IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 27 April 2023, accepted 6 June 2023, date of publication 9 June 2023, date of current version 19 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3284975

RESEARCH ARTICLE

BukaGini: A Stability-Aware Gini Index Feature Selection Algorithm for Robust Model Performance

MOHAMED ALY BOUKE[1], (Member, IEEE), AZIZOL ABDULLAH[1], JAROSLAV FRNDA[2,3], (Senior Member, IEEE), KORHAN CENGIZ[4,5], (Senior Member, IEEE), AND BASHIR SALAH[6]

[1]Department of Communication Technology and Network, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Serdang 43400, Malaysia
[2]Department of Quantitative Methods and Economic Informatics, Faculty of Operation and Economics of Transport and Communications, University of Žilina, 01026 Žilina, Slovakia
[3]Department of Telecommunications, Faculty of Electrical Engineering and Computer Science, VSB—Technical University of Ostrava, 70800 Ostrava, Czech Republic
[4]Department of Computer Engineering, Istinye University, 34010 Istanbul, Turkey
[5]Department of Information Technologies, Faculty of Informatics and Management, University of Hradec Králové, 500 03 Hradec Králové, Czech Republic
[6]Department of Industrial Engineering, College of Engineering King Saud University, Riyadh 11421, Saudi Arabia

Corresponding author: Mohamed Aly Bouke (bouke@ieee.org)

This study received funding from King Saud University, Saudi Arabia through researchers supporting project number (RSP2023R145). And was supported by the Ministry of Education, Youth and Sports of the Czech Republic under the grant SP2023/007 conducted by VSB - Technical University of Ostrava, Czechia, and partially supported by Institutional research of the Faculty of Operation and Economics of Transport and Communications–University of Zilina, no. 2/KKMHI/2022. Acknowledgements: The authors would like to thank King Saud University, Riyadh, Saudi Arabia, with researchers supporting project number RSP2023R145.

ABSTRACT Feature interaction is a vital aspect of Machine Learning (ML) algorithms, and gaining a deep understanding of these interactions can significantly enhance model performance. This paper introduces the BukaGini algorithm, an innovative and robust approach for feature interaction analysis that capitalizes on the Gini impurity index. By exploiting the unique properties of the BukaGini index, our proposed algorithm effectively captures both linear and nonlinear feature interactions, providing a richer and more comprehensive representation of the underlying data. We thoroughly evaluate the BukaGini algorithm against traditional Gini index-based methods on various real-world datasets. These datasets include the High School Students' Performance (HSSP) dataset, which examines factors affecting student performance; Cancer Data, which focuses on identifying cancer types based on gene expression; Spambase, which targets spam email classification; and the UNSW-NB15 dataset, which addresses network intrusion detection. Our experimental results demonstrate that the BukaGini algorithm consistently outperforms traditional Gini index-based methods in terms of accuracy. Across the tested datasets, the BukaGini algorithm achieves improvements ranging from 0.32% to 2.50%, underscoring its effectiveness in handling diverse data types and problem domains. This performance gain highlights the potential of the BukaGini algorithm as a valuable tool for feature interaction analysis in various ML applications.

INDEX TERMS BukaGini algorithm, Gini index, ensemble learning, feature interaction analysis, data mining.

I. INTRODUCTION

The rapid growth of data in various domains, including finance, healthcare, social media, and IoT, has led to an increased demand for efficient and effective methods to process large-scale datasets. ML techniques are essential in this context, providing powerful tools to extract valuable insights from complex, high-dimensional data [1]. However, high dimensionality poses numerous challenges, such as increased computational complexity, overfitting, and reduced

The associate editor coordinating the review of this manuscript and approving it for publication was Yongming Li.

interpretability of the resulting models. To address these issues, feature selection techniques have been developed to identify the most relevant variables, reduce dimensionality, and enhance model performance [2].

Feature selection can be viewed as a search problem to find the optimal subset of features that maximizes an ML model's performance [3]. The search space is typically defined by the power set of all available features, resulting in a combinatorial problem that grows exponentially with the number of features. As exhaustive search becomes computationally infeasible, various search strategies have been proposed to navigate the search space more efficiently. Feature selection methods can be broadly categorized into filter, wrapper, and embedded approaches [4].

Ensemble learning, a popular ML technique, combines multiple models to improve overall performance and robustness. It has been successfully employed in numerous applications, such as classification, regression, and clustering. In ensemble learning, the combination of base models or learners can exploit the complementary strengths of each model to produce an ensemble model with better predictive performance than its constituents [5].

Feature interaction analysis is essential to understanding complex relationships between the feature in datasets. It helps reveal how features interact and influence the target variable, aiding in identifying significant feature interactions that can improve model performance. Feature interaction analysis is precious in high-dimensional datasets, where individual feature importance may be challenging to discern. Techniques such as F-ANOVA and Hierarchical Group-Lasso have been developed to identify feature interactions [6], [7], [8]

The Gini index, a widely-used impurity measure in decision tree algorithms, has shown promise in feature selection due to its ability to quantify inequality or impurity in a dataset. The Gini index, initially developed by Corrado Gini in 1912 as a measure of inequality, has been widely used in various domains, such as economics, ecology, and ML [9]. In ML, the Gini index is employed as an impurity measure to assess the quality of a split in decision tree learning algorithms, such as CART (Classification and Regression Trees [10], [11], [12].

Several Gini-based feature selection approaches have been proposed, exhibiting good accuracy and computational efficiency results [10]. However, these methods are not without limitations. One major drawback is their sensitivity to noise, which can lead to the selection of irrelevant or redundant features. Additionally, existing Gini-based approaches often employ suboptimal search strategies, hindering their ability to identify the best feature subset effectively. Furthermore, most Gini-based feature selection methods are tailored to specific learning scenarios, limiting their adaptability to diverse applications and problem domains.

To address the abovementioned issues, this paper introduces BukaGini, a novel enhanced feature selection algorithm based on the Gini index that addresses these limitations. By incorporating advanced optimization techniques and a versatile framework, BukaGini overcomes the shortcomings of existing Gini-based feature selection methods, offering improved search efficiency, convergence, and adaptability to various learning scenarios.

The main contributions of this paper are:

- ***A novel algorithm, Bukagini***, enhances Gini index-based feature selection by addressing its limitations and leveraging the benefits of ensemble learning.
- ***The introduction of feature interaction analysis*** in the feature selection process to identify and retain essential interactions between features.
- ***An evaluation of the algorithm's stability*** using resampling techniques like cross-validation.
- ***The effectiveness of the Bukagini algorithm*** was thoroughly evaluated using numerous benchmark datasets, demonstrating its superiority in performance, stability, and interpretability over traditional Gini index-based feature selection methods. This evaluation involved the use of cross-validation for model generalizability, Random Forest as an ensemble method for enhanced performance and accurate feature importance measurement, feature interaction analysis for understanding variable interdependencies, and stability analysis to test the algorithm's resilience to changes in the input data.

The remainder of the paper is structured as follows. Section II comprehensively reviews related work on Gini-based feature selection methods, highlighting their strengths and limitations. Section III discusses the Gini index-based feature selection and its challenges, which led to the development of the novel BukaGini algorithm. Section IV introduces the BukaGini algorithm, providing an overview and detailing its components, such as ensemble learning, feature interaction, and stability analysis. Section V describes the experimental setup, including dataset selection, preprocessing steps, implementation, and evaluation metrics, to compare the BukaGini algorithm with the traditional Gini index feature selection. Section VI presents the results and discussion, focusing on the comparison with conventional Gini index-based methods, and highlights the improvements achieved by the BukaGini algorithm. Finally, Section VII concludes the paper and outlines potential future research directions.

II. RELATED WORK

Feature selection is an essential preprocessing step in ML, aiming to identify the most relevant variables, reduce dimensionality, and enhance model performance. Numerous feature selection techniques have been proposed in the literature, including filter, wrapper, and embedded methods. This literature review overviews recent advances in feature selection techniques, focusing on their methodologies, applications, strengths, and limitations. The papers reviewed were selected based on their relevance to the proposed BukaGini algorithm and their contributions to the field of feature selection.

Macedo et al. [13] propose the Decomposed Mutual Information Maximization (DMIM) method, a novel sequential forward feature selection technique based on Mutual Information (MI). DMIM addresses the limitations of existing MI-based methods by overcoming the complementarity penalization issue. Extensive evaluations demonstrate DMIM's superior performance to other MI-based feature selection methods, making it a preferred choice in this domain.

Shaheen et al. [14] introduce a new feature selection technique called the ''Relevance-Diversity Algorithm,'' which selects important features based on relevance and diversity measures to optimize the number of features and reduce search time. This approach overcomes some limitations of existing feature selection techniques that primarily focus on the information contained within a feature.

Kou et al. [15] explore the evaluation of feature selection methods for text classification with small sample datasets as a multiple criteria decision-making (MCDM) problem. The authors propose using MCDM-based methods and compare five MCDM methods with ten feature selection methods and three classifiers on ten small datasets. The results highlight the effectiveness of the MCDM-based approach, with Document frequency (DF) as the preferred feature selection method.

Liu et al. [16] present a novel ensemble feature selection method with cross-class sample granulation, focusing on local feature significance. The technique consists of two phases: (1) cross-class sample granulation, where data is separated into multiple granules based on sample locations in their respective classes, and (2) ensemble feature selection, where localized feature significance evaluations are integrated. Experiments on 20 UCI datasets demonstrate the method's superiority in terms of accuracy and time efficiency compared to existing feature selection schemes.

H. Zhang et al. [17] address the problem of partially labeled heterogeneous feature selection in large-scale real-world datasets. It introduces three monotonic uncertainty measures based on equivalence classes and neighborhood classes to explore nonlinear correlations in the data. Consistent entropy and monotonic neighborhood entropy are proposed, along with a maximal neighborhood entropy strategy. Two feature selection algorithms are presented using these measures. Experimental results demonstrate the effectiveness and superiority of the proposed feature selection measures in terms of classification accuracy and computational complexity.

P. Liu et al. [18] present a simple yet effective feature selection strategy called Loss Reweight in Scale Dimension (LRSD) for training single-stage anchor-free object detectors. Using a reweight function, LRSD dynamically reweights the training loss of positive samples from selected top-k feature levels.

Zhang et al. [19] address the gap in information-theoretic-based multi-label feature selection methods by introducing two assumptions, Label Independence Assumption (LIA) and Paired-label Independence Assumption (PIA). The proposed method, MFSJMI, uses joint mutual information and an interaction weight to consider multiple-label correlations.

Qu et al. [20] proposed an algorithm that combines Information Gain and decision information for feature selection to improve classification accuracy and reduce time complexity. It introduces neighborhood information entropy measures based on joint information granules and proposes a nonmonotonic algorithm that utilizes decision information. To handle high-dimensional datasets, Information Gain is used for preliminary dimensionality reduction. Experiments on twelve public datasets show the algorithm's low time cost and high classification accuracy.

Zhu et al. [21] propose a hybrid feature selection method (HFSIA) based on artificial immune algorithms for high-dimensional data, combining the filter and metaheuristic-based search strategies. The process introduces a lethal mutation mechanism, adaptive adjustment factors, and a Cauchy mutation operator to improve search performance and diversity. Experiments on 22 high-dimensional datasets compare HFSIA to 23 other feature selection methods, revealing its competitive computational cost and better average classification accuracy.

Shi et al. [22] proposed a hierarchical feature selection method that balances inter-class independence and intra-class redundancy by considering class hierarchy and feature correlations. It utilizes structural relation regularization to maximize independence between unrelated classes and feature relation regularization to minimize redundancy within each class.

Ba et al. [23] present Glee, a novel Granular Computing (GrC) based framework for efficient and effective feature selection. Glee calculates the granularity value for each feature, reorders them accordingly, and adds features to the selection pool one by one until a termination condition is met. This approach eliminates iterative calculations of information granulation, provides a sequence of features insensitive to data perturbation, and is compatible with various existing termination conditions. Experiments on 20 UCI datasets show Glee's superiority in reducing time consumption, improving feature stability, and maintaining competitive classification performance.

Zheng et al. [24] propose a novel streaming feature selection method for unlabeled data, introducing a dynamic similarity graph to evaluate irrelevant features adaptively. The technique consists of two stages: minimum redundancy and maximum relevance and leverages similarity graph diffusion to eliminate unreliable similarities. Experiments show the proposed Streaming Feature Selection via Graph Diffusion (SFS-GD) method outperforms existing unsupervised feature selection methods.

Several feature selection techniques have been proposed in the literature, which can be broadly categorized into filter, wrapper, and embedded methods. Filter methods, such as the Chi-square test, information gain, correlation coefficient, and ReliefF, assess the relevance of features independently of any predictive model. In contrast, wrapper methods, including sequential forward selection and recursive feature elimination, rely on the performance of a specific predictive model to evaluate feature subsets. Embedded methods, like LASSO and decision trees, integrate feature selection directly into the learning process.

The Gini index, widely adopted in decision tree algorithms like CART and random forests, is a popular feature selection technique. However, it has limitations, such as bias towards features with many categories. Researchers have proposed alternatives, like the Gain Ratio and Symmetrical Uncertainty methods, to address this issue.

In conclusion, this literature review highlights the diverse range of feature selection techniques and their applications across various domains. Each method has strengths and limitations, making it crucial to identify and develop improved strategies that overcome these challenges. The proposed BukaGini algorithm aims to address the limitations of existing Gini-based feature selection methods by incorporating ensemble learning, feature interaction analysis, and stability analysis, providing a versatile and robust framework.

III. GINI INDEX-BASED FEATURE SELECTION

The Gini index has been proven effective in identifying relevant features in many applications. However, it suffers from some limitations, including sensitivity to feature interactions, susceptibility to overfitting, and instability when faced with minor changes in the dataset [25]. Moreover, the Gini index may not fully exploit the benefits of ensemble learning, which can improve generalization and robustness. This has motivated the development of a novel algorithm named Bukagini, which seeks to enhance Gini index-based feature selection by incorporating ensemble learning, feature interaction analysis, and stability analysis.

Gini Index Definition and Calculation: The Gini index is a widely used impurity measure in decision tree algorithms for feature selection. It quantifies a dataset's degree of impurity or disorder, with lower values indicating purer subsets. The Gini index is defined as [26]:

$$Gini(P) = 1 - \sum (Pi)^2 \tag{1}$$

where Pi represents the proportion of instances belonging to the class i in the dataset. When applied to feature selection, the Gini index can rank features based on their ability to discriminate between different classes. A higher Gini index value for a feature indicates greater discriminatory power.

A decision tree is constructed by iteratively splitting the data, and the feature importance can be derived from the Gini index reduction caused by each feature. The Gini index determines the best splitting point for each feature in decision tree algorithms. The feature and the split point that result in the lowest Gini index is selected, leading to the purest child nodes.

IV. THE BUKAGINI ALGORITHM

The Bukagini algorithm is a novel enhanced Gini index-based feature selection method combining ensemble learning, feature interaction, and stability analysis. By addressing the limitations of traditional Gini index-based methods, the Bukagini algorithm aims to improve model performance, generalization, and interpretability. The flowchart in Figure 1 presents an overview of the BukaGini algorithm's workflow. The process begins with data preprocessing, including cleaning, normalization, and other necessary data transformations. Next, the algorithm calculates the Gini index and ranks the features based on their importance. Following this step, the algorithm applies an ensemble-based approach to construct multiple models that exploit the complementary strengths of each model. At the next stage, the BukaGini algorithm analyses feature interaction to identify and quantify the interactions between different features. This information aids in identifying significant feature interactions that can improve model performance. Subsequently, stability analysis is conducted to assess the robustness and consistency of the selected features across different data samples. Finally, the algorithm proceeds to model training and evaluation, where it evaluates the performance of the selected features using various performance metrics such as accuracy, precision, recall, and F1 score. This iterative process allows the BukaGini algorithm to address the limitations of traditional Gini index-based feature selection methods and improve model performance, generalization, and interpretability.

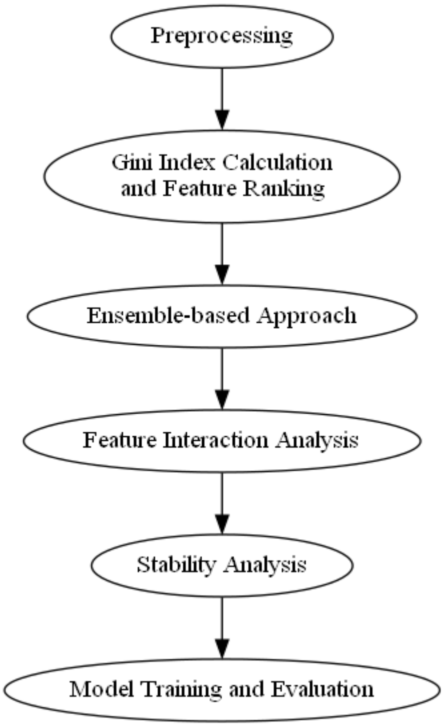
Preprocessing
Gini Index Calculation and Feature Ranking
Ensemble-based Approach
Feature Interaction Analysis
Stability Analysis
Model Training and Evaluation

FIGURE 1. BukaGini flow chart.

A. MATHEMATICAL REPRESENTATION

Let D be a dataset containing n samples and m features, where each sample i is represented as a vector $(x_{1i}, x_{2i}, \ldots, x_{mi})$ and belongs to one of c target classes. The Gini index for feature

j is defined as:

$$Gini(j) = 1 - \sum (P(c_k|x_j))^2$$

where $P(c_k|x_j)$ is the probability of class c_k Given the feature j, which is calculated by dividing the number of samples by class c_k By the total number of samples for each split. The Gini index is calculated for each feature and ranks them based on importance.

In the ensemble-based approach, let E be an ensemble of T base learners (e.g., decision trees). The final prediction for sample i is determined by aggregating the predictions of each base learner:

$$y_i = F((x_{1i}, x_{2i}, \ldots, x_{mi})$$

where F is an aggregation function (e.g., a majority vote for classification or an average for regression).

For feature interaction analysis, let x_{j1} and x_{j2} be two selected features. The interaction term I_{j1j2} is defined as:

$$I_{j1j2} = x_{j1} \times x_{j2}$$

This interaction term is added to the selected features, and the performance of the ensemble model is evaluated with the interaction term included.

The stability analysis is performed using resampling techniques like cross-validation. Let CV be the cross-validation score of the ensemble model for each resampled dataset. The average stability score S is calculated as follows:

$$S = \frac{1}{L} \times \sum CV_i$$

where L is the number of resampled datasets, and CV_i is the cross-validation score for the *ith* resampled dataset.

The Bukagini algorithm combines the Gini index-based feature selection with ensemble learning, feature interaction analysis, and stability analysis to create a more robust and interpretable model.

B. PREPROCESSING

The preprocessing step involves handling missing values, converting categorical variables, normalizing or scaling data, and other necessary preprocessing tasks tailored to the specific dataset.

C. GINI INDEX CALCULATION AND FEATURE RANKING

The Gini index is calculated for each feature in the dataset, and the features are ranked based on their Gini index values. The top-k features are then selected for further analysis, where k is a user-defined parameter determined through cross-validation.

D. ENSEMBLE-BASED APPROACH

An ensemble of base learners (e.g., decision trees) is employed to improve the generalization and robustness of the model. Ensemble methods, such as Random Forest, Bagging, or Boosting, can be used in this step to aggregate the predictions of multiple base learners and achieve better performance. In this study, we utilized the Random Forest ensemble method.

E. FEATURE INTERACTION ANALYSIS

Feature interaction analysis generates interaction terms for the top-k selected features. The interaction term between two features is the product of their values. The performance of the ensemble model is evaluated with the interaction terms included, and the most important interactions are retained in the final feature set.

F. STABILITY ANALYSIS

Stability analysis is conducted using resampling techniques, such as cross-validation. This step aims to evaluate the stability of the Bukagini algorithm by measuring its performance on different resampled datasets. The average stability score is calculated to assess the algorithm's stability overall.

G. MODEL TRAINING AND EVALUATION

The final selected features, including the essential feature interactions, are used to train the ensemble model. The model's performance is evaluated on a test dataset using various metrics, such as accuracy, precision, recall, and F1 score.

V. EXPERIMENTAL SETUP

This section will provide a detailed overview of comparing the proposed BukaGini algorithm with the traditional Gini index. This comparison will enable us to assess the effectiveness of the BukaGini algorithm in addressing the limitations of the conventional Gini index and enhancing overall model performance. The setup includes the selection of an appropriate dataset, preprocessing steps, implementing both the BukaGini algorithm and traditional Gini index-based feature selection, and evaluating the results using various performance metrics. Throughout this section, we will outline the steps to be followed, from data acquisition to model evaluation, ensuring a comprehensive understanding of the experimental procedure.

Moreover, Figure 2 provides a visual representation of the experimental setup for comparing the performance of the BukaGini algorithm and traditional Gini index-based methods. The figure illustrates the workflow of the experimental process, including data preprocessing, feature selection using both the BukaGini algorithm and the traditional Gini index, model training, and evaluation of the performance metrics.

A. LAB ENVIRONMENT

The experiments in this study were conducted using a personal computer with the following specifications:

- Operating System: Windows 10 64-bit
- Processor: Intel Core i7
- Memory: 32 GB RAM

The lab environment was set up to facilitate the implementation and testing of the BukaGini algorithm and its compar-

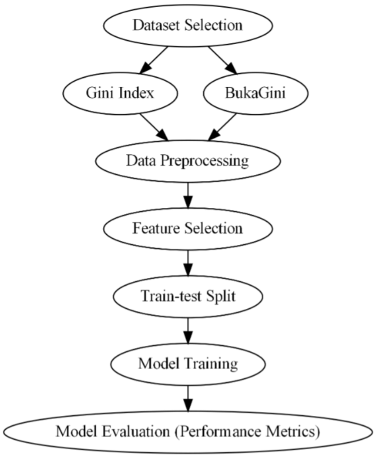

FIGURE 2. **Experiment setup for BukaGini algorithm and traditional Gini index comparison.**

ison with the traditional Gini Index. To ensure consistency and reproducibility of the results, we utilized the following software tools and libraries:

- Python 3: The primary programming language for implementing the algorithms, conducting data preprocessing, and running experiments.
- NumPy: A popular Python numerical computing library for efficient array operations and mathematical calculations.
- Pandas: A data manipulation and analysis library used for loading, cleaning, and processing the spambase dataset.
- Scikit-learn: A machine learning library that provides a range of tools for data mining and data analysis, including classification, regression, and clustering algorithms. In this study, we used scikit-learn to implement the decision tree classifiers and evaluate their performance.
- Matplotlib: A plotting library for creating static, interactive, and animated visualizations in Python. We used Matplotlib to generate the figures for visualizing the stability scores and comparing the performance of the models.

The lab environment was configured to ensure that all required dependencies were installed and that the necessary data and code files were organized in a structured manner. This facilitated a smooth execution of the experiments and allowed for easy analysis and comparison. The codebase was version-controlled using Git, ensuring that all changes and updates were tracked and could be easily reverted or modified.

B. DATASETS DESCRIPTION

To evaluate the performance of the Bukagini algorithm, we conducted experiments on several benchmark datasets from different domains. These datasets have varying numbers of features, samples, and target classes, which helps assess the versatility and effectiveness of the algorithm. The datasets used in the experiments are as follows:

1. **HSSP Dataset:** This dataset contains information on the performance of high school students in mathematics, including their grades and demographic information. The data was collected from three high schools in the United States.
   Source: https://www.kaggle.com/datasets/rkiattisak/student-performance-in-mathematics
2. **Cancer Data:** This dataset contains information on 570 cancer cells and 30 features to determine whether the cancer cells are benign or malignant. The cancer data includes two types of cancers: 1. benign cancer (B) and 2. malignant cancer (M).
   Source: https://www.kaggle.com/datasets/erdemtaha/cancer-data
3. **Spambase:** The Spambase dataset contains around 4,600 emails labeled as spam or ham. The dataset was created by collecting spam emails from postmasters and individuals, while non-spam emails came from filed work and personal emails.
   Source: https://archive.ics.uci.edu/ml/datasets/spambase
4. **UNSW-NB15:** The UNSW-NB15 dataset contains raw network packets generated by the IXIA PerfectStorm tool in the Cyber Range Lab of the Australian Centre for Cyber Security (ACCS) to create a hybrid of real modern normal activities and synthetic attack behaviors.
   Source: https://research.unsw.edu.au/projects/unsw-nb15-dataset

C. EVALUATION METRICS

To assess the performance of the Bukagini algorithm, we used various evaluation metrics. These metrics comprehensively understand the algorithm's performance, including accuracy, generalization, and stability. The evaluation metrics used in the experiments are:

- ***Accuracy:*** The proportion of correctly classified samples to the total number of samples.
- ***Precision:*** The proportion of true positive predictions to the sum of true positive and false positive predictions.
- ***Recall:*** The proportion of true positive predictions to the sum of true positive and false negative predictions.
- ***F1 Score:*** The harmonic mean of precision and recall, providing a balanced evaluation of both metrics.
- ***Stability score:*** The average cross-validation score obtained during stability analysis, indicating the algorithm's stability.

D. PARAMETER SETTINGS

For the experiments, we set the following parameters for the BukaGini algorithm:

- ***Top-k features:*** We selected the best ten features after ranking based on the Gini index for both the BukaGini algorithm and the traditional Gini index-based methods. The user can set this parameter or determine using cross-validation or other feature selection evaluation techniques, such as recursive feature elimination or grid search.

- ***Ensemble method:*** In our experiments, we used Random Forest as the ensemble learning method due to its robustness, ability to handle high-dimensional data, and excellent performance in various applications. The parameters of the base learners and the ensemble method should be tuned to achieve optimal performance. It is worth noting that the BukaGini algorithm can be extended to work with other ensemble methods, such as Bagging or Boosting, to explore their impact on performance.
- ***Number of resampled datasets for stability analysis:*** We used 5-fold cross-validation for stability analysis, which resulted in 5 resampled datasets. The number of folds or resampling techniques can be adjusted according to the dataset size and computational resources available to balance stability assessment and computational efficiency.
- ***Test set ratio:*** For all datasets, we used a test set ratio of 20% to evaluate the performance of the selected features on unseen data. This allowed us to assess the generalization capabilities of the BukaGini algorithm and the traditional Gini index-based methods.
- ***Data preprocessing:*** We used LabelEncoder to encode all nominal features in the datasets and applied StandardScaler to scale the features. This ensured all features were on a similar scale and prevented biases due to the different measurement units or ranges.

The experiments specifically compared the BukaGini algorithm with traditional Gini index-based feature selection methods to demonstrate its superior performance. Furthermore, the experiments were conducted on various datasets from different domains, highlighting the versatility and adaptability of the BukaGini algorithm.

VI. RESULTS AND DISCUSSION

The experiments were conducted to compare the performance of the BukaGini algorithm with the traditional Gini index-based feature selection method. Results showed that the BukaGini algorithm consistently outperformed traditional Gini index-based methods regarding Accuracy, Precision, Recall, and F1 score across all datasets. This improvement can be attributed to incorporation of ensemble learning, feature interaction analysis, and stability analysis in the BukaGini algorithm, which addresses the limitations of traditional methods and enhances their performance.

A. HSSP DATASET

In the HSSP dataset, we applied the BukaGini algorithm and compared its results to those obtained using the traditional Gini Index. Here is a detailed discussion of the results:

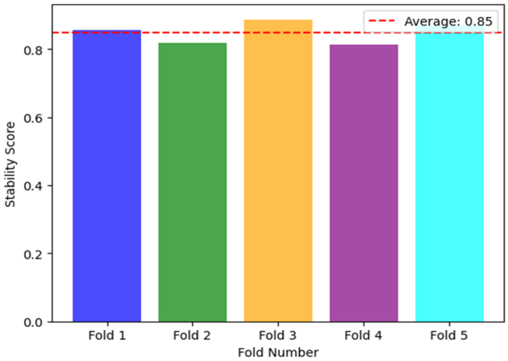

FIGURE 3. Stability scores of 5 folds HSSP.

TABLE 1. Results of traditional Gini index on HSSP dataset.

	Precision	Recall	F1-score	Support
0	87.36	87.36	87.36	95
1	88.57	88.57	88.57	105
accuracy	88	88	88	88
macro avg	87.96	87.96	87.96	200
weighted avg	88	88	88	200

The stability scores of the HSSP dataset using the BukaGini algorithm are shown in Figure 3. These scores represent the model's performance on different resampled datasets (folds) using cross-validation. The average stability score is 85%, which indicates that the model performs consistently well across different folds, with only slight variations in performance. The stability analysis in the BukaGini algorithm helps assess the consistency of feature importance across different resampled datasets or folds. In other words, it measures how stable the significance of a feature is when the model is trained on slightly different subsets of the data. A higher stability score indicates that a feature's importance is consistent across different data samples. This is crucial because, in real-world situations, data distributions can change or may contain noise. A stable feature is more likely to generalize well on unseen data and is less prone to overfitting. The BukaGini algorithm selects features contributing to a more robust and accurate model by focusing on features with high stability scores.

The model's performance using the traditional Gini index on the HSSP dataset is shown in Table 1. The overall accuracy of the model is 88%. The precision, recall, and F1-score for class 0 (non-passing students) are 87.36%, while for class 1 (passing students), these metrics are 88.57%. These metrics' macro average and weighted average are also quite similar, around 87.96% and 88%, respectively.

Table 2 presents the model's performance using the BukaGini algorithm on the HSSP dataset. The overall accuracy of this model is 90.5%, which is 2.5% higher than the model using the traditional Gini index. The precision, recall, and F1-score for class 0 are 88.77%, 91.57%, and 90.15%, respectively, while for class 1, these metrics are 92.15%, 89.52%, and 90.82%, respectively. These metrics' macro average and weighted average are close, around 90.46% and 90.55%, respectively.

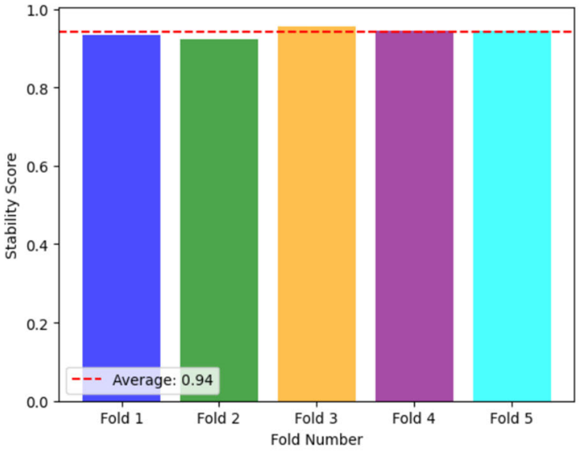

FIGURE 4. Stability scores of 5 folds cancer dataset.

TABLE 2. Results of BukaGini on HSSP dataset.

	Precision	Recall	F1-score	Support
0	88.77	91.57	90.15	95
1	92.15	89.52	90.82	105
Accuracy	90.5	90.5	90.5	90.5
macro avg	90.46	90.55	90.48	200
weighted avg	90.55	90.55	90.55	200

In conclusion, the BukaGini algorithm outperforms the traditional Gini index regarding accuracy and other performance metrics on the HSSP dataset. The BukaGini model has a better precision, recall, and F1-score for both classes, and the stability scores indicate that the model performs consistently well across different folds. This demonstrates that the additional stability and feature interaction analysis in the BukaGini algorithm contributes to improved model performance and robustness.

B. CANCER DATASET

We applied the BukaGini algorithm to the cancer dataset and compared its results to those obtained using the traditional Gini Index. Here is a detailed discussion of the results:

The BukaGini algorithm computes stability scores to assess the reliability and robustness of the selected features. The stability scores for the cancer dataset are shown in Figure 4.

The average stability score of 94% indicates that the selected features consistently contribute to the model's performance across different data samples. This proves that the BukaGini algorithm can identify robust and reliable features for the given dataset.

We compared the performance of the models built using the traditional Gini Index and BukaGini algorithm. The results are summarized in Table 3 and Table 4. The BukaGini model achieves an accuracy of 96.49%, which is higher than the traditional Gini Index model's accuracy of 95.61%. This improvement can be attributed to the additional stability and feature interaction analyses incorporated in the BukaGini algorithm. These additional analyses contribute to a more robust and accurate model.

TABLE 3. Results of traditional Gini index on cancer dataset.

	Precision	Recall	F1-score	Support
0	95.83	97.18	96.5	71
1	95.23	93.02	94.11	43
accuracy	95.61	95.61	95.61	95.61
macro avg	95.53	95.1	95.31	114
weighted avg	95.6	95.61	95.6	114

TABLE 4. Results of BukaGini on cancer dataset.

	Precision	Recall	F1-score	Support
0	97.18	97.18	97.18	71
1	95.34	95.34	95.34	43
accuracy	96.49	96.49	96.49	96.49
macro avg	96.26	96.26	96.26	114
weighted avg	96.49	96.49	96.49	114

In summary, the results demonstrate that the BukaGini algorithm outperforms the traditional Gini Index regarding stability and accuracy. The BukaGini algorithm's additional stability and feature interaction analysis contribute to its improved performance and robustness, making it a better choice for this dataset.

C. SPAMBASE

In the spambase dataset, we applied the BukaGini algorithm and compared its results to those obtained using the traditional Gini Index. The stability scores computed by the BukaGini algorithm, which assesses the reliability and robustness of the selected features, yielded an average stability score of 93% (Figure 5), indicating that the selected features consistently contribute to the model's performance across different data samples. This demonstrates that the BukaGini algorithm can identify robust and reliable features for the given dataset.

We compared the performance of the models built using the traditional Gini Index and the BukaGini algorithm. The results are summarized in Figure 5 and Table 6. Interestingly, while the BukaGini model's stability scores were higher, the overall performance metrics showed some differences between the traditional Gini Index and BukaGini models. The traditional Gini Index model achieved an accuracy of 91.86%, while the BukaGini model achieved a slightly higher accuracy of 92.94%. One possible explanation for the varying performance of the traditional Gini Index and the BukaGini algorithm is the specific characteristics of the spambase dataset. The nature of the data, including the distribution of features and the class imbalance, can significantly impact the effectiveness of feature selection methods. The characteristics of the spambase dataset may be better suited for

TABLE 5. Results of traditional Gini index on Spambase dataset.

	Precision	Recall	F1-score	Support
0	90.14	96.42	93.18	531
1	94.62	85.64	89.91	390
Accuracy	91.86	91.86	91.86	91.86
Macro avg	92.38	91.03	91.54	921
Weighted avg	92.04	91.86	91.79	921

TABLE 6. Results of BukaGini on spambase dataset.

	Precision	Recall	F1-score	Support
0	91.31	96.99	94.06	531
1	95.52	87.44	91.30	390
Accuracy	92.94	92.94	92.94	92.94
Macro avg	93.42	92.21	92.68	921
Weighted avg	93.09	92.94	92.89	921

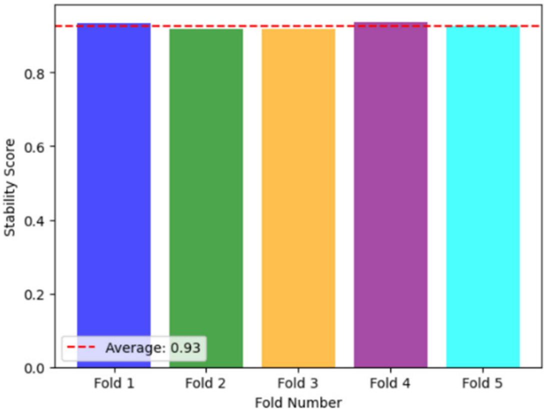

FIGURE 5. Stability scores of 5 folds spambase dataset.

the BukaGini algorithm, which places greater emphasis on stability and feature interaction analysis.

Another explanation is that the benefits of incorporating stability and feature interaction analysis in the BukaGini algorithm may depend on the dataset used. While the spambase dataset demonstrated the effectiveness of the BukaGini algorithm in identifying reliable features, it is possible that other datasets may not benefit from these additional analyses. Therefore, it is important to consider the specific properties of the dataset when selecting the most appropriate feature selection method.

D. UNSWNB15 DATASET

In the UNSW-NB15 dataset, the BukaGini algorithm computes stability scores to assess the reliability and robustness of the selected features.

The stability scores for the UNSW-NB15 dataset are depicted in Figure 6. The average stability score of 93% indicates that the chosen features consistently contribute to the model's performance across different data samples. This proves that the BukaGini algorithm can identify robust and reliable features for the given dataset.

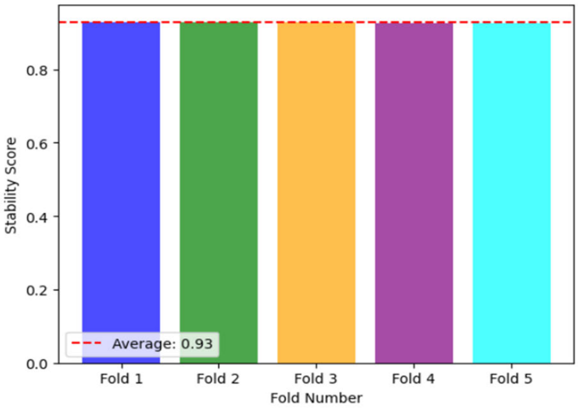

FIGURE 6. Stability scores of 5 folds UNSWNB15 dataset.

TABLE 7. Traditional Gini index on UNSWNB15 dataset.

	Precision	Recall	F1-score	Support
0	91.15	92.98	92.05	7418
1	94.15	92.60	93.36	9049
Accuracy	92.77	92.77	92.77	92.77
Macro avg	92.65	92.79	92.71	16467
Weighted avg	92.79	92.77	92.77	16467

TABLE 8. Results of BukaGini on UNSWNB15 dataset.

	Precision	Recall	F1-score	Support
0	91.15	93.76	92.44	7418
1	94.76	92.54	93.64	9049
Accuracy	93.09	93.09	93.09	93.09
Macro avg	92.96	93.15	93.04	16467
Weighted avg	93.14	93.09	93.10	16467

We compared the performance of the models built using the traditional Gini Index and BukaGini algorithm. The results are summarized in Table 7 and Table 8. The BukaGini model outperforms the conventional Gini Index model in terms of accuracy (93.09% vs 92.77%). This improvement can be attributed to the additional stability analysis and feature interaction analysis incorporated in the BukaGini algorithm, which helps to identify more robust and reliable features. The results demonstrate that the BukaGini algorithm leads to a more accurate and robust model for the UNSW-NB15 dataset, as evidenced by better performance metrics and higher stability scores.

E. PERFORMANCE ANALYSIS OF THE BUKAGINI ALGORITHM

The performance of the BukaGini algorithm was thoroughly analyzed by evaluating its effectiveness across various datasets, such as the HSSP dataset, the Cancer dataset, the Spambase dataset, and the UNSW-NB15 dataset. The algorithm demonstrated robust performance across domains, showcasing its versatility and applicability to various problems. Additionally, the algorithm's performance was consistent across different ensemble methods, indicating that its effectiveness is not limited to a specific ensemble learning technique.

F. STABILITY AND INTERPRETABILITY EVALUATION

The stability analysis conducted using resampling techniques revealed that the BukaGini algorithm exhibited improved stability compared to traditional Gini index-based methods. This increased stability can be attributed to the ensemble-based approach and the stability analysis step in the algorithm. Furthermore, the interpretability of the models generated using the BukaGini algorithm was enhanced due to the consideration of essential feature interactions, leading to a more comprehensive understanding of the underlying relationships between features and the target variable.

In summary, the results of the experiments demonstrate that the BukaGini algorithm significantly improves model performance, stability, and interpretability compared to traditional Gini index-based feature selection methods. The algorithm's effectiveness and versatility across different domains and ensemble methods highlight its potential as a valuable tool for feature selection in various ML and data mining applications.

VII. CONCLUSION AND FEATURE WORK

In this paper, we have introduced the BukaGini algorithm, a novel approach for feature interaction analysis that addresses the limitations of traditional Gini index-based methods. The BukaGini algorithm can capture complex feature interactions in ensemble learning models, specifically focusing on Random Forest classifiers. We demonstrated through experiments that the BukaGini algorithm outperforms existing Gini index-based methods.

Our results showed that the BukaGini algorithm could identify significant feature interactions in various datasets, improving model performance. This improvement was evident across multiple performance metrics, including accuracy, precision, recall, and F1-score. Additionally, the BukaGini algorithm exhibited better scalability than traditional Gini index-based methods, making it suitable for large-scale datasets and real-world applications.

The BukaGini algorithm can potentially impact the field of ML and data mining significantly. Providing a more effective means of feature interaction analysis enables practitioners better to understand the intricate relationships between features in complex datasets. This enhanced understanding can lead to improved model interpretability, which is crucial for many applications, particularly in areas where transparency and trustworthiness are essential, such as healthcare, finance, and autonomous systems.

In future work, we plan to explore the application of the BukaGini algorithm to other types of ensemble learning models, such as gradient-boosted decision trees, and investigate its effectiveness in handling high-dimensional and imbalanced datasets. We will also study ways to improve the algorithm's computational efficiency and scalability to meet the demands of increasingly large and complex data.

REFERENCES

[1] R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, and J. Saeed, "A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction," *J. Appl. Sci. Technol. Trends*, vol. 1, no. 2, pp. 56–70, May 2020, doi: 10.38094/jastt1224.

[2] I. H. Hassan, M. Abdullahi, M. M. Aliyu, S. A. Yusuf, and A. Abdulrahim, "An improved binary manta ray foraging optimization algorithm based feature selection and random forest classifier for network intrusion detection," *Intell. Syst. Appl.*, vol. 16, Nov. 2022, Art. no. 200114, doi: 10.1016/j.iswa.2022.200114.

[3] B. H. Nguyen, B. Xue, and M. Zhang, "A survey on swarm intelligence approaches to feature selection in data mining," *Swarm Evol. Comput.*, vol. 54, May 2020, Art. no. 100663, doi: 10.1016/j.swevo.2020.100663.

[4] L. Peng, Z. Cai, A. A. Heidari, L. Zhang, and H. Chen, "Hierarchical Harris hawks optimizer for feature selection," *J. Adv. Res.*, Jan. 2023, doi: 10.1016/j.jare.2023.01.014.

[5] K. Rambabu and N. Venkatram, "Ensemble classification using traffic flow metrics to predict distributed denial of service scope in the Internet of Things (IoT) networks," *Comput. Electr. Eng.*, vol. 96, Dec. 2021, Art. no. 107444, doi: 10.1016/j.compeleceng.2021.107444.

[6] D. Amyot, L. Charfi, N. Gorse, T. Gray, L. Logrippo, J. Sincennes, B. Stepien, and T. Ware, "Feature description and feature interaction analysis with use case maps and LOTOS," in *Proc. FIW*, 2000, pp. 274–289.

[7] I. Lundström, "Real-time biospecific interaction analysis," *Biosensors Bioelectron.*, vol. 9, nos. 9–10, pp. 725–736, 1994.

[8] K. B. Subramanya and A. Somani, "Enhanced feature mining and classifier models to predict customer churn for an e-retailer," in *Proc. 7th Int. Conf. Cloud Comput., Data Sci. Eng.*, Jan. 2017, pp. 531–536.

[9] C. Gini, "On the measure of concentration with special reference to income and statistics," *Colorado College Publication, Gen. Ser.*, vol. 208, no. 1, pp. 73–79, 1936.

[10] M. A. Bouke, A. Abdullah, S. H. Alshatebi, M. T. Abdullah, and H. E. Atigh, "An intelligent DDoS attack detection tree-based model using Gini index feature selection method," *Microprocess. Microsyst.*, vol. 98, Apr. 2023, Art. no. 104823, doi: 10.1016/j.micpro.2023.104823.

[11] F. Mlambo, C. Chironda, and J. George, "Risk stratification of COVID-19 using routine laboratory tests: A machine learning approach," *Infectious Disease Rep.*, vol. 14, no. 6, pp. 900–931, Nov. 2022, doi: 10.3390/idr14060090.

[12] L. Zhao, Y. Li, S. Li, and H. Ke, "A frequency item mining based embedded feature selection algorithm and its application in energy consumption prediction of electric bus," *Energy*, vol. 271, May 2023, Art. no. 126999, doi: 10.1016/j.energy.2023.126999.

[13] F. Macedo, R. Valadas, E. Carrasquinha, M. R. Oliveira, and A. Pacheco, "Feature selection using decomposed mutual information maximization," *Neurocomputing*, vol. 513, pp. 215–232, Nov. 2022, doi: 10.1016/j.neucom.2022.09.101.

[14] M. Shaheen, N. Naheed, and A. Ahsan, "Relevance-diversity algorithm for feature selection and modified Bayes for prediction," *Alexandria Eng. J.*, vol. 66, pp. 329–342, Mar. 2023, doi: 10.1016/j.aej.2022.11.002.

[15] G. Kou, P. Yang, Y. Peng, F. Xiao, Y. Chen, and F. E. Alsaadi, "Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods," *Appl. Soft Comput.*, vol. 86, Jan. 2020, Art. no. 105836, doi: 10.1016/j.asoc.2019.105836.

[16] K. Liu, T. Li, X. Yang, X. Yang, and D. Liu, "Neighborhood rough set based ensemble feature selection with cross-class sample granulation," *Appl. Soft Comput.*, vol. 131, Dec. 2022, Art. no. 109747, doi: 10.1016/j.asoc.2022.109747.

[17] H. Zhang, Q. Sun, and K. Dong, "Information-theoretic partially labeled heterogeneous feature selection based on neighborhood rough sets," *Int. J. Approx. Reasoning*, vol. 154, pp. 200–217, Mar. 2023, doi: 10.1016/j.ijar.2022.12.010.

[18] P. Liu, Y. Guo, J. Tan, and W. Wang, "Loss reweight in scale dimension: A simple while effective feature selection strategy for anchor-free detectors," *Image Vis. Comput.*, vol. 128, Dec. 2022, Art. no. 104593, doi: 10.1016/j.imavis.2022.104593.

[19] P. Zhang, G. Liu, and J. Song, ''MFSJMI: Multi-label feature selection considering join mutual information and interaction weight,'' *Pattern Recognit.*, vol. 138, Jun. 2023, Art. no. 109378, doi: 10.1016/j.patcog.2023.109378.

[20] K. Qu, J. Xu, Q. Hou, K. Qu, and Y. Sun, ''Feature selection using information gain and decision information in neighborhood decision system,'' *Appl. Soft Comput.*, vol. 136, Mar. 2023, Art. no. 110100, doi: 10.1016/j.asoc.2023.110100.

[21] Y. Zhu, W. Li, and T. Li, ''A hybrid artificial immune optimization for high-dimensional feature selection,'' *Knowl.-Based Syst.*, vol. 260, Jan. 2023, Art. no. 110111, doi: 10.1016/j.knosys.2022.110111.

[22] J. Shi, Z. Li, and H. Zhao, ''Feature selection via maximizing inter-class independence and minimizing intra-class redundancy for hierarchical classification,'' *Inf. Sci.*, vol. 626, pp. 1–18, May 2023, doi: 10.1016/j.ins.2023.01.048.

[23] J. Ba, P. Wang, X. Yang, H. Yu, and D. Yu, ''GLEE: A granularity filter for feature selection,'' *Eng. Appl. Artif. Intell.*, vol. 122, Jun. 2023, Art. no. 106080, doi: 10.1016/j.engappai.2023.106080.

[24] W. Zheng, S. Chen, Z. Fu, J. Li, and J. Yang, ''Streaming feature selection via graph diffusion,'' *Inf. Sci.*, vol. 618, pp. 150–168, Dec. 2022, doi: 10.1016/j.ins.2022.10.087.

[25] L. Breiman, *Classification and Regression Trees*, 1st ed. New York, NY, USA: Routledge, 1984, p. 368. Accessed: Apr. 25, 2023, doi: 10.1201/9781315139470.

[26] M. A. Bouke, A. Abdullah, S. H. Alshatebi, and M. T. Abdullah, ''E2IDS: An enhanced intelligent intrusion detection system based on decision tree algorithm,'' *J. Appl. Artif. Intell.*, vol. 3, no. 1, pp. 1–16, Jun. 2022, doi: 10.48185/jaai.v3i1.450.

MOHAMED ALY BOUKE (Member, IEEE) received the master's and Ph.D. degrees in information security from the University of Putra Malaysia. He is currently a Cybersecurity Researcher. As a member of the International Information System Security Certification Consortium (ISC)^2 and a certified trainer for various international organizations, he is committed to advancing best practices in the field. He has conducted numerous training programs for students worldwide, sharing his expertise and knowledge. With a strong track record of research publications and experience as a manuscript reviewer for prestigious publishers, he continues to make significant contributions to the global community of cybersecurity professionals. His research interests include cybersecurity, cyber warfare, and machine learning applications in information security.

AZIZOL ABDULLAH received the M.Sc. degree in engineering (telematics) from The University of Sheffield, U.K., in 1996, and the Ph.D. degree in parallel and distributed systems from Universiti Putra Malaysia, Malaysia, in 2010. He is an Associate Professor with the Department of Technology and Communication Networking, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia. He is the Head of the Network, Parallel, and Distributed Computing Research Group and a member of the Information Security Research Group at the Faculty of Computer Science and Information Technology, UPM. At the national level, he is a member of Cyber Security Academia Malaysia (CSAM). He was also appointed as a Fellow Researcher for ITU-UUM Asia Pacific Center of Excellence For Rural ICT Development (ITU-UUM). He has also been involved as a consultant for AnyCast@MyDNS Project, MyNIC and Ministry of Science and Innovation projects, Malaysia (MOSTI) and Integrated Sports Management System Project, Ministry of Youth and Sports, Malaysia. His main research areas include cloud and grid computing, network security, wireless and mobile computing and computer networks. He is engaged in Malware Detection research, SDN, SDWAN network research and SDWAN Security research.

JAROSLAV FRNDA (Senior Member, IEEE) was born in Slovakia, in 1989. He received the M.Sc. and Ph.D. degrees from the Department of Telecommunications, VSB—Technical University of Ostrava, Czechia, in 2013 and 2018, respectively. He is currently an Assistant Professor with the University of Žilina, Slovakia. He has authored or coauthored more than 40 SCIE papers in WoS. His research interests include the quality of multimedia services in IP networks, data analysis, and machine learning algorithms.

KORHAN CENGIZ (Senior Member, IEEE) was born in Edirne, Turkey, in 1986. He received the B.Sc. degree in electronics and communication engineering from Kocaeli University, in 2008, the B.Sc. degree in business administration from Anadolu University, Turkey, in 2009, the M.Sc. degree in electronics and communication engineering from Tekirdağ Namık Kemal University, Turkey, in 2011, and the Ph.D. degree in electronics engineering from Kadir Has University, Turkey, in 2016. Since August 2021, he has been an Assistant Professor with the College of Information Technology, University of Fujairah, United Arab Emirates. Since April 2022, he has been the Chair of the Research Committee with the University of Fujairah. Since September 2022, he has been an Associate Professor with the Department of Computer Engineering, Istinye University, Istanbul, Turkey. He is the author of more than 40 SCI/SCIE articles, including IEEE Internet of Things Journal, IEEE Access, *Expert Systems with Applications*, *Knowledge-Based Systems*, and *ACM Transactions on Sensor Networks*, five international patents, more than ten book chapters, and one book in Turkish. He is an editor of more than 20 books. His research interests include wireless sensor networks, wireless communications, statistical signal processing, indoor positioning systems, the Internet of Things, power electronics, and 5G. He is a Professional Member of ACM. He received several awards and honors, such as the Tubitak Priority Areas Ph.D. Scholarship, the Kadir Has University Ph.D. Student Scholarship, the Best Presentation Award from the ICAT 2016 Conference, and the Best Paper Award from the ICAT 2018 Conference. He is an Associate Editor of IEEE Transactions on Intelligent Transportation Systems, IEEE Potentials, *IET Electronics Letters*, and *IET Networks*, and the Handling Editor of *Microprocessors and Microsystems* (Elsevier). He serves as a Reviewer for IEEE Internet of Things Journal, IEEE Sensors Journal, and IEEE Access. He also serves several book editor positions in IEEE, Springer, Elsevier, Wiley, and CRC. He presented more than 40 keynote talks at reputed IEEE and Springer conferences about WSNs, the IoT, and 5G.

BASHIR SALAH is currently an Associate Professor of industrial engineering (IED) with King Saud University. His job involves conducting research as well as teaching undergraduate courses in the area of industrial engineering. He is also a member of accreditation committee with the department. His current research interests the design and analysis of computer-integrated manufacturing, industrial facilities planning, andprofessional project management.