JOSEIT JOURNAL OF SYSTEMS ENGINEERING AND INFORMATION TECHNOLOGY
http://jurnal.iaii.or.id/index.php/JOSEIT

An Optimal Solution to the Overfitting and Underfitting Problem of Healthcare Machine Learning Models

Anil Kumar Prajapati[1], Umesh Kumar Singh[2]
[12]Institute of Computer Science, Vikram University Ujjain, Madhya Pradesh, India
anilprajapatiujjain@gmail.com[1], umeshsingh@rediffmail.com[2]

Abstract

In the current technological era, artificial intelligence is becoming increasingly popular. Machine learning, as the branch of AI is taking charge in every field such as healthcare, the Stock market, Automation, Robotics, Image Processing, and so on. In the current scenario, machine learning and/or deep learning are becoming very popular in medical science for disease prediction. Much research is underway in the form of disease prediction models by machine learning. To ensure the performance and accuracy of the machine learning model, it is important to keep some basic things in mind during training. The machine learning model has several issues which must be rectified duration of the training of the model so that the learning model works efficiently such as model selection, parameter tuning, dataset splitting, cross-validation, bias-variance tradeoff, overfitting, underfitting, and so on. Under- and over-fitting are the two main issues that affect machine learning models. This research paper mainly focuses on minimizing and/or preventing the problem of overfitting and underfitting machine learning models.

Keywords: Machine learning, Underfitting, Overfitting, Bias-Variance, Cross-validation, Data Splitting, Parameter Tuning, Loss Function

1. Introduction

The machine learning models are trained on the basis of previous datasets and knowledge, and they are being deployed for a large number of industrial uses. The performance of the model depends on the dataset and/or prior knowledge used in model training. Nowadays, machine learning and/or deep learning are being used extensively in disease detection. Machine learning and/or deep learning techniques for disease detection use a variety of disease symptom datasets, which contain a large amount of heterogeneity. For example, image and text datasets are used for breast cancer and/or skin cancer, while text and numerical datasets are used for heart disease and/or liver disease, which have a lot of variation.

There is a wide range of real-world datasets that are highly diverse including text, numeric, audio, video, and image datasets. Managing models built on these different datasets brings additional challenges. Training a model is also challenging because of the high degree of variation in the numerical dataset. Due to the accurate prediction and strong computational power, machine learning and deep learning disease prediction models are trained on datasets with high variance, which also have challenges such as underfitting and overfitting. It is very lengthy and tough work to understand the dataset and their variance. Most datasets contain text and numerical, video and text, audio and text, and video and numerical [1].

Training the model by straightening text, audio, video and images as raw data through mathematical computation is a difficult task and presents challenges [2-3]. In order to improve the performance, accuracy, and consistency of the disease prediction model, it is very important to minimize such problems as bias variance, feature extraction, overfitting, and underfitting. The paper is organized into the following sections. Section 2 introduces the model training steps. Section 3 describes the problems faced in model training, and Section 4 describes the available remedies. Section 5 discusses the proposed remedies, and Section 6 presents the conclusions.

2. Model Training Phases

To build an effective, intelligent, and robust disease diagnostic machine learning model for real issues and/or problems, the diagnostic model works on various kinds of dataset such as text, numerical, audio, video, image, and so on. Model training is a very tough task according to the data set. Hence, it is very important to keep in mind some basic steps which are shown below.

1.1. Model Selection

Selecting the most appropriate model for a particular problem is known as model selection in machine learning. In machine learning techniques, there are many models available and it is very difficult to select the appropriate

Accepted by editor: 26-09-2023 | Final revision: 30-09-2023 | Online publication: 30-09-2023

model for a specific disease prediction. Different models have different specialties, and they provide performance according to their use. Therefore, when selecting a model, it is very important to take into account certain factors such as the type of dataset, the function to be performed by the model, the nature of the model, and so on.

A. Based on data sets

There are different types of data set available such as image and video, text or speech data, numerical data, etc. So, the CNN model is more reliable for images and video [4-5], the RNN model for text and speech [6], and logistic regression, SVM, random forest decision tree, etc. model for the numerical data set.

B. Based on Functionality

Models are based on their functionality such as classification tasks, regression tasks, and clustering tasks. In the Classification, only two choices which are yes or no. It means the classification model classifies the problem into two categories and generates the result. The regression model works on a linear problem such as an increase in the x-axis value, then the y-axis value also increases and generates results on the basis of regularity, and the clustering model develops clusters based on the similarities in the data set and generates the result [7].

1.2. Hyperparameter Tuning

Machine learning models have several parameters that determine their consistency, accuracy, and performance. There are two types of parameters: model parameters and hyperparameters. The model parameters are known as the internal parameters of a machine learning model, which can be determined by training with the training data. The basic model parameters are bias and weights. hyperparameter is known as an external parameter that can be managed during model training. These parameter values control the learning process that is adjustable and used for developing optimum machine learning models. There are certain types of hyperparameters such as learning rate, data split ratio, number of epochs, number of branches, clustering, cross-validation, etc. [8-9].

Hyperparameter tuning is the most important concept in machine learning model training because the application of wrong and inaccurate parameters makes the machine learning model inefficient and less effective. It is essential to manage the hyperparameters of the model according to the requirements of the dataset. Operates on the parameter data set and it is necessary to apply parameters according to the nature of the data set, otherwise the model will give an ineffective and incorrect prediction [10].

1.3. Loss function

The loss function works to calculate the performance of the machine learning model. The model calculates the difference between how far the estimated prediction value is from its true prediction value. For example, if a person has a blood pressure level of 165 and the model predicts a 145 then the difference is 20 which is known as the loss function. The loss function determines the model and parameter efficiency that can be better for particular data sets [11]. If the loss value of any model is close to zero, then the model has a good performance but if the loss value increases, then the model is not a good fit and needs to fix the parameter.

1.4. Data Splitting

There are two basic methods for training a machine learning model, the first is cross-validation, and the second is train-test split.

A. Cross-validation

In the cross-validation method, the entire data set is divided into k segments of equal size. The first or last segment is used as the test data set and the remaining k-1 segments are used as the training data set. For example, if the value of k is 5 then the data set is divided into 5 segments and each time a different segment is used as the test data set. The main drawback of this method is that the execution time is longer when the number of data sets is large [12-13].

B. *Train-Test Split*

In this method, the overall data set is divided into two parts and the division is based on the mindset of the model trainer. A small part of the data set is used as a test data set and the test data set is used for training. The splitting of the dataset may be like that 20/80, 30/30, 40,60, etc. The execution time of this technique is very shorter and is suitable for large datasets [14,15,16].

3. Model Training Problems

To develop an effective and intelligent machine learning model, it is very essential to understand the hidden threat and missing values of the model. The machine learning model is developed in various stages such as model selection, data gathering, data filtering, data spitting, data validation, and performance indices. To effectively train a model, each step must be understood and validated. There are several issues and/or problems associated with machine learning for Model training that are listed below.

3.1 *Overfitting*

When a machine learning model generates accurate results on a training dataset but is not a good fit for a new dataset and generates wrong results, it is called overfitting. Overfitting occurs when the model covers all the points in the training data set. The main reason for overfitting is noisy data and data overload when the data set contains irrelevant and noisy information [17]. Figure 2.0 underlines the overfitting problem which is described below.

Figure 1. Overfitted model

Causes of Overfitting: The fewer data and/or unfiltered data to be used in the machine learning model then the overfitting problem occurs; the second one is when we use a more complex model for a simple problem. When the layer is increased, the model going to be very complex and the problem of overfitting occurs. Hence, the use of a large number of layers in neural networks causes an overfitting model [18].

3.2 *Underfitting*

When a Machine learning model is trained on a limited number of datasets and/or features, then the model does not work efficiently and generates the wrong result, this is known as underfitting. The main reason for underfitting is low variance and high bias in model training. The overfitting problem occurs when the machine learning model does not cover all valuable datasets and/or features [19]. Figure 2.0 underlines the underfitting problem which is described below.

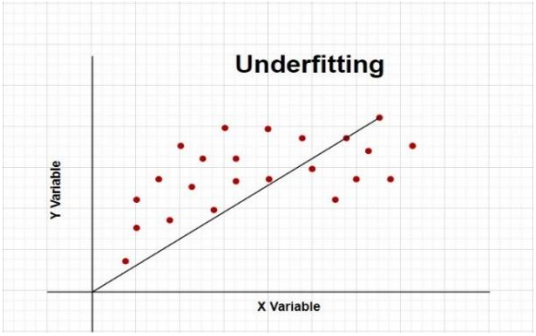

Figure 2. Underfitted model

Causes of underfitting: Underfitting occurs when the model does not learn enough from the data. when the accuracy of training data is very low then the accuracy of test data is also low, which is the main cause of the underfitting problem in machine learning. There are certain reasons for the underfitting problem, such as wrong model selection, less complexity of the model, and less variance but high bias [20].

3.3 Bias-Variance Tradeoff

The bias in the machine learning model is the difference between the prediction that is achieved by the machine learning model and the actual prediction that we are trying to predict. When the bias is high and the variance is low, the machine learning model is affected by the underfitting problem. A high-variance problem occurs when a

machine learning model is trained on a dataset with unvalued features or noise; the model trains very well but provides incorrect results when applied to new data. When the variance in the machine learning model is high and the bias is low, the machine learning model suffers from the underfitting problem [21-22]. Figure 3.0 underlines the high bias that is described in the following.

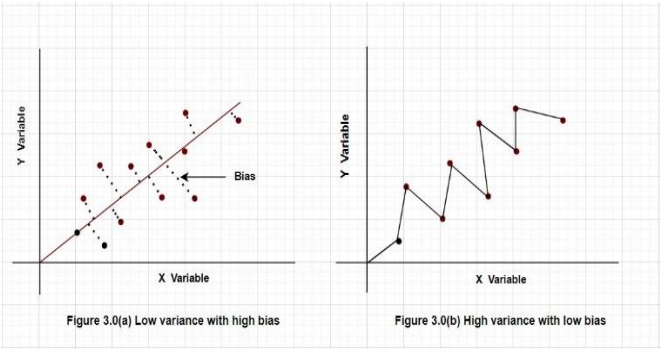

Figure 3. Bias-Variance Tradeoff

The above graph shows that when the bias is high, the variance is low (i.e., figure 3.0 a), and when the bias is low, the variance is high (i.e., figure 3.0 b). Hence, for the optimum model, the variance and bias have to be kept equal.

4. Existing remedies

Machine learning techniques are currently being used in many areas for disease diagnosis and/or prediction. For better efficiency and accuracy in the field of disease diagnosis, it is essential to deal with the problems of overfitting, underfitting, bias-variance tradeoffs, etc. In this section, some existing remedies that can be used to deal with these problems are highlighted.

4.1 Prevent Overfitting

Overfitting occurs when there is noise in the data and improper model selection for the particular problem; another reason is low bias and high variance. To avoid those issues, we take the steps as follows.

A. Model Selection: Need to understand the dataset, what type of dataset is available, and which model will provide better results for the available dataset. The CNN model for the video dataset, The RNN model for text and speech, regression for the numerical problems, machine random forest, or other models perform better.

B. Filter data set and feature extraction: The machine learning model is trained on the dataset and features. Therefore, reliable features and filtered datasets overcome the problem of overfitting. For feature extraction, four techniques are available such as FCBF, LASSO, MRMR, and RELIEF [23,24,25]. These techniques extract valuable features for model training.

C. Reducing Layers/ Complexity: the dense and more internal layer creates a neural network too complex. When the model has more complexity and or complex layers, the mode got a low bias, which also creates an overfitting problem. Therefore, it is very necessary to reduce and/or minimize the internal layers in neural networks.

D. Early Stopping: In the training session, the model learns the data multiple times and trains itself. Therefore, the model overtraining the training data causes overfitting. It is necessary to reduce the repetition of training and/or to stop the model early once the model is trained on the training data; this is called early stopping.

E. Use Dropouts: A neural network contains multiple layers, and each layer contains multiple neurons. To prevent overfitting some neurons, randomly drop, which is known as a dropout. This technique is more beneficial for the neural network.

4.2 Prevent Underfitting

Overfitting occurs when the machine learning model trains on less and/or a small number of datasets. The low variance and high bias are also the reason for the underfitting. There are some remedies to overcome the underfitting problem that are listed below.

A. Model Selection Maybe the dataset has high and/or low numerical values, good and/or bad video quality, high and/or speech quality, small and/or big image size, etc. Therefore, it is very necessary to choose a suitable machine learning model to deal with the problem of underfitting.

B. Increase Complexity and/or Layers: A lower number of data sets and a very small and less complex model do not train properly. Therefore, using complex models and increasing the dataset rectify the under-fitting issue.

C. Increasing parameter: This is another way to rectify underfitting issues. In this technique, some parameters are increased during model training, which increases the complexity of the model and removes the underfitting problem of the model.

4.3 Bias-Variance Tradeoff

The bias and variance are the backbones of model training that can be managed through parameter tuning, data splitting, cross-validation, model selection, feature selection, etc.

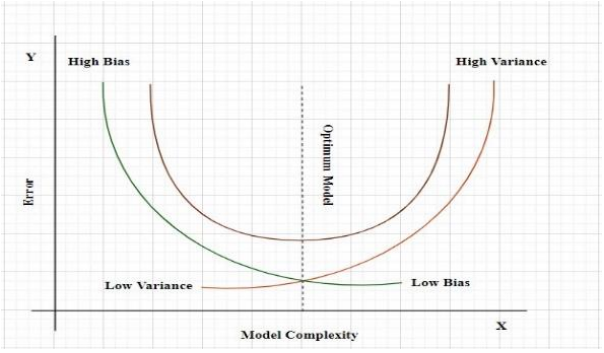

Figure 4. Optimum Complexity for Bias Variance

The diagram 4.0 provides the overview and optimum line for avoiding bias-variance trade-off problems. The centerline in the diagram represents the optimum level of bias and variance of the ML model. This line represents the complexity level of each ML model. There is a certain thing to managing bias and variance of the ML model, which are listed below.

a) Regularization: Regularization tries to reduce the value of the coefficient, which reduces the model complexity, and the model gets high performance.
b) Dimensionality Reduction: A large number of features and/or dimensions are available in the dataset. Dimension reduction is done so that the model does not train on useless features during training.

5. Proposed remedies

Bagging Ensemble Techniques: The performance and accuracy of the ML models depend on model training, feature selection, parameter tuning, and datasets. Hence, it is essential to understand and tune all parameters according to the problem statement. To minimize and/or remediation underfitting, and overfitting problems we have proposed and/or suggested the bagging method of ensemble techniques. In the bagging technique, the dataset is divided into several sub-datasets, and predictions are generated on each. The final prediction is generated based on the average of the prediction generated from each and every subdataset. This process is also known as bootstrap and aggregation. Figure 5.0 represents the working process of the proposed ensemble bagging method.

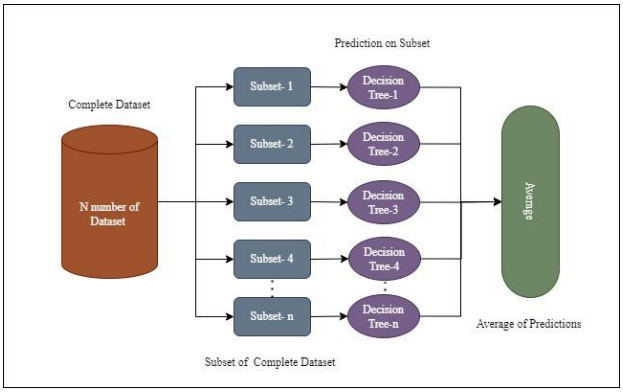

Figure 5. Ensemble model for the bagging method

Experimental Setup & Results: Experimental Setup & Results: For practical implementation, we have used Python programming language (Version 3.9.12) in Jupyter Notebook (version 6.4.8) on Anaconda Navigator. In the above practical setup, we have used four different disease datasets such as cardiovascular, diabetes, liver disease, and pregnancy datasets which are freely available on the Kaggle website for practical purposes only. We used 1025 for heart disease, 583 for liver disease, 768 for diabetes and 2000 for pregnancy as a data set and

obtained different precision scores in the Decision Tree, Random Forest, Bagging Decision Tree and Bagging Random Forest ML algorithm. In the data validation technique, we have used 5-fold cross-validation for data splitting. The accuracy scores of the ML algorithm and bagging techniques are listed in Table 1.0.

Table 1. Accuracy Score of an ML Algorithm and Bagging Ensemble Technique

Sl.	Disease Data Sets	No. of Features	No. Samples	Accuracy score			
				RF	DT	BRF (Ensemble)	BDT (Ensemble)
01	CVD	14	1025	93.10	94.21	99.18	99.26
02	Diabetes	09	768	70.17	70.16	76.57	75.57
03	Liver Disease	10	583	66.10	65.51	71.18	70.66
04	Pregnancy	09	2000	94.22	95.52	99.28	99.47

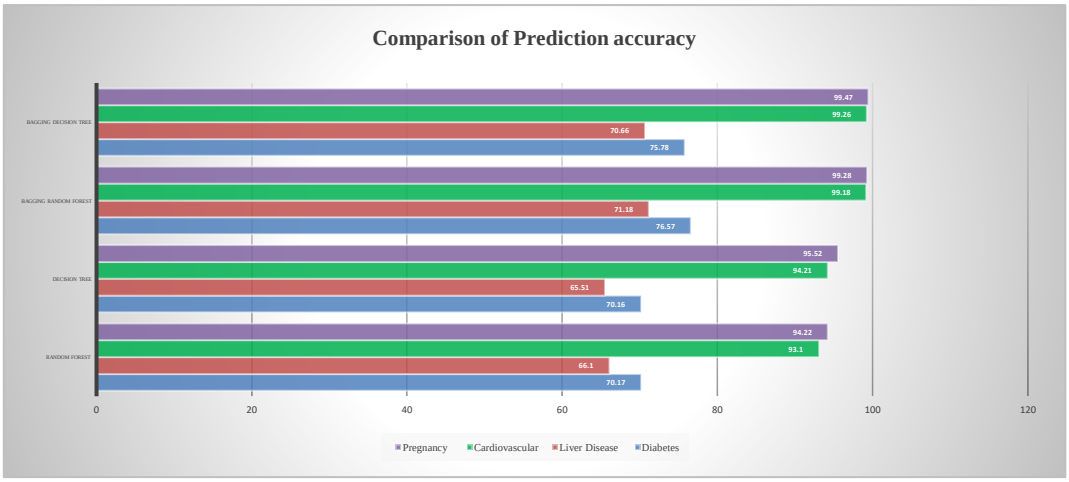

Figure 6. Comparison of bagging ensemble techniques and ML algorithm

Observation: In the proposed model we used the bagging ensemble technique with a random forest and decision tree ML algorithm. We observe that the result and consistency of bagging techniques are much better compared to the ML algorithm. We evaluated the proposed technique on four different disease data sets and the proposed bagging technique provided the best accuracy on these data sets. Graph 1.0 provides a detailed analysis of the ML algorithm and the proposed bagging technique. This graph represents the accuracy score of the ML technique and bagging ensemble technique.

6. Conclusions

The machine learning technique became very popular in medical science for disease prediction and/or diagnosis. For the diagnosis of the disease, it is very important to have an accurate identification and/or prediction of the disease, only then is its correct treatment possible. The problems mentioned above in machine learning models, if they are ignored during machine learning model training, then the ML model may generate wrong decisions or predictions, which is a major threat to human life. Disease detection and/or prediction are the classification problems. There are two possibilities that the disease is present or not. In this situation, the bagging ensemble method provides good results. The bagging method prepares small sub-datasets of the entire dataset, generates predictions from each dataset prepared and gives the final result by generating the mean of all the predictions. The problem of bias and variance is eliminated by generating subdatasets. Machine learning/deep learning is being used in all fields; there is a need for a more comprehensive approach to model training problems such as overfitting, underfitting, and bias-variance trade-offs before machine learning can be used in medical science. So that better and more robust machine learning models can be prepared for disease diagnosis and/or prediction.

References

[1] A. Paullada, I. D. Raji, E. M. Bender, E. Denton, and A. Hanna, " Data and its (dis)contents: A survey of dataset development and use in machine learning research", DOI:https://doi.org/10.1016/j.patter.2021.100336, Patterns open access, Vol. 2, Issue 11, Pp. 1-14, November 2021

[2] O. A. M. López, A. M. López, and J. Crossa, (2022) "Multivariate Statistical Machine Learning Methods for Genomic Prediction", https://doi.org/10.1007/978-3-030-89010-0, Springer Cham Publisher, Biomedical and Life Sciences book chapter - 4, Vol. , Issue, Pp. 109-122, November 2021

[3] A. K. Prajapati and U. K. Singh, "An empirical analysis of ML techniques and/or algorithms for disease diagnosis prediction from the perspective of cardiovascular disease (CVD)," International Journal of Computing Algorithm, vol. 11, no. 2, Pp. 6-16, Dec. 2022, doi: 10.20894/ijcoa.101.011.002.002.

[4] S. Zha F. Luisier, W. Andrews, N. Srivastava, and R. Salakhutdinov, "Exploiting Image-trained CNN Architectures for Unconstrained Video Classification", https://doi.org/10.48550/arXiv.1503.04144, Arxiv publishing, Cornell University, Vol. 3, Issue, Pp. 1-9, May 2015.

[5] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson "CNN Architectures for large-scale audio classification", DOI: 10.1109/ICASSP.2017.7952132, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), June 2017.

[6] E. Sharma, G. Ye, W. Wei, R. Zhao, Y. Tian, J. Wu, L. He, E. Lin and Y. Gong "Adaptation of RNN transducer with text-to-speech technology for keyword spotting", DOI: 10.1109/ICASSP40776.2020.9053191, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), April 2020.

[7] I. Lee, and Y. J. Shin, "Machine learning for enterprises: Applications, algorithm selection, and challenges", https://doi.org/10.1016/j.bushor.2019.10.005, Business Horizons, Elsevier Journal, Vol. 63, Issue 2, Pp. 1-14, April 2020.

[8] T. Yu, and H. Zhu, "Hyper-Parameter Optimization: A Review of Algorithms and Applications", https://doi.org/10.48550/arXiv.2003.05689, Arxiv publishing, Cornell University, Vol. 1, Issue, Pp. 1-56, March 2020.

[9] L. Yang, and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice" https://doi.org/10.1016/j.neucom.2020.07.061, Neurocomputing, Elsevier Journal, Vol. 415, Issue, Pp. 1-22, November 2020.

[10] X. Du, H. Xu, and F. Zhu, "Understanding The Effect of Hyperparameter Optimization on Machine Learning Models for Structure Design Problems", https://doi.org/10.1016/j.cad.2021.103013, Computer-Aided Design, Elsevier Journal, Vol. 135, Issue, Pp. 1-16, June 2021.

[11] Q. Wang, Y. Ma, K. Zhao, and Y. Tian, "A Comprehensive Survey of Loss Functions in Machine Learning". https://doi.org/10.1007/s40745-020-00253-5" Annals of Data Science. Vol., Issue, Pp. 1-26, March 2020

[12] B. U. Bawankar and K. Chinnaiah, "Implementation of ensemble method on dna data using various cross validation techniques", https://doi.org/10.17993/3ctecno.2022.v11n2e42.59-69, 3c Technology innovation glosses applied to SMEs, . Vol 11., Issue 2, Pp. 1-11, December 2022

[13] M. Rafało, "Cross validation methods: Analysis based on diagnostics of thyroid cancer metastasis", https://doi.org/10.1016/j.icte.2021.05.001, ICT Express (ScienceDirect), Vol 8., Issue 2, Pp. 1-6, June 2022

[14] V. Singh, M. Pencina, A. J. Einstein, J. X. Liang, D. S. Berman & P. Slomka "Impact of train/test sample regimen on performance estimate stability of machine learning in cardiovascular imaging" https://doi.org/10.1038/s41598-021-93651-5, www.nature.com/scientificreports, Vol 11., Issue, Pp. 1-8, July 2021

[15] A. Rácz, D. Bajusz and K. Héberger, "Effect of Dataset Size and Train/Test Split Ratios in QSAR/QSPR Multiclass Classification", https://doi.org/10.3390/ molecules26041111, Molecules [MDPI], Vol 26., Issue 4, Pp. 1-16, February 2021

[16] Anil Kumar Prajapati,Umesh Kumar Singh. Cardiovascular disease (CVD) prediction through Artificial Neural network in the perspective of Deep Learning, International Journal of Computing Algorithm, Vol. 11, Issue 2, 2022, pp. 1-7, DOI: 10.20894/IJDMTA.102.011.002.001, ISSN: 2278-2397

[17] L. Li, and M. Spratling "Understanding and combating robust overfitting via input loss landscape analysis and regularization" https://doi.org/10.1016/j.patcog.2022.109229, Pattern Recognition, Vol. 136, Issue, Pp. 1-11, April 2023

[18] G. K. Gupta, and D. K. Sharma, "A Review of Overfitting Solutions in Smart Depression Detection Models" 10.23919/INDIACom54597.2022.9763147, 9th International Conference on Computing for Sustainable Global Development (INDIACom), March 2022

[19] F. Heintz, M. Milano, and B. O'Sullivan "Trustworthy AI – Integrating Learning, Optimization and Reasoning", https://doi.org/10.1007/978-3-030-73959-1, First International Workshop, TAILOR 2020 Virtual Event (Springer Cham), Pp. 31-42, September 2020.

[20] A. D. Gavrilov, A. Jordache, M. Vasdani, and J. Deng "Preventing Model Overfitting and Underfitting in Convolutional Neural Networks", DOI: 10.4018/IJSSCI.2018100102, International Journal of Software Science and Computational Intelligence, Vol. 10, Issue 4, Pp. 1-10, December 2018

[21] T. Aotani, T. Kobayashi, and K. Sugimoto," Meta-Optimization of Bias-Variance Trade-Off in Stochastic Model Learning" 10.1109/ACCESS.2021.3125000, IEEE Acess, Vol. 9, Issue 4, Pp. 1-10, November 2021.

[22] Y. Dar, V. Muthukumar, and R. G. Baraniuk "A Farewell to the Bias-Variance Tradeoff? An Overview of the Theory of Overparameterized Machine Learning" https://doi.org/10.48550/arXiv.2109.02355, Arxiv publishing, Cornell University, Vol. 1, Issue, Pp. 1-48, September 2021.

[23] Y. Muhammad, M. Tahir, M. Hayat and K. Chong, "Early and accurate detection and diagnosis of heart disease using intelligent computational model", https://doi.org/10.1038/s41598-020-76635-9, Published by Scientific Reports, 10, Issue- 4, PP. 1-18, November 2020.

[24] P. Ghosh, S. Azam, M Jonkman, S. Karim, F. M. J. M Shamrat, E. Ignatius, S. Sultana, A. R. Beeravolu, and A. F. De Boer, "Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques", DOI: 10.1109/ACCESS.2021.3053759, IEEE ACCESS, Vol. 09, Issue-, PP. 1-23, February 2021.

[25] M. Ahmad, M. Alfayad, S. Aftab,M. A. Khan, A. Fatima, B. Shoaib, M. Sh. Daoud and N. S Elmitwally "Data and Machine Learning Fusion Architecture for Cardiovascular DiseasePrediction", CMC-Computers, Materials & Continua, doi:10.32604/cmc.2021.019013, Vol. 69, Issue-02, PP. 1-15, April 2021.