A COMPARATIVE STUDY OF FEATURE SELECTION TECHNIQUES TO IMPROVE MACHINE LEARNING-BASED SOFTWARE FAULT PREDICTION

TIRUGULLA NEELIMA
Research Scholar
IEC University-H.P

Dr. PRASADU PEDDI
Research Supervisor
IEC University-H.P

ABSTRACT

This study presents a systematic literature review (SLR) that investigates recent advancements in Software Fault Prediction (SFP) methodologies. The quality of a fault prediction model depends on the software metrics that are used to build the prediction model. Feature selection represents a process of selecting a subset of relevant features that may lead to build improved prediction models. Feature selection methods, particularly wrapper techniques, are often employed to improve predictive accuracy. Evaluation of models predominantly relies on confusion matrix-based metrics such as Accuracy, Precision, Recall, and F1-Score. The review focuses on key dimensions including techniques, datasets, feature selection methods, software metrics, and evaluation criteria. Findings reveal that machine learning approaches— particularly neural networks, deep learning, and ensemble methods— are increasingly employed due to their capability to manage the complexity of software fault data. By analyzing significant studies from renowned digital libraries such as ACM, IEEE, Springer Link, and Science Direct, five research questions were defined to guide the assessment of current trends in SFP research. Public datasets, notably those from the PROMISE and NASA MDP repositories, are widely utilized, underlining the importance of dataset diversity for enhancing model performance.

Keywords: *systematic literature review (SLR), Software Fault Prediction (SFP), prediction models, datasets, feature selection methods.*

INTRODUCTION

Software systems, however, are prone to errors that may result in negative outcomes such system failures, data loss, security breaches, and monetary losses. Early detection of software system defects may improve software quality and save overall expenses. SFP is a technique for estimating and predicting the probability of software system defects. Today, a wide variety of SFP approaches are accessible, each with unique advantages and disadvantages. Machine learning, hybrid, and statistical techniques are some of these approaches. SFP has two subcategories: static analysis and dynamic analysis. Static analysis looks at the software code without actually running it, and dynamic analysis looks at the program while it is operating. Code inspection, code review, and code analysis are examples of static analysis methods; testing, debugging, and profiling are examples of dynamic analysis approaches. The complexity and diversity of software systems make SFP a challenging task. Software systems are composed of many layers and components, some of which may interact in complex ways. Furthermore, it might be challenging to adequately portray the general behaviour of software systems since they may display different features depending on the situation. Many SFP techniques are now available, including as hybrid methodologies, machine learning algorithms, and statistical methodology. Each strategy's benefits and drawbacks determine when and where it may be used to various test scenarios.

LITERATURE REVIEW

Asgarov, E. (2024) One of the most significant issues facing the globe today is heart disease. It is an ongoing issue that is a contributing factor to the world's top cause of mortality. Early heart disease identification is essential to resolving this

problem. This project's main objective is to ascertain if heart disease may be predicted using supervised machine learning methods. This research carefully reviewed 30 articles published between 1997 and 2023 that discussed machine learning methods for cardiac disease prediction. The issue is that various authors train and assess these models using different data sets and varying quantities of parameters. The accuracy of the model may be impacted by these two variables. To avoid bias, I only utilized studies that examine many approaches using the same data when comparing models. However, with a minimal accuracy rate of 88%, hybrid models provide dependable accuracy, indicating that they could be a more successful method of predicting cardiac disease.

Haque, F. M. A. (2024) Through commercial and consumer loans, banks play a vital role in economic growth in any financial ecosystem. To lower the likelihood of default, banks must assess the applicant's financial situation in relation to the loan risks. Recently, a lot of companies have improved their decision-making processes by using data analytic and contemporary technology. A predictive modelling strategy that uses machine learning techniques prescribes the chance of return. The most effective algorithm among them was Ada Boosting, which had an astounding accuracy rate of 99.99%. Consequently, the results demonstrate the effectiveness of ensemble learning in enhancing the ability to forecast loan acceptance choices.

Chana, A., (2023) This study aims to help farmers use IoT and AI (smart farming) to make better informed crop choices by considering soil fertility and weather forecasts. Based on this procedure, the Internet of Things using the Message Queuing Telemetry Transport (MQTT) protocol, weather forecast APIs for crop prediction and recommendations, and a machine learning approach based on Random Forest were used. The prototype forecasts inputs including temperature, humidity, pH, phosphorus, potassium, and nitrogen using Internet of Things sensors and the weather API. The Yaounde area of Cameroon served as the testing ground for the technology. This project's examination of projected climate factors, such as temperature, wind, and rainfall, produced better crop selection recommendations. The Internet of Things system may be accessed at any time and from any location using a web browser.

Saha, V. (2023) According to there has been a discernible increase in interest in bitcoin price prediction as a consequence of the rising significance of digital assets in the financial sector. This essay looks closely at machine learning techniques for bitcoin price prediction. It makes use of historical data from several bitcoin exchanges that are openly accessible. In order to handle missing data and maintain the data set's completeness and dependability, interpolation methods are used. Four technical indicators have been chosen as prediction characteristics. The findings highlight the advantages and disadvantages of the various strategies as well as the importance of feature engineering and algorithm selection in producing precise predictions of the price of bitcoin. Traders and investors may find the study's analytical data on the dynamic and sometimes moving sector of cryptocurrency price predictions helpful in their decision-making, despite the difficulties presented by the bitcoin market.

Ojo, S. (2022) For precise signal delivery in wireless channels, path loss prediction methods are essential. Route loss has not been well predicted by either deterministic or empirical models. Because machine learning techniques provide a flexible network design and make it possible to handle enormous amounts of data, we used them in this research to forecast route loss. To forecast route loss in the scenarios under investigation, we used radial basis function (RBF) and support vector regression (SVR) models. Numerous input parameters may be handled by the SVR model without adding complexity to the network design. For its part, the RBF offers a fair approximation of the function. Additionally, we compared the two machine learning approaches using the Cost-231 W-I, SUI, Egli, Frees-pace, and Cost-231 models. Path loss was exaggerated in the analytical models. Overall, the machine learning models outperformed the empirical models in evaluating route loss.

Machine Learning

The study of algorithms and methods that enable computers to automatically "learn" from experiences is known as machine learning. Machine learning incorporates ideas and methods from a wide range of disciplines, including as artificial intelligence, statistics, information theory, biology, philosophy, and cognitive psychology. The two main categories into which learning may be separated are deductive and inductive learning. Algorithms for inductive machine learning extrapolate or uncover previously unidentified patterns and rules from data samples. Conversely, deductive learning builds on previously taught material to infer new information.

Supervised Machine Learning

The training data's class labels are preset in supervised learning. Pairs of an input item and its anticipated outcome, such the class label, serve as the representation for the training examples. To predict the classes of fresh data, a supervised learner has to identify a function that roughly represents the mapping between training data and their classes. For supervised learning, several methods have been developed, including support vector machines (SVMs), random forests, K-nearest neighbour, decision trees, naïve Bayes classifiers, artificial neural networks, and others.

Unsupervised Machine Learning

Unsupervised learning differs from supervised learning due to the absence of easily available class labels for training data sets. Whether an item belongs in a single class is determined via unsupervised learning approaches. In other words, pupils independently study the material. Knearest Neighbour, self-organism maps (SOMs), and data clustering techniques (such fuzzy c-means clustering and K-means clustering) are often used for unsupervised learning applications. Since it significantly affects the quality of the taught model, a correct representation of the input item is essential. In order to characterize the entity, the input object is often transformed into a vector of attributes or characteristics.

Feature Selection

In machine learning, feature selection—also referred to as variable selection or subset selection—is a commonly used strategy to address the high dimension issue. For a more straightforward and condensed data representation, it chooses a subset of significant characteristics and eliminates extraneous, redundant, and noisy information. Feature selection has a number of advantages. First, by eliminating unnecessary and duplicate components,

feature selection significantly reduces the amount of time needed for a learning process. Second, learning algorithms may concentrate on the most crucial elements of the data and generate simpler but more accurate data models when unnecessary, redundant, and noisy information is removed from the equation. The categorization performance is enhanced as a result. Third, feature selection might improve our understanding of the task's central idea and assist us in creating a more straightforward and thorough model.

The process of feature extraction, also known as feature transformation, which blends the original features to produce new features, is distinct from feature selection. Techniques for feature modification include principal component analysis (PCA), locally linear embedding (LLE), and linear discriminant analysis (LDA). However, feature selection preserves the original meanings of the chosen features, which is advantageous in a variety of fields.

Random Feature Selection

The selection of random attributes is the third Random Forest method premise. One technique to stop decision trees from over fitting is the random feature selection method, often referred to as feature bagging or the random subspace approach. This is accomplished by reducing the "weight" of traits that are limited yet discernible. One kind of ensemble learning is random feature selection, where the main goal is to increase the accuracy of the results by combining many models rather than depending just on one. Random feature selection and bagging are comparable, especially in that they may both be used to tiny training sample numbers. The classifiers are built in random sub-spaces of the data feature space in the random subspace approach. In order for the model to categorize the input, these classifiers are often included into the final decision rule by a straightforward majority voting procedure.

Support Vector Machine

The most advanced form of the Support Vector Machine technique, as previously stated, makes use of an algorithm that maximizes a certain mathematical function with respect to a particular set of data. Like the Random Forest method, the Support Vector Machine algorithm may be understood via a set of fundamental ideas that provide its theoretical basis.

Ensemble Learning

The fourth Random Forest fundamental is ensemble learning. Based on characteristics obtained from several data projections, ensemble learning makes early, imprecise predictions using a range of machine learning methods. These results are then aggregated using various voting techniques to get better results than any of the individual algorithms. As part of ensemble learning, Random Forests builds many smaller decision trees using subsets of the dataset's attributes. Based on the forecasts of each tree, a final decision is then made by a majority vote. Ensemble learning increases classification accuracy, reduces the chance of over fitting and under-fitting, and enhances performance.

RESEARCH METHODOLOGY

Data sets may be assessed or discussed according to their "features." Considerations may include size, location, age, time, colour, and other elements. Features are also known as attributes, variables, fields, and characteristics, and they show up as columns in datasets. One definition of a data feature in machine learning is a quantifiable element or component of an observed phenomenon. This research makes use of three datasets.

In order to evaluate the effectiveness of the suggested approaches, a number of experiments are developed and carried out using datasets from actual software projects, such as NASA software projects, which are often used by academics looking at software defect prediction. NASA collects its datasets via the publicly available PROMISE repository. NASA Datasets PC1, PC2, KC1, and KC3 are used. These datasets include Boolean variables that indicate if a module is prone to issues, as well as software metrics like Halstead and McCabe metrics. Eclipse is an open-source Java project that offers tools for software development, deployment, and life-cycle management in addition to configurable frameworks and run-times. The fact that Eclipse is an open-source system does not restrict the study's conclusions to just open-source platforms since it employs a different open-source paradigm than Linux. The Eclipse project is centrally organized by a group of experts from IBM Corporation.

RESULTS AND DISCUSSIONS

The accuracy findings in Table 1 show that Ada Boost and GBM were 86.69% and 86.56% less accurate than XG Boost, respectively. On the F-measure, XG Boost scored 0.92, whereas Ada Boost and GBM scored 0.84 and 0.83, respectively. This suggests that the ensemble approaches outperformed the individual classifiers. While Ada-Boost, GBM, XG Boost, and J48 have outstanding accuracy (0.85), all three ensemble classifiers have outstanding recall (0.87).

Table 1: Performance Comparison for the Eclipse 2.0 dataset

Eclipse 2.0					
Classifier Type		**Precision**	**Recall**	**F-Measure**	**Accuracy**
State-of Art	**Naïve Bayes**	0.84	0.76	0.79	76.68
	Decision Table	0.82	0.85	0.81	85.80
	J48	0.85	0.86	0.82	85.51
	Random Tree	0.78	0.83	0.80	83.23
Ensemble Learning	**AdaBoost**	0.85	0.87	0.83	86.69
	GBM	0.85	0.87	0.84	86.56
	XGBoost	0.85	0.87	0.92	**87.11**

The accuracy results in Table 2 showed that XG Boost did the best (87.19%), followed by Ada Boost and GBM (89.17% and 89.09%, respectively). In the ensemble technique, Ada Boost outperformed the other classifiers with a F measure, recall, and accuracy score of 0.89.

Table 2: Performance Comparison for the Eclipse 2.1 dataset

Eclipse 2.1					
Classifier Type	**Algorithm**	**Precision**	**Recall**	**F-Measure**	**Accuracy**
State-of Art	**Naïve Bayes**	0.85	0.78	0.81	78.11
	Decision Table	0.87	0.89	0.85	88.90
	J48	0.78	0.88	0.83	88.32
	Random Tree	0.81	0.87	0.84	87.15
Ensemble Le	**AdaBoost**	0.89	0.89	0.89	89.17
	GBM	0.87	0.89	0.85	89.09
	XGBoos	0.85	0.89	0.84	**89.1**

ar ni ng	**t**				**9**

According to Table 3, XG Boost has the highest accuracy score (85.98), followed by Ada Boost (85.23) and GBM (85.24). In terms of accuracy and recall, the ensemble methods outperformed the individual classifiers. However, the Decision Table classifier produced the greatest results when the F-measure was 0.85.

Table 3: Performance Comparison for the Eclipse 3.0 dataset

Eclipse 2.1					
Classifier Type	**Algorithm**	**Precision**	**Recall**	**F-Measure**	**Accuracy**
State-of Art	**Naïve Bayes**	0.85	0.78	0.81	78.11
	Decision Table	0.87	0.89	0.85	88.90
	J48	0.78	0.88	0.83	88.32
	Random Tree	0.81	0.87	0.84	87.15
Ensemble Learning	**AdaBoost**	0.89	0.89	0.89	89.17
	GBM	0.87	0.89	0.85	89.09
	XGBoost	0.85	0.89	0.84	**89.19**

Even though FNR and FPR are lower in all three ensemble classifiers, Table 4 demonstrates that TPR and TNR are greater. Eclipse 2.0's XG Boost ensemble classifier has superior TPR and TNR but poor FNR and FPR in comparison to earlier classifiers. In Eclipse 2.1, XG Boost and Decision Tree both exhibit low FPR and high TNR. The low FNR and high TNR of XG Boost in Eclipse 3.0 are comparable to those of Naive Bayes. The confusion matrix study shows that the ensemble classifier XG Boost performs better than the others in the majority of cases.

Table 4: Confusion Matrix Analysis for Eclipse dataset

Eclipse 2.1					
Classifier Type	**Algorithm**	**Precision**	**Recall**	**F-Measure**	**Accuracy**
State-of Art	**Naïve Bayes**	0.85	0.78	0.81	78.11
	Decision Table	0.87	0.89	0.85	88.90
	J48	0.78	0.88	0.83	88.32
	Random Tree	0.81	0.87	0.84	87.15
Ensemble Learning	**AdaBoost**	0.89	0.89	0.89	89.17
	GBM	0.87	0.89	0.85	89.09
	XGBoost	0.85	0.89	0.84	**89.19**

In terms of precision, recall, accuracy, and f-measure, the ensemble classifier's performance is contrasted with that of the most advanced classifiers.

CONCLUSIONS

A comparison between ensemble classifiers and the most advanced classifiers in terms of software error prediction has been attempted. The experiments for this topic are carried out in a Google Collaborators notebook using Python 3.6.9. We have compared Ada-boost, GBM, XG Boost, J48, Random Tree, Naive Bayes, and Decision Tree. Performance is often assessed using four metrics: F1 scores, recall, accuracy, and precision. An Eclipse bug data collection that was made publicly accessible was used for the exploratory

investigation. The features were selected using the Random Forest feature selection procedure. When compared to other classifiers, analysis revealed that ensemble techniques may improve SFP's classification performance, with the XG Boost method often yielding the best results. An attempt is made to build a functional deep neural network model, taking into account parameters like the number of hidden layers and the number of nodes in each layer, as well as training features like learning rate and regularization techniques (like L2 Regularization and Dropout Regularization). Furthermore, an attempt has been made to show how important hyper-parameter tweaking is to the creation of effective deep neural network models. The primary goal of this research is to compare the results with those of other machine learning methods and improve the parameters, including the number of hidden layers and nodes in each layer, as well as the training elements, including regularization procedures and learning rate.

REFERENCES

1. *Asgarov, E. (2024), "A Comprehensive Analysis of Machine Learning Techniques for Heart Disease Prediction." Open Access Library Journal, Vol.11 No.4, pages.1-17. ISSN: 2333-9721.*
2. *Haque, F. M. A., (2024), "Bank Loan Prediction Using Machine Learning Techniques." American Journal of Industrial and Business Management, Vol.14 No.12, pages.1690-1711. ISSN: 2164-5175.*
3. *Chana, A (2023), "Real-Time Crop Prediction Based on Soil Fertility and Weather Forecast Using IoT and a Machine Learning Algorithm." Agricultural Sciences, Vol.14, No.5, pages. 645-664, ISSN: 2156-8561.*
4. *Saha, V. (2023), "Predicting Future Crypto currency Prices Using Machine Learning Algorithms." Journal of Data Analysis and Information Processing, Vol.11, No.4, pages. 400-419. ISSN: 2327-7203.*
5. *Ojo, S. (2022), "Path Loss Modeling: A Machine Learning Based Approach Using Support Vector Regression and Radial Basis Function Models." Open Journal of Applied Sciences, Vol.12, No.6, pages. 990-1010. ISSN: 2165-3925.*
6. *Zhao, S. (2021), "Prediction of Protein Expression and Growth Rates by Supervised Machine Learning." Natural Science, Vol.13, No.8, pages. 301-330. ISSN: 2150-4105.*
7. *Zheng, H. (2018), "Analysis of Global Warming Using Machine Learning." Computational Water, Energy, and Environmental Engineering, Vol.7, No.3, pages. 127-141. ISSN: 2168-1570.*
8. *Zhu, Y (2015), "Chemometric Feature Selection and Classification of Ganoderma lucidum Spores and Fruiting Body Using ATR-FTIR Spectroscopy", American Journal of Analytical Chemistry, Vol.6, No.10, pages. 830-840. ISSN: 2156-8278.*
9. *Zhu, Z. (2024), "Stock Type Prediction Based on Multiple Machine Learning Methods." Journal of Intelligent Learning Systems and Applications, Vol.16 No.3, pages. 242-261. ISSN: 2150-8410.*
10. *Zhuang, Z. (2023), "Decipher Clinical and Genetic Underpins of Breast Cancer Survival with Machine Learning Methods." Advances in Breast Cancer Research, Vol.12 No.4, pages. 163-185. ISSN: 2168-1597.*