JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A REVIEW OF RECENT ADVANCES AND ENSEMBLE METHODS FOR MACHINE LEARNING

J. P. Pramod[1], Kanneboina Vaishnavi Yadav[2] & Bhagavatula Veda Bharati[3]

[1]Asst Professor, Dept of Physics

Stanley College of Engineering and Technology for Women

[2&3]B.Tech Student Dept of Information Technology

Stanley College of Engineering and Technology for Women

ABSTRACT

Researchers around the globe are working on improving the machine learning algorithms for modeling prediction and analytics problems. No single best machine learning algorithm is present which is applicable for all the possible cases of problems. Corporate distress signals are important for both institutions and banks when evaluating firms' performances. This has led to the rise of a paradigm shift in machine learning called federated learning (FL) that allows for decentralized model training over distributed data sources. Cervical cancer is a serious public health issue worldwide, and early identification is crucial for better patient outcomes. Recent study has investigated how ML and DL approaches may be used to increase the accuracy of vagina tests. Stocks in the Chinese stock market can be divided into ST stocks and normal stocks, so to prevent investors from buying potential ST stocks, this paper first performs SMOTEENN oversampling data preprocessing for the ST stock category, and selects 139 financial indicators and technical factor as predictive features. Then, it combines the Boruta algorithm and Copula entropy method for feature selection, effectively improving the machine learning model's performance in ST stock classification, with the AUC values of the two models reaching 98% on the test set.

Keywords: Machine Learning, modeling prediction, Machine Learning Techniques, Federated Learning, analytics problems.

INTRODUCTION:

A software defect is a bug, fault, or error in a program that causes improper outcomes. Software defects are programming errors that may occur because of errors in the source code, requirements, or design. Defects negatively affect software quality and software reliability. Hence, they increase maintenance costs and efforts to resolve them. Software development teams can detect bugs by analyzing software testing results, but it is costly and time-consuming by testing entire software modules. As such, identifying defective modules in early

stages is necessary to aid software testers in detecting modules that required intensive testing. In the field of software engineering, software defect prediction (SDP) in early stages is vital for software reliability and quality. The intention of SDP is to predict defects before software products are released, as detecting bugs after release is an exhausting and time-consuming process. In addition, SDP approaches have been demonstrated to improve software quality, as they help developers predict the most likely defective modules. SDP is considered a significant challenge, so various machine learning algorithms have been used to predict and determine defective modules. The mango, an exotic fruit with a delicious taste and many virtues, is the subject of particular attention throughout its supply chain. From the selection of varieties by producers to the choices made by consumers, precise identification is of vital importance. It guarantees taste quality, authenticity, and a fair price for each variety while meeting the individual expectations of mango lovers. Manual recognition of mango varieties, although common, is a laborious and imperfect process, especially when large volumes are involved. Machine learning offers a revolutionary alternative, providing an automated and robust solution to this crucial task. Based on the analysis of digital images of mangoes, machine learning algorithms learn to identify the distinctive characteristics of each variety, such as shape, color, texture, and size. These models, fed by labeled image datasets, become experts in variety recognition, often surpassing the accuracy of the human eye. Adopting machine learning to recognize mango varieties brings a host of benefits, including increased accuracy and tenfold efficiency.

LITERATURE REVIEW

Faye, D. (2023) The world's agricultural production suffers huge losses estimated between 20% and 40% annually. 40% to 50% of such losses are due to pest and diseases which cause significant economic losses every year. Precise assessment of severity is crucial for suitable management of crop diseases. It helps famers to avoid yield losses, reduce production costs, ensure good disease management and so on. This study is a review of plant diseases severity estimation solutions proposed by researchers the last few years and based on Image Processing Techniques (IPT), classical Machine Learning (ML) and Deep Learning (DL) algorithms. The analysis of these solutions has allowed us to identify their limitations and potential challenges in plant disease severity assessment.

Rahman, M. (2021) Data clustering plays a vital role in object identification. In real life we mainly use the concept in biometric identification and object detection. In this paper we use Fuzzy Weighted Rules, Fuzzy Inference System (FIS), Fuzzy C-Mean clustering (FCM), Support Vector Machine (SVM) and Artificial Neural Network (ANN) to distinguish three types of Iris data called Iris-Setosa, Iris-Versicolor and Iris-Virginica. Each class in the data table is identified by four-dimensional vector, where vectors are used as the input variable called: Sepal Length (SL), Sepal Width (SW), Petal Length (PL) and Petal Width (PW). The combination of five machine learning methods provides above 98% accuracy of class identification.

Bkheet, S. and Agbinya, J. (2021) The Internet of Things (IOT) is a recent technology originating from the field of sensor networks. It has received significant attention because it is involved in most aspects of our daily lives. The IOT vision makes objects of various kinds become part of the Internet by assigning each object a unique identifier, enabling objects to communicate with each other in the same or different environments. IOT can collect, process, and exchange data via a data communication network. There are

many methods for identifying objects; some have existed since the beginning of IOT innovation, such as Radio Frequency Identification (RFID), Barcode/2D code, IP address, Electronic Product Codes (EPC), etc. Continuous development in IOT domain and the large number of objects connected to the Internet daily require an improved identification method to cope with the rapid development in this field.

Kalathas, I. , (2021) The extraction of useful information supports the process of making business decisions. In every mechanical process, application or service, the periodic maintenance of the necessary equipment is an expensive process and therefore the technicians and the supervisors have the responsibility of the proper decision making. At the railway companies, a huge amount of data is produced which, with the appropriate processing and smart business systems, can attribute quality information and knowledge. In this study, the benefits of the business intelligence are presented with the techniques of machine learning and data mining involved of the Greek railway companies, which use obsolete procedures of maintenance. In addition, a study of the present situation is held as well as a record of the needs and requirements of the railway companies. At the same time, tools of machine learning and data mining are examined that can assist on the creation of a new strategic support of decisions for the development of the predictive maintenance of the Greek railways making a new complete system of business intelligence.

Alegado, R. and Tumibay, G. (2020) This study aimed to find a suitable model for forecasting the appropriate stock of vaccines to avoid shortage and over-supply. The Auto-Regressive Integrated Moving Average (ARIMA) and Multilayer Perceptron Neural Network (MLPNN) models were used for forecasting time series data. The monthly vaccination coverage was used to develop the models from January 2014 until December 2019. The dataset consists of 72 months of observation, the 60 months of data are used for model fitting from January 2014 to December 2019, and the remaining 12 months of data from January 2019 to December 2019 are used to test the accuracy of the forecast. The most suitable forecast model was selected based on the lowest Root Mean Square Error (RMSE) value and the Mean Absolute Error (MAE). The analytical result shows that the MLPNN model outperformed the ARIMA model in forecasting monthly demand for vaccines. The results will help policymakers improve the proper use of vaccination resources.

Inside Ensemble Methods for Machine Learning you will find:

- Methods for classification, regression, and recommendations
- Sophisticated off-the-shelf ensemble implementations
- Random forests, boosting, and gradient boosting
- Feature engineering and ensemble diversity
- Interpretability and explainability for ensemble methods

Ensemble machine learning trains a diverse group of machine learning models to work together, aggregating their output to deliver richer results than a single model. Now in Ensemble Methods for Machine Learning you'll discover core ensemble methods that have proven records in both data science competitions and real-world applications. Hands-on case studies show you how each algorithm works in production. By the time you're done, you'll know the benefits, limitations, and practical methods of applying ensemble machine learning to real-world data, and be ready to build more explainable ML systems**.**

The bayesian machine learning advanced forecast ensemble (bamcafe) framework

The ensemble forecast is a popular prediction approach for turbulent systems that employ a collection, known as the "ensemble," of multiple individual forecasts from a parametric model. A skillful ensemble forecast requires an accurate representation of the underlying dynamics as well as a reliable forecast initialization. DA is used to generate a more accurate initialization by combining partial and noisy observations with the given imperfect model. However, the forecast error can grow up quickly as time evolves due to the model bias, which often results in an inaccurate quantification of the forecast uncertainty as well.

In the Bayesian Machine Learning Advanced Forecast Ensemble (BAMCAFE) framework, the intrinsic error in the imperfect physics-based forecast model is alleviated by training an ML model (e.g., a NN) based on a set of assimilated trajectories, which are obtained by applying a Bayesian sampling method (i.e., a Bayesian ensemble DA) to the imperfect physics-based model with the help from the available partial and noisy observational time series. Combining the information from both the imperfect model and the noisy observations, the assimilated trajectories achieve trajectory-wise improvement compared with the signals generated from the imperfect model in terms of both the dynamical and statistics features. Specifically, the BAMCAFE involves the following four steps:

1. generating the ML training data using a Bayesian sampling approach,
2. training an ML model (e.g., a NN) utilizing the training data from step 1,
3. employing a generalized DA for the initialization of the ML model,
4. applying an ML ensemble forecast.

Theoretical and advanced machine learning with TensorFlow

Before starting on the learning materials below, be sure to:

1. Complete our curriculum Basics of machine learning with TensorFlow, or have equivalent knowledge
2. Have software development experience, particularly in Python

This curriculum is a starting point for people who would like to:

1. Improve their understanding of ML
2. Begin understanding and implementing papers with TensorFlow

   You should already have background knowledge of how ML works or completed the learning materials in the beginner curriculum Basics of machine learning withTensorFlow before continuing.

Deep Learning and Neural Networks

Deep learning is a subset of machine learning that uses neural networks with many layers to analyze various forms of data. These advanced algorithms uncover patterns that traditional algorithms might miss and excel at processing and making sense of enormous datasets. Deep learning is behind advancements in areas such as image recognition and natural language processing (NLP). Two advancements in deep learning include convolutional neural networks (CNNs) and recurrent neural networks (RNNs). CNNs can easily parse visual information, so they're widely used in image recognition systems. They simulate the way the human brain processes information by breaking down images into components and analyzing them layer by layer to identify patterns and features. RNNs are designed to understand sequential data, so they're ideally suited for NLP tasks. They can remember previous inputs in the data sequence, which allows them to use predictive

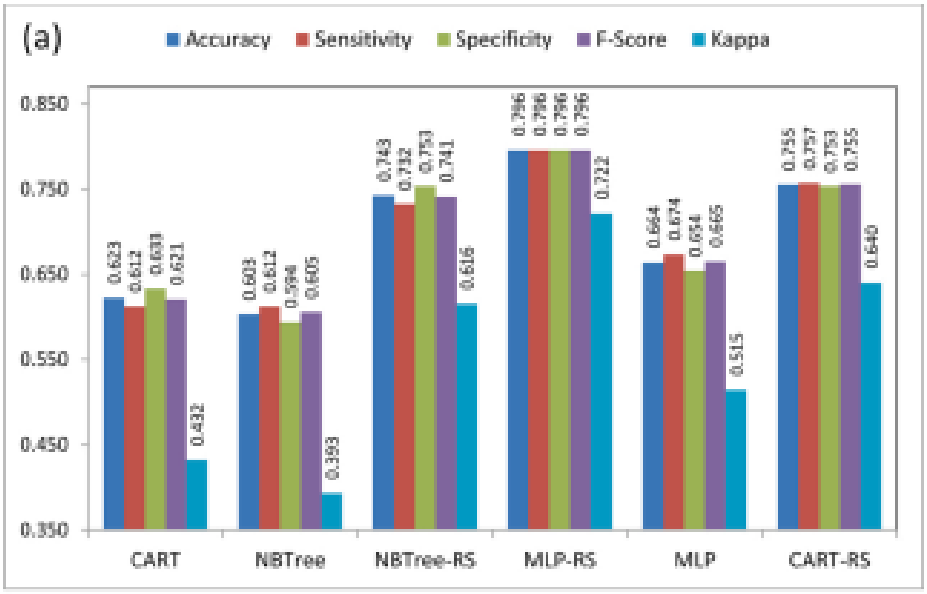
analytics to produce contextually informed text—a critical element in speech recognition and the generation of human language.

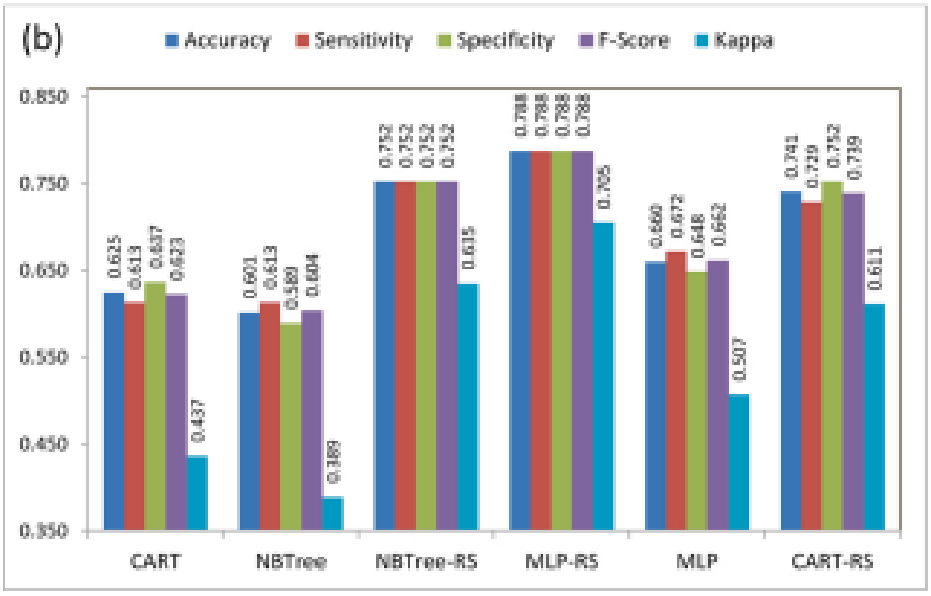
RESEARCH METHODOLOGY

The evaluation used different measures like simple averaging and winner-takes-all measures on classification tasks and simple average combination methods for regression problems. Here, counting is done based on regression-based ensemble learning from a pool of convolutional feature mapped weak regressors. More specifically, the combination of multiple ML models with similar performance on a predictive task can eventually lead to models with higher accuracy and fewer errors. To obtain the data needed to carry out the review, the Scopus and PubMed online electronic databases were searched to return the relevant literature. The search process is detailed. In recent years, ensemble learning models are becoming increasingly popular among researchers in the field of predictive modeling leading to an overall improved classification performance. For the experiments, 10 well-known software defect datasets were selected. The majority of related works used these datasets to evaluate the performance of their SDP techniques and this is the reason behind selecting the above-mentioned dataset for further comparisons. The main idea behind this is to introduce the NCL concept in deep architectures. Robust regression via deep NCL is an extension of in which theoretical insights about the Rademacher complexity are given and extended to more regression-based problems.

RESULTS AND DISCUSSIONS

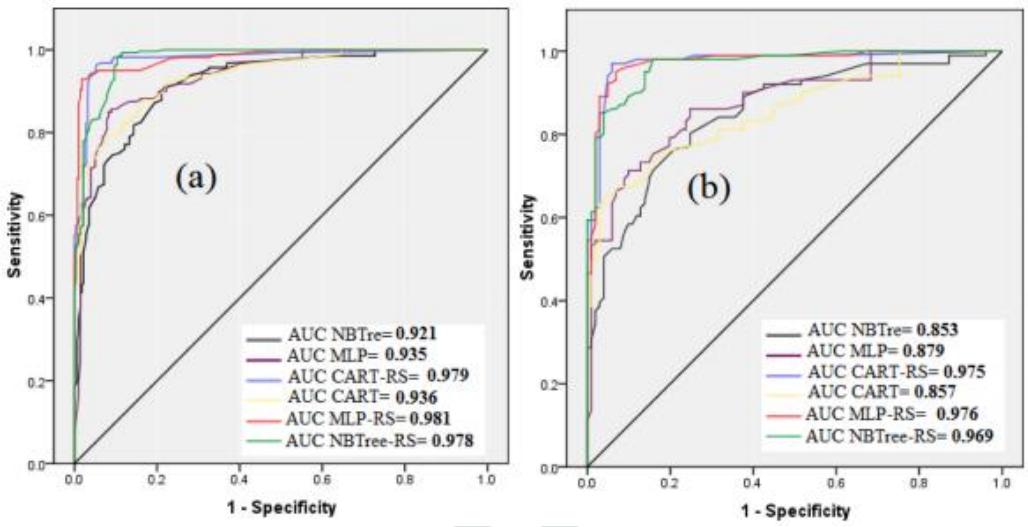
The GWP maps produced from the ensemble models were compared to those generated by the three standalone models, using the ROC curve, SST, SPF, F-score, accuracy (ACC), and kappa (K; Graph. 1a and 1b). The models were evaluated and validated using the confusion matrix for the validation (30 % of all data) and training (70 % of all data) datasets of spring locations. The results indicate that the MLP-RS model (Graph. 2a and b) had an AUC-value of 0.981 based on the training dataset, and an AUC-value of 0.976 based on the validation dataset. This model appears to outperform the other models when producing GPMs. Based on the other five statistics, the MLP-RS ensemble made the best prediction of GWP [Graph1 a and b; (training, validation)],





Graph 1: Validation of results, using (a) training datasets, and (b) validation datasets

The other standalone models achieved acceptable GPM results as well; MLP [see Graph. 2a and b] the models. Furthermore, the Wilcoxon signed-rank test was used to evaluate pair wise algorithms. This test uses p-value and z-value parameters such that when $p < 0.05$ and $-1.96 < z < 1.96$, this suggests that the efficiency of the models in preparing the GWP maps differed substantially.



Graph 2: models, using (a) training datasets and (b) validation datasets

Our test findings, Supplementary, explicitly determine that the efficiency of all GWP models had statistically significant pair wise.

CONCLUSIONS

All six ML models in this study achieved good accuracy (AUC > 0.85); however, there were differences in their performances when using different validation methods. MLP-RS had the best accuracy the future potential of ML applications is promising, not only in improving model prediction accuracies, but in aiding in the understanding of the underlying physiological changes within an athlete's heart. To this end, this paper includes a related works section where several studies about network traffic monitoring and classification were presented. These studies refer to contemporary solutions machine learning techniques with both wavelet analysis and ensemble techniques, namely the bootstrap and boosting techniques, to predict drought. This work proposed to evaluate the performance of two machine learning (ML) dynamic ensemble methods, using

wind speed and solar irradiance data separately as inputs. Machine learning has great potential for improving estimates of future suicidal behaviour and monitoring changes in risk over time. Further research can address important challenges and potential opportunities that may contribute to significant advances in suicide prediction. To overcome the challenges of CNS drug discovery, researchers have utilized AI/ML-based methods.

REFERENCES

1. S. Dutta, [2010], "Machine Learning Algorithms and Their Application to Ore Reserve Estimation of Sparse and Imprecise Data," Journal of Intelligent Learning Systems and Applications, Vol. 2 No. 2, 2010, pp. 86-96. doi: 10.4236/jilsa.2010.22012.
2. Rahman, M. (2021), "Data Classification Using Combination of Five Machine Learning Techniques", Journal of Computer and Communications, vol.9, pages.48-62. doi: 10.4236/jcc.2021.912004.
3. Faye, D. (2023), "Plant Disease Severity Assessment Based on Machine Learning and Deep Learning: A Survey", Journal of Computer and Communications, vol.11, pages.57-75. doi: 10.4236/jcc.2023.119004.
4. Ryu, S.. (2018) "A Comparative Study of Machine Learning Algorithms and Their Ensembles for Botnet Detection", Journal of Computer and Communications, vol.6, pages.119-129. doi: 10.4236/jcc.2018.65010.
5. Yang, Y. and Ma, G. (2010), "Ensemble-based active learning for class imbalance problem", Journal of Biomedical Science and Engineering, vol.3, pages.1022-1029. doi: 10.4236/jbise.2010.310133.
6. Bkheet, S. and Agbinya, J. (2021), "A Review of Identity Methods of Internet of Things (IOT)", Advances in Internet of Things, vol.11, pages.153-174. doi: 10.4236/ait.2021.114011.
7. S. Kalhori and X. Zeng, [2013], "Evaluation and Comparison of Different Machine Learning Methods to Predict Outcome of Tuberculosis Treatment Course," Journal ofIntelligent Learning Systems and Applications, Vol. 5 No. 3, pp. 184-193. doi: 10.4236/jilsa.2013.53020.
8. .Kalathas, I., (2021), "Business Intelligence and Machine Learning Methods for Predictive Maintenance in Greek railways", Open Journal of Applied Sciences, vol.11, pages.20-35. doi: 10.4236/ojapps.2021.111A003.
9. Alegado, R. and Tumibay, G. (2020), "Statistical and Machine Learning Methods for Vaccine Demand Forecasting: A Comparative Analysis", Journal of Computer and Communications, vol.8, pages.37-49. doi: 10.4236/jcc.2 020.810005.
10. Liu, S. (2024), "An Application of Machine Learning to Thalassemia Diagnosis", Journal of Computer and Communications, vol.12, pages.211-230. doi: 10.4236/jcc.2024.122013.