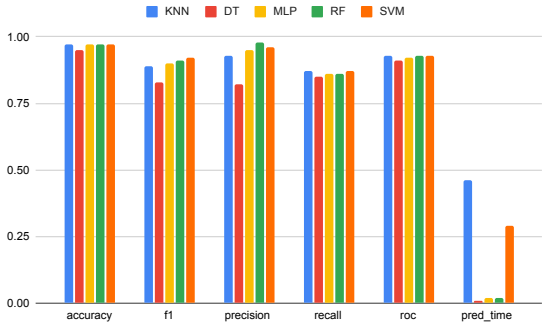

Figure 8: Performance Analysis Dataset 3

Table 1
Performance Analysis of Dataset 1

Metrics	KNN	DT	MLP	RF	SVM
Accuracy	1.00	0.92	0.82	1.00	0.89
F1	1.00	0.95	0.89	1.00	0.94
Pricision	1.00	0.93	0.82	1.00	0.88
Recall	1.00	0.97	1.00	1.00	1.00
ROC	1.00	0.84	0.50	1.00	0.71
Time	0.05	0.02	0.02	0.76	0.02

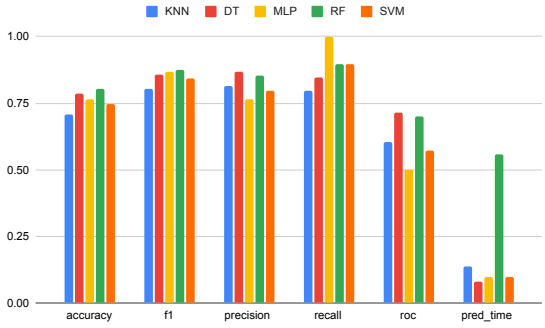

Figure 7: Performance Analysis Dataset 2

Table 2
Performance Analysis of Dataset 2

Metrics	KNN	DT	MLP	RF	SVM
Accuracy	0.71	0.78	0.76	0.80	0.75
F1	0.81	0.86	0.87	0.88	0.84
Precision	0.82	0.67	0.76	0.85	0.80
Recall	0.79	0.85	1.00	0.90	0.90
ROC	0.61	0.71	0.50	0.70	0.57
Time	0.13	0.07	0.09	0.55	0.09

from Figure 8, shows that almost all models performed satisfactorily on this dataset. The system selected the SVM model as best suited mode.

Table 3
Performance Analysis of Dataset 3

Metrics	KNN	DT	MLP	RF	SVM
Accuracy	0.97	0.95	0.97	0.97	0.97
F1	0.89	0.83	0.90	0.91	0.92
Precision	0.93	0.82	0.95	0.98	0.96
Recall	0.87	0.85	0.86	0.86	0.87
ROC	0.93	0.91	0.92	0.93	0.93
Time	0.46	0.01	0.02	0.02	0.29

4.3. Key Findings

From the data displayed in previous section, we saw how the system works on different datasets. These are a few key findings we obtained from that knowledge.

1. The system successfully works on different scales of data.
2. Systems performance does not depends on the number of records.
3. Systems performance is dependent on the number of features of a dataset.
4. The system is prone to overfitting depending on the scale of the dataset. Specifically, KNN and RF models tend to overfit with the small scale of data.
5. The system successfully uses the tuning parameters to select the most suited model from the dataset.

4.4. Benefits of the System

The system can be used for any scale of data. It allows users with limited prior knowledge easier access to machine learning technology. As seen in previous section, the system can train multiple models effectively.

The user-defined fine-tuning parameters lead to the selection of the best-suited model for particular tasks. This model can be used to generate predictions in the case of similar datasets. The performance of all models is stored for future evaluation of the system.

4.5. Improvements On the System

Currently, the system is limited to only five machine learning algorithms. With the smaller scale of data, specifically with a small number of features, the system tends to

overfit the models. Allowing users to implement their model templates will solve both of these problems.

These algorithms are supervised learning algorithms. The supervised nature of these algorithms limits the training dataset to the labeled dataset. By providing support to unsupervised learning algorithms, systems can accommodate various types of training datasets.

The selection parameters are adjusted before the training process. These predefined parameters restrict the selection choices of the system. Allowing users to tweak selection parameters can increase systems choices.

5. Conclusion and Future Work

The system performed satisfactorily during the tests. The application was able to select the model for the provided dataset. The best-suited model showed great accuracy during the prediction process. The whole process required minimum human interaction.

Future work focuses on the implementation of unsupervised learning algorithms. With this, the system will be able to use unlabeled datasets for training. Implementation of continuous learning will be another focus. This will massively increase the adaptability and efficiency of the system.

References

Ağbulut, Ü., Gürel, A.E., Biçen, Y., 2021. Prediction of daily global solar radiation using different machine learning algorithms: Evaluation and comparison. Renewable and Sustainable Energy Reviews 135, 110114.

Alfaras, M., Soriano, M.C., Ortín, S., 2019. A fast machine learning model for ecg-based heartbeat classification and arrhythmia detection. Frontiers in Physics 7, 103.

Ayat, N.E., Cheriet, M., Suen, C.Y., 2005. Automatic model selection for the optimization of svm kernels. Pattern Recognition 38, 1733–1745.

Biswas, D., (2019, May). Parkinson's Disease (PD) classification, Version 2. https://www.kaggle.com/datasets/dipayanbiswas/parkinsons-disease-speech-signal-features. Accessed on 20 Apr 2022.

Brazdil, P.B., Soares, C., Da Costa, J.P., 2003. Ranking learning algorithms: Using ibl and meta-learning on accuracy and time results. Machine Learning 50, 251–277.

Burkart, N., Huber, M.F., 2021. A survey on the explainability of supervised machine learning. Journal of Artificial Intelligence Research 70, 245–317.

Chadaga, K., Prabhu, S., Vivekananda, B.K., Niranjana, S., Umakanth, S., 2021. Battling covid-19 using machine learning: A review. Cogent Engineering 8, 1958666.

Debasis, S., (2020, Oct). Parkinson Disease Detection, Version 2. https://www.kaggle.com/datasets/debasisdotcom/parkinson-disease-detection?sort=votes. Accessed on 20 Apr 2022.

Deo, R.C., 2015. Machine learning in medicine. Circulation 132, 1920–1930.

Fazeli, S., (2018, Jun). ECG Heartbeat Categorization Dataset, Version 1. https://www.kaggle.com/datasets/shayanfazeli/heartbeat. Accessed on 15 Dec 2020.

Flah, M., Nunez, I., Ben Chaabene, W., Nehdi, M.L., 2021. Machine learning algorithms in civil structural health monitoring: a systematic review. Archives of computational methods in engineering 28, 2621–2643.

Ghaderzadeh, M., Asadi, F., Hosseini, A., Bashash, D., Abolghasemi, H., Roshanpour, A., 2021. Machine learning in detection and classification of leukemia using smear blood images: a systematic review. Scientific Programming 2021.

Ghazal, T.M., Hasan, M.K., Alshurideh, M.T., Alzoubi, H.M., Ahmad, M., Akbar, S.S., Al Kurdi, B., Akour, I.A., 2021. Iot for smart cities: Machine learning approaches in smart healthcare—a review. Future Internet 13, 218.

Greener, J.G., Kandathil, S.M., Moffat, L., Jones, D.T., 2022. A guide to machine learning for biologists. Nature Reviews Molecular Cell Biology 23, 40–55.

Guvenir, H.A., Acar, B., Demiroz, G., Cekin, A., 1997. A supervised machine learning algorithm for arrhythmia analysis, in: Computers in Cardiology 1997, IEEE. pp. 433–436.

Ibrahim, I., Abdulazeez, A., 2021. The role of machine learning algorithms for diagnosing diseases. Journal of Applied Science and Technology Trends 2, 10–19.

Kim, I.K., Lee, K., Park, J.H., Baek, J., Lee, W.K., 2021. Classification of pachychoroid disease on ultrawide-field indocyanine green angiography using auto-machine learning platform. British Journal of Ophthalmology 105, 856–861.

Lee, J.H., Lin, C.J., 2000. Automatic model selection for support vector machines. CSIE, NTU .

Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., Tang, J., 2021. Self-supervised learning: Generative or contrastive. IEEE Transactions on Knowledge and Data Engineering .

Luo, G., 2016. A review of automatic selection methods for machine learning algorithms and hyper-parameter values. Network Modeling Analysis in Health Informatics and Bioinformatics 5, 1–16.

Mandal, I., Sairam, N., 2014. New machine-learning algorithms for prediction of parkinson's disease. International Journal of Systems Science 45, 647–666.

Maschler, B., Weyrich, M., 2021. Deep transfer learning for industrial automation: a review and discussion of new techniques for data-driven machine learning. IEEE Industrial Electronics Magazine 15, 65–75.

Maseer, Z.K., Yusof, R., Bahaman, N., Mostafa, S.A., Foozy, C.F.M., 2021. Benchmarking of machine learning for anomaly based intrusion detection systems in the cicids2017 dataset. IEEE access 9, 22351–22370.

Mei, J., Desrosiers, C., Frasnelli, J., 2021. Machine learning for the diagnosis of parkinson's disease: A review of literature. Frontiers in aging neuroscience 13, 184.

Pande, S., Khamparia, A., Gupta, D., Thanh, D.N., 2021. Ddos detection using machine learning technique, in: Recent Studies on Computational Intelligence. Springer, pp. 59–68.

Prashanth, R., Roy, S.D., Mandal, P.K., Ghosh, S., 2016. High-accuracy detection of early parkinson's disease through multimodal features and machine learning. International journal of medical informatics 90, 13–21.

Qayyum, A., Qadir, J., Bilal, M., Al-Fuqaha, A., 2020. Secure and robust machine learning for healthcare: A survey. IEEE Reviews in Biomedical Engineering 14, 156–180.

Sarker, I.H., 2021. Machine learning: Algorithms, real-world applications and research directions. SN Computer Science 2, 1–21.

Shimpi, P., Shah, S., Shroff, M., Godbole, A., 2017. A machine learning approach for the classification of cardiac arrhythmia, in: 2017 international conference on computing methodologies and communication (ICCMC), IEEE. pp. 603–607.

Soman, T., Bobbie, P.O., 2005. Classification of arrhythmia using machine learning techniques. WSEAS Transactions on computers 4, 548–552.

Zhang, D., Shen, Y., Huang, Z., Xie, X., 2022. Auto machine learning-based modelling and prediction of excavation-induced tunnel displacement. Journal of Rock Mechanics and Geotechnical Engineering .

Zuranski, A.M., Martinez Alvarado, J.I., Shields, B.J., Doyle, A.G., 2021. Predicting reaction yields via supervised learning. Accounts of chemical research 54, 1856–1865.