IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 9 May 2024, accepted 28 May 2024, date of publication 3 June 2024, date of current version 10 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3408869

RESEARCH ARTICLE

A Dynamic Selection Hybrid Model for Advancing Thyroid Care With BOO-ST Balancing Method

MAJDI KHALID[1], F. M. JAVED MEHEDI SHAMRAT[2], (Member, IEEE), HANAN ALSHANBARI[1], MAJED FARRASH[1], AND THAMIR M. QADAH[3], (Member, IEEE)

[1]Department of Computer Science and Artificial Intelligence, College of Computing, Umm Al-Qura University, Makkah 21955, Saudi Arabia
[2]Department of Computer System and Technology, Universiti Malaya, Federal Territory of Kuala Lumpur 50603, Malaysia
[3]Department of Computer and Network Engineering, College of Computing, Umm Al-Qura University, Makkah 21955, Saudi Arabia

Corresponding author: Hanan Alshanbari (hsshanbari@uqu.edu.sa)

ABSTRACT Recently, thyroid disease has been a leading cause of mortality, underscoring the importance of early diagnosis to mitigate its impact. Researchers have randomly employed static selection ensemble methods aiming to forecast the disease in its initial stages. However, the use of such ensemble methods in healthcare diagnosis poses challenges related to performance consistency and potential mismatches with new data characteristics. Hence, this paper proposes a novel approach by introducing the Dynamic Selection Hybrid Model (DSHM) that leverages the most effective conventional classifiers using an appropriate ensemble method. Instead of going the conventional way, we evaluate various baseline classifiers to demonstrate their impact on the characteristics selected by two robust feature selection techniques. This evaluation employs an explainable AI (XAI) method, Permutation Feature Importance (PFI), and selects the most effective classifiers based on their characteristics impact. Then the selected classifiers are integrated by an appropriate ensemble method, based on a comparative evaluation between four efficient ensemble methods. Allowing the proposed DSHM to dynamically adjust its composition based on selecting conditions can potentially achieve robust performance by better adaptability on unseen data. Before the training DSHM, the methodology begins by addressing the dataset's imbalance issue using an effective data balancing method BOO-ST. To demonstrate the superiority of DSHM, various performance evaluation matrices, and a statistical test are employed. The experimental results reveal the effectiveness of our proposed DSHM, outperformed with an impressive 99.33% accuracy. Finally, to enhance transparency, trust, and patient outcomes, we applied the Local Interpretable Model-agnostic Explanations (LIME) to explain DSHM-provided outcomes. With a robust classification performance, our proposed DSHM aims to explain its outcomes, contributing to improved clinical decision-making processes and ultimately enhancing patient care.

INDEX TERMS Dynamic selection hybrid classifier, feature selection, explainable AI, thyroid disease.

I. INTRODUCTION

Thyroid disease is a widespread global health concern that arises when the thyroid gland is unable to produce a sufficient relevant hormone. Individuals affected by this condition may exhibit specific symptoms, such as loss of hair, lethargy, weight gain, an accelerated heart rate, and dry skin [1]. Left untreated, thyroid disease can lead to various health complications, including joint pain, infertility, obesity, and heart disease [2]. The prevalence of this condition in the general population ranges from 0.3% to 3.7% in the USA and 0.2% to 3.5% in Europe, with a gradual increase observed annually [3]. The American Cancer Society estimates that the aberrant proliferation of thyroid gland cells accounted for 2120 more deaths in the USA in 2023 [4].

Detecting or predicting thyroid disease at an early stage is crucial for preventing severe consequences. However, traditional laboratory tests for diagnosing this disease are intricate and demand extensive knowledge and expertise. Moreover, the manual diagnostic process is time-consuming and may yield inaccurate results. Machine Learning (ML) has emerged as a widely accepted approach to tackling these challenges and providing early warnings for disease prevention. Despite its potential, these studies for disease

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

forecasting pose significant issues, as discussed in the following subsection.

A. PROBLEM STATEMENT AND RESEARCH GAP

Previous research efforts primarily concentrated on addressing the imbalance issues [5], [6], [7], [8], training models using conventional ML algorithms [9], [10], [11], [12], [13], and employing traditional ensemble methods randomly [14], [15], [16], [17], [18]. For instance, the study by [5] utilized the synthetic minority oversampling technique (SMOTE) to balance the thyroid dataset before initiating the training process. Similarly, Sultana and Islam [6] employed SMOTE to tackle the class imbalance problem. The use of SMOTE remained consistent across studies [7], [8], given its well-established reputation for data balancing. It is imperative to acknowledge that SMOTE may result in noisy and uninformative data, resulting in, potentially compromising the model's efficiency [19]. Subsequently, Savcı and Nuriyeva [9] employed six machine learning (ML) algorithms, including Random Forest (RF) and Support Vector Machine (SVM), to classify thyroid stages. Olatunji et al. [10] also utilized RF and SVM for this task on the Saudi Arabian thyroid dataset. In addition to these, other standalone ML classifiers, such as Decision Tree (DT) [11], Gradient Boost (GB), K-Nearest Neighbors (KNN) [12], and Ada Boost (AB) [13] were applied. Though these studies achieve extraordinary outcomes for the task, however, standalone ML algorithms exhibit limited effectiveness in handling complex and diverse datasets and are not stable due to the algorithm's stochastic nature [20]. Consequently, Solmaz et al. [14] focused on a hybrid ML classifier using the ensemble method Boosting (BS). The study [15] employed another ensemble technique named Bagging (BG). Hybrid models were created by Dharamkar et al. [16] and Akhtar et al. [17] utilizing Voting (VT), while Yadav and Pal [18] employed the Stacking (ST) ensemble method in their study. It's noteworthy that the selection process of base classifiers in these hybrid models is static, as the classifier is predefined before the training phase and remains unchanged. This limitation hampers their ability to maintain optimal performance in situations where the relationships between features and the target variable may vary over time [21]. Moreover, the authors randomly selected different ensemble methods without considering their compatibility with the characteristics of the dataset, which lies in the potential lack of optimization and adaptability to the characteristics of the dataset.

B. NOVELTY AND CONTRIBUTION

In response to the concerns mentioned above, we initially employed an effective data-balancing method named BOO-ST [22]. This method initially utilizes BS to improve the minority class's representatives before applying SMOTE. BS assists SMOTE in mitigating the issues related to producing noisy and irrelevant samples by enhancing the model's discriminative ability, focusing on informative features, prioritizing the correct classification of minority class instances, and improving generalization to unseen data [23], [49]. Subsequently, Tomek links were also considered to remove any raucous and useless synthetic samples if still generated by SMOTE [22], [24]. By employing these ways, we effectively address the issues related to SMOTE in our study. Then, instead of static selection conventional classifiers and randomly selected ensemble methods, we proposed a Dynamic Selection Hybrid Model (DSHM) considering effective baseline classifiers and an appropriate ensemble method. To establish a dynamic selection process, we conducted a comparative analysis of six well-known ML classifiers (e.g., DT, KNN, SVM, RF, GB, and AB), commonly used in thyroid classification tasks. An explainable AI (XAI) technique named Permutation Feature Importance (PFI) is utilized to assess the significance of each classifier on selected feature sets [50]. Based on this analysis, we selected half of the classifiers as base estimators for this predictive task [51]. Additionally, we make a comparative evaluation between the predictive results of four ensemble algorithms (e.g., BS, BG, VT, and ST). From this analysis, we choose the most appropriate ensemble method for the task, and applying it we integrate the selected baseline classifiers, resulting in presented DSHM. This proposed model possesses the ability to adjust their ensemble of base classifiers dynamically based on the evolving nature of the input data. The dynamic adaptability and responsiveness make it more suitable where the data distribution is non-stationary or subject to changes over time [21]. Overall, DSHM offers greater adaptability, performance optimization, robustness to model drift, and enhanced diversity compared to static selection methods. This enables the model to better cope with the dynamic nature of real-world data and tasks, leading to improved predictive performance and reliability [52]. The key contributions of the research work are as follows:

- We employed an effective data balancing method named BOO-ST to address the limitations associated with SMOTE.
- We applied two robust feature selection methods, namely Univariate and Information Gain, to identify and retain the most relevant features while reducing dimensionality.
- Incorporated a dynamic selection phrase using an explainable AI, named PFI and introduced the Dynamic Selection Hybrid Model.
- We conducted a comprehensive and intuitive comparison between the static and our proposed dynamic model. Wherein, DSHM emerged as the overall superior model, achieving an impressive accuracy of 99.33%.
- Utilized the LIME and SHapley Additive exPlanations (SHAP) to clarify the outcome reasons of DSHM and global behavior of characteristics.

II. RELATED WORK

In recent years, a significant number of researchers have turned their attention to the detection of thyroid diseases.

TABLE 1. A summary of recent machine learning-based studies on the diagnosis of thyroid disease and their core limitations.

Author and year	Data size (row, column)	Target classes	Performed classifier	Study gap and limitations
Mohammad [5] 2023	3772, 29	2	SVM, KNN, DT BS, BG	The use of SMOTE has the potential to generate the noisy and irrelevant synthetic instances
Sultana [6] 2023	2800, 28	2	SVM, KNN, DT RF, GB, AB	The use of SMOTE has the potential to generate the noisy and irrelevant synthetic instances
Haneet [7] 2023	7200, 21	3	Linear Discriminant Analysis, BG	The use of SMOTE has the potential to generate the noisy and irrelevant synthetic instances
Saima [8] 2022	3163, 25	2	RF, DT, SVM, KNN, ANN	The use of SMOTE has the potential to generate the noisy and irrelevant synthetic instances
Ege [9] 2022	7200, 21	3	RF, SVM, KNN, ANN	Standalone ML algorithms exhibit limited effectiveness in handling diverse datasets and are not stable
Olatunji [10] 2021	218, 15	2	RF, SVM ANN	Standalone ML algorithms exhibit limited effectiveness in handling diverse datasets and are not stable
Tahir [11] 2022	3163, 29	2	RF, DT KNN, ANN	Standalone ML algorithms exhibit limited effectiveness in handling diverse datasets and are not stable
Diganta [12] 2023	3772, 29	2	RF, SVM, KNN, GB	Standalone ML algorithms exhibit limited effectiveness in handling diverse datasets and are not stable
Rajasekhar [13] 2022	9172, 31	5	RF, SVM AB	Standalone ML algorithms exhibit limited effectiveness in handling diverse datasets and are not stable
Ramazan [14] 2020	7200, 21	3	BS, AB, DT	Static selection hybrid model hampers their ability to maintain optimal performance and potential to lack of adaptability
Yufei [15] 2023	9172, 26	6	BG, VT	Static selection hybrid model hampers their ability to maintain optimal performance and potential to lack of adaptability
Bhavna [16] 2020	7547, 30	2	VT, C4.5, RF	Static selection hybrid model hampers their ability to maintain optimal performance and potential to lack of adaptability
Akhtar [17] 2021	309, 11	3	DT, GB, RF, BS, BG	Static selection hybrid model hampers their ability to maintain optimal performance and potential to lack of adaptability
Yadav [18] 2020	12204, 12	3	BS, BG VT, ST	Static selection hybrid model hampers their ability to maintain optimal performance and potential to lack of adaptability

Some have initiated efforts to address dataset imbalances. For instance, Alshayeji [5] proposed a thyroid classification system that employed SMOTE to balance the dataset before applying ML classification algorithms. The proposed model demonstrated satisfactory outcomes in disease prediction. Sultana and Islam [6] utilized a thyroid dataset from the University of California Irvine (UCI) repository, addressing the imbalance issue through SMOTE. They applied two feature selection techniques to reduce the dimensionality of synthetic samples, and RF achieved the highest accuracy of 99% within one of the subsets. Kour et al. [7] introduced a bagged-based ensemble model, incorporating SMOTE to identify thyroid disorders. They utilized two thyroid datasets and achieved accuracy rates of 85.45% and 82.71% from their proposed model. Islam et al. [8] employed the Sick-euthyroid dataset from the UCI repository, implementing both oversampling and under-sampling using SMOTE on the raw dataset. Their over-sampled dataset yielded the highest result, with 95.87% accuracy.

Subsequently, the authors also emphasize the utilization of different ML classifiers, conducting a comparative evaluation among them. For instance, Savcı and Nuriyeva [9] explored various ML algorithms, with the Artificial Neural Network (ANN) achieving the highest accuracy of 98% compared to others. Olatunji et al. [10] presented an ML-based tool using a Saudi Arabian dataset, where among several conventional algorithms, RF emerged as the top performer with 90.91% accuracy. Alyas et al. [11] analyzed various ML classifiers, where the RF algorithm demonstrated a generalized accuracy of 94.8%. Dignata et al. [12] also presented an analysis of different ML classifiers, where RF consistently outperformed other classifiers with an accuracy of 99.14%. Chaganti et al. [13] employed several ML classifiers, highlighting that the Extra Tree (ET) features subset yielded the highest results, achieving 99% accuracy with the RF algorithm. Additionally, they asserted the essential role of ML classifiers over Deep Learning in terms of accuracy and complexity.

Moreover, several researchers have introduced static selection ensemble methods utilizing BS, BG, VT, and ST. For instance, Solmaz et al. [14] introduced an Android thyroid diagnosis application using a hybrid algorithm, achieving a 99.08% accuracy with the aid of the BS ensemble method. Xie et al. [15] proposed another static hybrid selection model using the BG ensemble method, considering multiple base estimators, which demonstrated an acceptable score of 92.3% accuracy in the experimental section. Subsequently, Dharamkar et al. [16] introduced a hybrid model combining C4.5 and RF by VT, named CCTML, which achieved an accuracy of 96%. Similarly, Akhtar et al. [17] proposed a homogeneous hybrid model activating both BS and BG using both soft and hard VT. Finally, Yadav and Pal [18]

introduced a static hybrid system using the ST ensemble method. An overall summary of these studies is presented in Table 1.

Furthermore, the prevalence of developing hybrid or ensemble ML models is also increasing in various biomedical data. For example, Shahid et al. [43] proposed a Multi-Feature Representation and Genetic Algorithm-Based Deep Ensemble Model to identify the Anti-Tubercular Peptides. Akbar et al. [44] introduced an ensemble deep neural network-based model for classifying anticancer peptides. Additionally, to identify antifreeze proteins [45], anti-inflammatory peptides [46], antiviral peptides [47], and anti-fungal peptides [48] the researchers also presented several efficient hybrid or ensemble ML models.

III. RESEARCH METHODOLOGY

In this section, we offer a thorough discourse on the methodologies and procedures utilized in the research. The working methodology is segmented into five primary components, encompassing data collection, data preprocessing, feature selection or dimension reduction, proposed dynamic selection classifiers, and performance evaluation. Figure 1 provides an overview of the entire working process.

A. DATA COLLECTION

A real-world dataset related to thyroid cases is sourced from the Kaggle data repository [25], comprising 30 features and 3,772 distinct case records. Table 2 provides details on their respective data types (such as string, float, integer, Boolean, and constant), short descriptions, and the count of identical values. Noteworthy features such as T3, TT4, TSH, T4U, and FTI assess various functions related to the disease and measure their levels in blood through lab tests. Additionally, clinical features like Pregnant, Sick, Thyroid Surgery, Goiter, Tumor, and Psych are included in the dataset. A very unbalanced dataset is indicated by the study of the test report, which found 231 positive thyroid cases and 3,541 negative thyroid cases in the target class.

B. DATA PREPOCESSING

Data preprocessing stands as a pivotal phase in the training of any machine learning model, facilitating the extraction of meaningful insights from the original dataset. Essentially, these processing techniques serve to convert the original data into a comprehensible and readable format. In our experimentation, six distinct preprocessing techniques are employed. Primarily, since the majority of features are categorical, a conversion into numeric vectors is imperative before feeding the data into some classification algorithms in ML [26]. Utilizing a Level encoder, the data is encoded without altering its dimension [27]. Subsequently, a considerable number of features are identified with missing values, including T3, TSH, Age, T4U, TT4, TBG, FTI, and Sex. Features with missingness exceeding 50% are excluded from further analyses [28]. For the remaining features, an imputation technique, specifically mean interpolation, is applied to address the missing values. In the subsequent subsection, we elaborate on the data-balancing method named BOO-ST, employed to mitigate issues related to data imbalance.

1) DATA BALANCING WITH BOO-ST

In contemporary times, the imbalance of datasets has become a prevalent issue, particularly in publicly accessible datasets. When the number of examples in one class greatly surpasses or falls short of those in another class, this situation occurs. This issue may cause the model to become biased in favor of the majority class, which would lead to poor performance of the minority class and misleading performance measures [29]. Consequently, researchers express considerable concern about this issue and aim to address it proactively before commencing data training. The synthetic minority oversampling technique (SMOTE) stands out as a renowned approach for balancing data, often favored by researchers [5], [6], [7], [8], [11]. However, it is important to highlight that this approach frequently introduces irrelevant and noisy data when creating synthetic instances [24].

In this study, we have systematically tackled both the challenges associated with the class imbalance and the intricacies of SMOTE through a three-stage process termed BOO-ST [22], encompassing BS, SMOTE, and Tomek link (TL). Minority classes often encounter misclassification issues due to their underrepresentation, lacking sufficient instances to adequately capture complex patterns. Hence, initially, we applied the BS to the imbalanced dataset, denoted as D, across I iterations. During this stage, D is trained with equal weights $(1/N)$ assigned to samples, and the learning rate (LR) is computed, where N represents the number of samples. The LR are then utilized to increase the weights for minority class samples, ensuring that in subsequent stages, greater emphasis is placed on minority instances. This deliberate weighting helps to produce more varied synthetic instances and improves the minority class's representation [23].

After adjusting the weights, we applied the SMOTE technique to D. This technique calculates the imbalance ratio as M/S, where M and S represent the number of minority classes and samples, respectively. Subsequently, it identifies the k nearest neighbors (KNN) from M and randomly selects these neighbors. The newly generated synthetic instances are then incorporated into the augmented dataset (AG). However, there is a potential drawback in this process, as the inclusion of noisy and irrelevant synthetic instances may introduce high complexity and hinder result reproducibility. Therefore, in the final stages, we address these concerns by applying TL to the AG. In this stage, we once again identify KNN from both majority and minority samples in AG. This entails figuring out the Euclidean distance between each instance in AG and the feature vector, then choosing the instances from both classes with the shortest distances. After that,

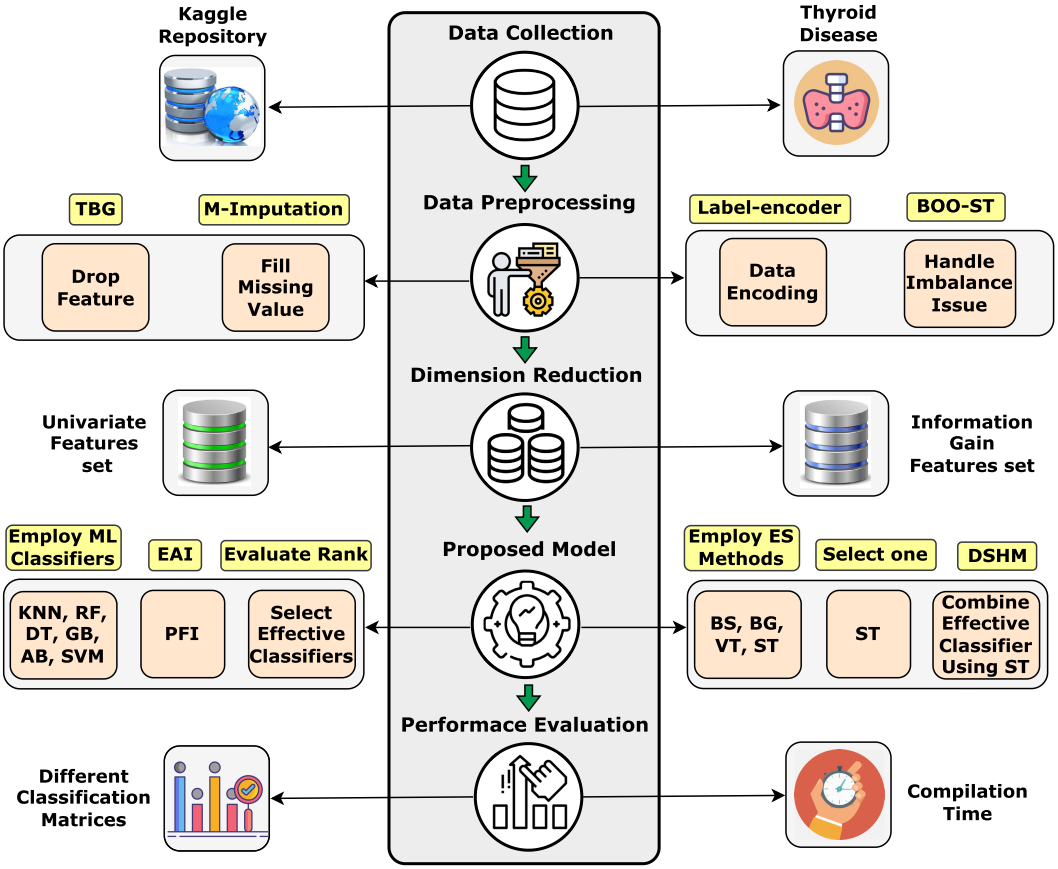

FIGURE 1. A flow diagram of our proposed study with different primary components.

we find and eliminate the majority class data samples that are most similar to the minority class data. Due to the removal of irrelevant and noisy samples, these approaches greatly minimize the complexity of AG [30]. The utilized BOO-ST method resulted in the generation of 2815 instances in the negative class and 3156 instances in the positive class. Additional details regarding the employed method can be found in the study [22].

C. DIMENSION REDUCTION

Dimension reduction, also known as feature selection, is a systematic procedure designed to eliminate irrelevant features and choose a subset of relevant ones from the original dataset. The goal is to retain the most informative features while reducing their number, all without compromising the model's ability to generalize [31]. This method not only reduces computational expenses but also enhances model performance by addressing the challenges associated with the curse of dimensionality. In this study, we have applied two well-known feature selection methods, Univariate and Information Gain, explanations are provided in the following subsections.

1) UNIVARIATE FEATURE SELECTION (UFS)

UFS is used to select the informative features from a dataset based on their relationship with target features. Here we set the f-statistic metric for the evaluation of each feature and rank them based on their scores. Features with higher scores are considered more relevant in the context of predictive tasks. Then, a selection threshold is determined, features with scores above this threshold are selected for the final feature subset, while those below the threshold are discarded.

2) INFORMATION GAIN FEATURE SELECTION (IGS)

IGS is an entropy-based feature selection method for identifying informative feature subsets. It assesses the reduction in entropy or disorder resulting from the transformation of a dataset. It primarily operates by computing the Information Gain of each feature concerning the target class. A higher Information Gain indicates that splitting the dataset based on that specific feature is more effective in reducing uncertainty.

From these dimension reduction methods, we have discarded half of the irrelevant features [32] and selected the rest of them for the predictive task. The table 3 shows the reprieved features from these selection techniques.

D. CLASSIFIERS DESCRIPTIONS

We have employed six conventional well-known ML classifiers: DT, SVM, KNN, RF, AB, and GB. Additionally, we proposed A Dynamic Selection Hybrid Model by fitting

TABLE 2. Details of the employed thyroid dataset.

Attributes	Value	Type	Identical Values	Descriptions
Age	Real Number	Integer	94	Age in years
Sex	Male, Female	Boolean	2	Gender Instance
On thyroxine	True, False	Boolean	2	Having triiodothyronine or not
Query on thyroxine	True, False	Boolean	2	Type of thyroid hormone
On antithyroid	True, False	Boolean	2	Prevent the thyroid form
Sick	True, False	Boolean	2	About a person's sickness
Pregnant	True, False	Boolean	2	Is patient pregnant
Thyroid surgery	True, False	Boolean	2	Remove a part or all of the thyroid gland
I131 treatment	True, False	Boolean	2	Radioactive iodine therapy
Query hypothyroid	True, False	Boolean	2	Patient's query hypothyroid
Query hyperthyroid	True, False	Boolean	2	Makes more hyperthyroid hormones
Lithium	True, False	Boolean	2	Used to treat Certain Psychiatric disorders
Goiter	True, False	Boolean	2	Swelling in front of the neck
Tumor	True, False	Boolean	2	An abnormal mass of tissue
Hypopituitary	True, False	Boolean	2	Fails to produce more hormones
Psych	True, False	Boolean	2	Combining forms denoting the mind
TSH measured	True, False	Boolean	2	Measure hormone is in patient's blood
TSH	Real number	Float	288	Thyroid-stimulating hormone
T3 measured	True, False	Boolean	2	Involves having patient blood drawn
TT4 measured	True, False	Boolean	2	Measures the label of total T4
TT4	Real number	Float	242	Total thyroxine
T4U	Real number	Float	147	Thyroxine utilization rate
T4U measured	True, False	Boolean	2	Measures or not
FTI	Real number	Float	235	Thyroxine utilization rate
FTI measured	True, False	Boolean	2	Measures or not
TBG	Null	(-)	(-)	Thyroxine-binding globulin
TBG measured	False	Constant	1	Thyroxine-binding globulin measured or not
T3	Real number	Float	70	Lab report for triiodothyronine
Referral source	SVHC, other, SVI, STMW, SVHD	String	5	Referencing source
Class	Negative, Sick	Boolean	2	Thyroid affected or not

TABLE 3. Selected feature subsets from the univariate and information gain feature selection method.

Univariate		Information Gain	
Feature	Score	Feature	Score
T3	618.4	T3	0.5406
Age	247.4	Age	0.4562
TT4	162.6	TT4	0.4285
T3 Measured	156.3	FTI	0.4218
On Thyroxine	53.8	TSH	0.3418
TSH Measured	23.2	Sex	0.1779
TT4 Measured	23.2	T3 Measured	0.0542
Query Hyperthyroid	20.4	TT4 Measured	0.0375
Psych	14.8	On Thyroxine	0.0373
T4U	12.8	Psych	0.0202
T4U Measured	9.2	TSH Measured	0.0157
Sex	9.1	Pregnant	0.0148
FTI	5.0	Tumor	0.0142
Thyroid Surgery	4.4	Hypopituitary	0.0125

an effective ensemble method. Subsequent sections describe each classifier briefly.

1) DECISION TREES (DT)

An algorithmic method is used to create DT, which finds the best methods to divide a dataset according to certain criteria. "Splitting" is the process of building the tree from the root node upwards, choosing the "Best Feature" from among the features that are present. The determination of the "Best Feature" involves calculating Entropy (E) and Information Gain (IG). The formulas for computing E and IG are expressed in Eq.1 and 2, where X is the attributes, Y represents the class level, and $(P+)$ and $(P-)$ denote positive and negative samples, respectively. It categorizes the data points from the root node to the terminal node, where the terminal node provides the classification of a particular feature. For every subtree rooted at the new node, this iterative procedure is repeated. Notably, the DT classifier is a suitable choice when dealing with datasets containing a significant amount of discrete, logical, or categorical data [33].

$$E(D) = -(P+)\, log_2\,(P+) - (P-)\, log_2\,(P+), \quad (1)$$

$$IG(X) = E(X) - E(X, Y). \quad (2)$$

2) SUPPORT VECTOR MACHINE (SVM)

SVM is an effective supervised learning technique that may be used for tasks involving both classification and regression. Using datasets with several classes, SVM seeks to determine the best decision boundary or hyperplane [34]. The principal aim is to identify the hyperplane with the highest margin, denoting the separation between the hyperplane and the closest data point for every class. The functional representation of SVM is depicted in Eq. 3, where X

represents the input, W signifies the weight, B denotes the bias, T stands for transpose, and $SIGN()$ is a function providing either 1+ or 1− based on the input data type.

$$SVM(X) = SIGN\{W^{(T)}X + B\}. \tag{3}$$

3) K NEAREST NEIGHBORS (KNN)

KNN algorithm aims to identify the optimal class for test data by assessing the distance between the test data and training points. KNN offers flexibility through various types of modifications, allowing for adaptations to specific scenarios. This algorithm demonstrates resilience to considerable noise in training data and proves effective when dealing with substantial training datasets [35]. The core operation of KNN involves computing the Euclidean distance (Eq. 4) between each set of raw training data and the test data. In this equation, (X_1, X_2) and (Y_1, Y_2) represent the coordinates of the first and second points, respectively.

$$Euclidean = \sqrt{(X_2 - X_1) + (Y_2 - Y_1)}. \tag{4}$$

4) RANDOM FOREST (RF)

RF classifier is constructed using an ensemble of Decision Trees (DTs). Each tree in the ensemble is generated from a sample drawn from the training set with replacement, known as the bootstrap sample [36]. For the classification task, a majority voting approach is employed, wherein the predicted class is determined by selecting the most frequently occurring class. The features are denoted as $X = \{x_1, x_2, x_3, \ldots, x_n\}$ with corresponding responses $Y = \{y_1, y_2, y_3, \ldots, y_n\}$ where n represents the number of samples. There is a lower limit of 1 and an upper limit of L for the index l. The prediction for samples is carried out by averaging the predictions for x^p given by each distinct tree for x, as shown in Eq. 5.

$$RF = \frac{1}{L}\sum_{l=1}^{L} l(x^p). \tag{5}$$

5) ADA BOOST (AB)

AB stands as an ensemble boosting classifier that assembles a robust and accurate classifier by combining multiple weak classifiers. The main advantage of the AB is that it is less prone to over-fitting and correct misclassification of weak learners [37]. The core idea behind AB classifiers involves adjusting the weights of the data and training the data samples with an initial weight of $1/F$, where F represents the frequency of training instances. Subsequently, after obtaining the outcome, the error is calculated as $(correct - F)/F$. Finally, the classification is measured using the Eq. 6, where e is the number of weak learners, $h_e(p)$ is the prediction of e, and a_e represents the weight of e.

$$AB = +/-\sum_{e=1}^{e}(a_e h_e(p)). \tag{6}$$

6) GRADIENT BOOST (GB)

GB emerges as an ensemble technique that combines multiple weak learning models to create a robust predictive model suitable for high-dimensional data [38]. Through optimization, this method aims to minimize the loss function. Consider a training dataset $\{(X_1, Y_1),\ldots,(X_n, Y_n)\}$, where the objective is to learn a function $F(X)$ for predicting dependent variables Y. Initially, a constant value $f_0(X)$ is initialize to compute residuals $\{(r_1 = Y_1 - f_0(X_1))\},\ldots,\{(r_n = Y_n - f_0(X_n))\}$. Subsequently, the model is updated by fitting a weak learner $h_1(X)$ to predict r for the current model. This process continues until the r can no longer be significantly reduced, incorporating new weak learners $h_n(X)$ into the r. The ultimate procedure is detailed in Eq. 7.

$$F(x) = f_0(X) + h_1(X) + h_2(X) +, \ldots, + h_n(X). \tag{7}$$

7) DYNAMIC SELECTION HYBRID MODEL (DSHM)

There are two possible approaches to choosing a hybrid classifier: static and dynamic. Static selection involves pre-selecting the classifier before the training phase. Subsequently, the ensemble formed is utilized for training and classifying all unseen data. On the contrary, dynamic approaches involve the selection of the ensemble of classifiers during the learning phase. A selection criterion is used to assess the base classifiers' competency for every training sample. To predict the label of the supplied test sample, only the classifier(s) that meet a predetermined degree of competence are used. Through dynamic ensemble selection, complex non-linear decision boundary classification problems can be addressed using only a few classifiers.

This research integrates a dynamic selection phase by employing an Explainable AI (EAI) technique known as Permutation Feature Importance (PFI) [40]. PFI serves as a model-independent global explanation method used for ranking features based on their influence on the predictions of trained ML models. We applied PFI to the six classifiers utilized in our study and assessed the significance of each feature selected by Univariate Feature Selection (UFS) and Information Gain Selection (IGS). Through this evaluation, we identified half of the classifiers demonstrating the highest positive impact on predictions. Subsequently, we employed four well-established ensemble methods, namely Boosting (BS), Bagging (BG), Voting (VT), and Stacking (ST), to construct a hybrid model with the dynamically selected classifiers. Our objective is to discern the most effective ensemble methods by evaluating their initial accuracy for the task.

Algorithm 1 depicted the procedure of our proposed DSHM. Where, initially, we trained six different classifiers (e.g., DT, SVM, KNN, RF, AB, and GB) and for each classifier, we measured the PFI rank on each feature, let as $PFI\left\{DT^{(X_{TR})}\right\}$, $PFI\left\{SVM^{(X_{TR})}\right\}$, $PFI\left\{KNN^{(X_{TR})}\right\}$, $PFI\left\{RF^{(X_{TR})}\right\}$, $PFI\left\{AB^{(X_{TR})}\right\}$, and $PFI\left\{GB^{(X_{TR})}\right\}$. Then we attempted to select the half-most efficient classifiers (HEC) with the highest PFI, as stated in Eq. 8, where B represents the

total conventional classifiers and X_{TR} is the training features of UFS and IGS.

$$HEC = B_i \in \left\{DT^{(X_{TR})}, \ldots, GB^{(X_{TR})}\right\}, \wedge \frac{B}{2}\left[argmax\left[PFI\left\{DT^{(X_{TR})}\right\}, \ldots, PFI\left\{GB^{(X_{TR})}\right\}\right]\right]. \quad (8)$$

In the second dynamic stage, we make a comparative evaluation between the ensemble methods based on their accuracy on UFS and IGS. We have selected four different ensemble methods (e.g., BS, BG, VT, and ST) and for each subset (UFS, IGS), we measured the initial accuracy, let as $ACC\left\{BS^{(X_{TR})}\right\}$, $ACC\left\{BG^{(X_{TR})}\right\}$, $ACC\left\{VT^{(X_{TR})}\right\}$, and $ACC\left\{VT^{(X_{TR})}\right\}$. Next, we select the most efficient ensemble method (EEM) with the highest accuracy, as stated in Eq. 9, where E represents the number of ensemble methods.

$$EEC = E_i \in \left\{BS^{(X_{TR})}, \ldots, ST^{(X_{TR})}\right\}, \wedge argmax\left[ACC\left\{BS^{(X_{TR})}\right\}, \ldots, ACC\left\{ST^{(X_{TR})}\right\}\right]. \quad (9)$$

From the numerical result analysis, we conclude that the ST ensemble methods outperformed others, showcasing the obtained results in the experimental result analysis section. Hence, we leveraged the collective power of multiple conventional classifiers using ST. The ST ensemble method is a potent strategy that integrates predictions from multiple models to enhance accuracy and resilience. In this method, a meta-model is trained based on the predictions generated by the base classifiers, let as HEC_1, HEC_2, and HEC_3. The individual predictions from these base classifiers serve as inputs for the meta-model, Eq. 10 stated the working procedures of these processes.

$$BE_{TE} = HEC_1(X_{TE}), HEC_2(X_{TE}), HEC_3(X_{TE}), \quad BE_{PR} = HEC_1(Y_{PR}), HEC_2(Y_{PR}), HEC_3(Y_{PR}). \quad (10)$$

where BE_{TE} and BE_{PR} represent the training of base estimators and prediction of base estimators, respectively. Logistic Regression (LG) is employed in this context to train the meta-model using the predictions from the base estimators. Finally, we apply BE_{PR} to classify X_{TR} as X_{NTR}, Eq. 11 states the procedure of meta model (MM), δ^0 is the intercept, and δ^1 to δ^n is the coefficient of the generated input features X^i_{NTR} to X^n_{NTR}.

$$MM = exp\left[-\left\{\delta^0 + \delta^1(X^1_{NTR}) + \ldots. + \delta^n(X^n_{NTR})\right\}\right] \quad (11)$$

The proposed method evaluates the outcome based on first-level prediction BE_{PR}, it used the first-level prediction as a new training set for MM. MM is trained on the outputs of BE_{PR}, allowing it to capture complex relationships and dependencies. The potential superiority of DSHM lies in its ability to adaptively learn and optimize the combination of diverse base models based on the characteristics of the data.

Algorithm 1 Showcasing Major Working Steps of DSHM.

Inputs: Dataset, $D = \sum_{i=1}^{M}(X_i, Y_i)$, Conventional Classifiers = CC, Ensemble Methods = EM.
Outputs: Classify whether the thyroid is affected or not.
$X_{TR}, Y_{TR}, X_{TS}, Y_{TS} \leftarrow TrainTestSplit(X_i, Y_i, 0.2)$
for $i = 1; i <= Number - of - CC; i++$ **do**
  $HEC_i \leftarrow PFI(CC(X_{TR}))$
end for
$HEC \leftarrow \frac{Number-of-CC}{2}\{argmax_i(HEC_i)\}$
for $i = 1; i <= Number - of - EM; i++$ **do**
  $EEC_i \leftarrow ACC(EM(X_{TR}))$
end for
$EEC \leftarrow argmax_i(EEC_i)$
while $(Train - different - HEC)$ **do**
  $BE^{(i)}_{TE} \leftarrow HEC_i(X_{TR}, Y_{TR})$
end while
while $(Pred - different - HEC)$ **do**
  $BE^{(i)}_{PR} \leftarrow HEC_i(X_{TS}, Y_{TS})$
end while
$BE_{PR} \leftarrow concatenate(BE^{(1)}_{PR}, \ldots, BE^{(n)}_{PR})$
for $i = 1; i <= M; i++$ **do**
  Apply BE_{PR} to classify X_{TR}
  $X_{NTR} \leftarrow BE_{PR}(X_{TR})$
end for
$MM = exp\left[-\left\{\delta^0 + \delta^1(X^1_{NTR}) + \ldots. + \delta^n(X^n_{NTR})\right\}\right]$
$Final_{PR} \leftarrow MM.predict(New - sample)$
Return $Final_{PR}$.

IV. EXPERIMENTAL ANALYSIS AND DISCUSSION

This section conducts a comprehensive evaluation of the experimental results obtained from our proposed methodology. In order to ensure a comprehensive examination, we have assessed multiple classification metrics for each of the three situations (All features, UFS-based features, and IGS-based features), such as accuracy, precision, recall, and f1-score [39], for all three scenarios (All features, UFS-based features, and IGS-based features). Additionally, we showcase the superiority of our proposed study through a comparative analysis between traditional classifiers (DT, SVM, KNN, RF, GB, and AB), a static hybrid model (combining ensemble methods BS, BG, and VT with all conventional classifiers, named A-BS, A-BG, and A-VT), and our proposed DSHM.

A. EXPERIMENTAL SETUP

The developed methods were constructed and prototyped using the cloud-based Jupyter Notebook environment (Colab NoteBook). The availability of numerous freely available and appropriate libraries for machine learning models, including Scikit-learn, Matplotlib, Keras, and others, led to this decision.

B. DYNAMIC PARAMETER ANALYSIS FOR PROPOSED DSHM

As detailed in the proposed methodology section, an initial comparative evaluation was conducted using the Permutation

TABLE 4. Parameter analysis in terms of the base estimators of our proposed DSHM.

Features	DT	SVM	KNN	RF	AB	GB
UFS-based Feature Subset						
T3	0.5072 ± 0.2335	0.5275 ± 0.2415	0.4575 ± 0.2213	0.6171 ± 0.2435	0.6314 ± 0.2645	0.6331 ± 0.2535
Age	0.0933 ± 0.1586	0.0813 ± 0.1308	0.0987 ± 0.1677	0.1583 ± 0.1960	0.1233 ± 0.1691	0.2243 ± 0.2301
TT4	0.0900 ± 0.0850	0.0408 ± 0.0378	0.0871 ± 0.0701	0.1103 ± 0.1055	0.1434 ± 0.1385	0.1159 ± 0.1102
T3 Measured	0.0145 ± 0.0267	0.0105 ± 0.0195	0.0095 ± 0.0109	0.0214 ± 0.0312	0.0199 ± 0.0301	0.0335 ± 0.0403
On Thyroxine	0.0185 ± 0.0245	0.0131 ± 0.0213	0.0208 ± 0.0319	0.0162 ± 0.0229	0.0275 ± 0.0383	0.0192 ± 0.0275
TSH Measured	0.0095 ± 0.0171	0.0108 ± 0.0189	0.0081 ± 0.0160	0.0185 ± 0.0252	0.0157 ± 0.0236	0.0203 ± 0.0293
TT4 Measured	0.0007 ± 0.0061	0.0004 ± 0.0053	0.0005 ± 0.0059	0.0055 ± 0.0110	0.0041 ± 0.0097	0.0069 ± 0.0131
Query Hyperthyroid	**0.0000 ± 0.0015**	**0.0000 ± 0.0009**	0.0002 ± 0.0033	0.0017 ± 0.0098	0.0033 ± 0.0103	0.0013 ± 0.0085
Psych	0.0070 ± 0.0121	0.0081 ± 0.0143	0.0034 ± 0.0091	0.0054 ± 0.0100	0.0060 ± 0.0109	0.0063 ± 0.0114
T4U	0.0878 ± 0.0920	0.0790 ± 0.0811	0.0778 ± 0.0803	0.0998 ± 0.1113	0.0799 ± 0.0855	0.1331 ± 0.1454
T4U Measured	0.0052 ± 0.0123	0.0016 ± 0.0081	0.0008 ± 0.0072	0.0041 ± 0.0106	0.0077 ± 0.0181	0.0092 ± 0.0196
Sex	0.0033 ± 0.0061	0.0045 ± 0.0098	0.0026 ± 0.0048	0.0130 ± 0.0156	0.0151 ± 0.0214	0.0145 ± 0.0192
FTI	0.0737 ± 0.0596	0.0523 ± 0.0424	0.0681 ± 0.0511	0.0737 ± 0.0596	0.0844 ± 0.0681	0.0922 ± 0.0723
Thyroid Surgery	**0.0000 ± 0.0000**	**0.0000 ± 0.0005**	**0.0000 ± 0.0003**	**0.0000 ± 0.0011**	0.0021 ± 0.0087	0.0015 ± 0.0073
IGS-based Feature Subset						
T3	0.5331 ± 0.2411	0.5512 ± 0.2487	0.4631 ± 0.2102	0.6414 ± 0.2511	0.6121 ± 0.2467	0.6635 ± 0.2627
Age	0.1001 ± 0.1702	0.0713 ± 0.1206	0.1123 ± 0.1831	0.1936 ± 0.2077	0.1421 ± 0.1798	0.2012 ± 0.2151
TT4	0.0838 ± 0.0915	0.0421 ± 0.0353	0.0920 ± 0.0782	0.1204 ± 0.1333	0.1337 ± 0.1411	0.1209 ± 0.1185
FTI	0.0836 ± 0.0926	0.0616 ± 0.0424	0.0703 ± 0.0585	0.0915 ± 0.0741	0.0829 ± 0.0648	0.0971 ± 0.0706
TSH	0.0316 ± 0.0382	0.0241 ± 0.0294	0.0211 ± 0.0282	0.0374 ± 0.0398	0.0343 ± 0.0359	0.0418 ± 0.0474
Sex	0.0102 ± 0.0183	0.0093 ± 0.0162	0.0110 ± 0.0194	0.0132 ± 0.0213	0.0121 ± 0.0195	0.0112 ± 0.0186
T3 Measured	0.0131 ± 0.0242	0.0105 ± 0.0197	0.0142 ± 0.0263	0.0169 ± 0.0307	0.0181 ± 0.0343	0.0175 ± 0.0321
TT4 Measured	0.0006 ± 0.0051	0.0011 ± 0.0076	**0.0000 ± 0.0011**	0.0009 ± 0.0067	0.0077 ± 0.0161	0.0052 ± 0.0131
On Thyroxine	0.0175 ± 0.0231	0.0202 ± 0.0282	0.0183 ± 0.0225	0.0232 ± 0.0301	0.0179 ± 0.0241	0.0193 ± 0.0284
Psych	0.0074 ± 0.0130	0.0070 ± 0.0128	0.0032 ± 0.0086	0.0034 ± 0.0070	0.0054 ± 0.0093	0.0063 ± 0.0114
TSH Measured	0.0065 ± 0.0143	0.0118 ± 0.0201	0.0087 ± 0.0171	0.0193 ± 0.0263	0.0177 ± 0.0281	0.0169 ± 0.0261
Pregnant	0.0021 ± 0.0093	0.0035 ± 0.0102	0.0007 ± 0.0028	0.0006 ± 0.0024	0.0049 ± 0.0134	0.0032 ± 0.0098
Tumor	**0.0000 ± 0.0003**	0.0005 ± 0.0043	0.0014 ± 0.0077	0.0002 ± 0.0016	0.0021 ± 0.0094	0.0037 ± 0.0099
Hypopituitary	0.0001 ± 0.0008	**0.0000 ± 0.0000**	**0.0000 ± 0.0003**	0.0013 ± 0.0042	0.0002 ± 0.0014	0.0015 ± 0.0073

Feature Importance (PFI) rank of features selected through Univariate Feature Selection (UFS) and Information Gain Selection (IGS). Table 4 presents the PFI rank results for six well-known machine learning classifiers employed in this evaluation. The first value in each rank signifies the extent to which the model's performance deteriorated due to random shuffling. This value is typically the mean or average of each feature rank, indicating the estimated importance or impact of the feature based on testing. The number following the ± sign represents the variation in performance across reshuffles, usually denoting the confidence interval or standard deviation. For example, the obtained rank for the feature T3 in IGS-based features is 0.6635 ± 0.2627 using the Gradient Boosting (GB) classifier, providing a central estimate of feature importance (0.6635) with an associated level of uncertainty (± 0.2627).

These mean and standard deviation values lead to the conclusion that features including T3, TT4, FTI, On Thyroxine, TSH, and T4U are the most relevant for the task. However, it is evident that for the UFS set, the features Query Hyperthyroid have no influence on the DT and SVM classifiers, and Thyroid Surgery has no impact on DT, SVM, KNN, and RF classifiers. Additionally, for the IGS set, TT4 Measured has no impact on the prediction of KNN, Tumor in DT, and Hypopituitary on SVM and KNN, these values are highlighted in the table. From this analysis, it is apparent that DT, SVM, and KNN classifiers have no impact on some of the features for this task, and RF is on the list once. Thereby we aim to select half of the classifiers [51], such as RF, AB, and GB, which have a consistent impact on the features of UFS and IGS (based on the PFI analysis), and these are chosen for the intended proposed DSHM.

Now we aim to combine their individual strengths, thereby conducting another analysis to explore the most effective ensemble method. For this evaluation, we trained four well-known ensemble methods (e.g., BS, BG, VT, and ST) and analyzed their obtained outcome for both UFS and IGS-based selected features. Table 5 states the accuracy of these ensemble methods on the selected feature sets. All the methods have achieved an acceptable accuracy for this task. However, ST outperformed with IGS-based selected features with an impressive 98.99% accuracy. Hence, we have selected the ST ensemble method for our proposed DSHM.

C. RESULT ANALYSIS

To demonstrate the significance of our proposed method over static hybrid and baseline classifier we have employed four classification metrics. Figure 2 showcases a comparative analysis of these classifiers in terms of their accuracy. This figure demonstrates the superiority of our proposed DSHM,

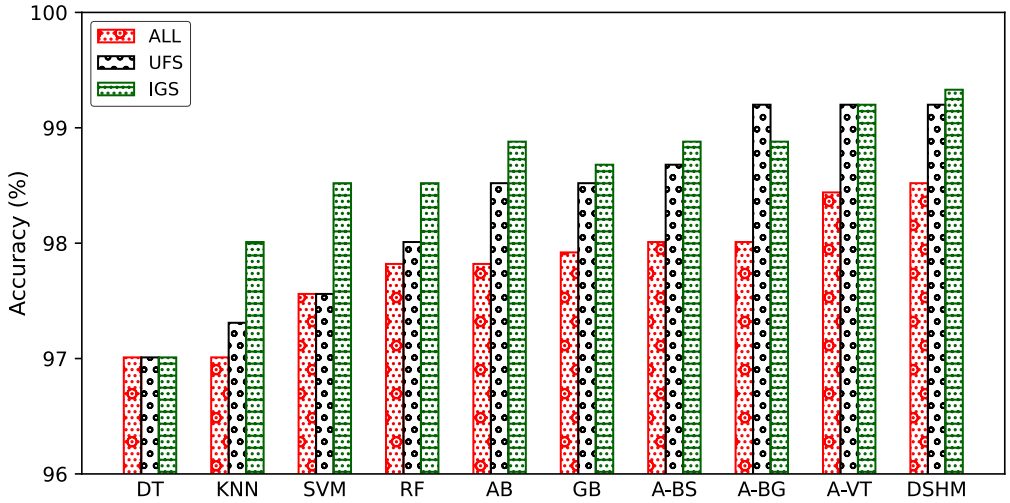

FIGURE 2. Accuracy measured for the conventional classifiers, static hybrid models, and proposed DSHM on ALL, UFS, and IGS features.

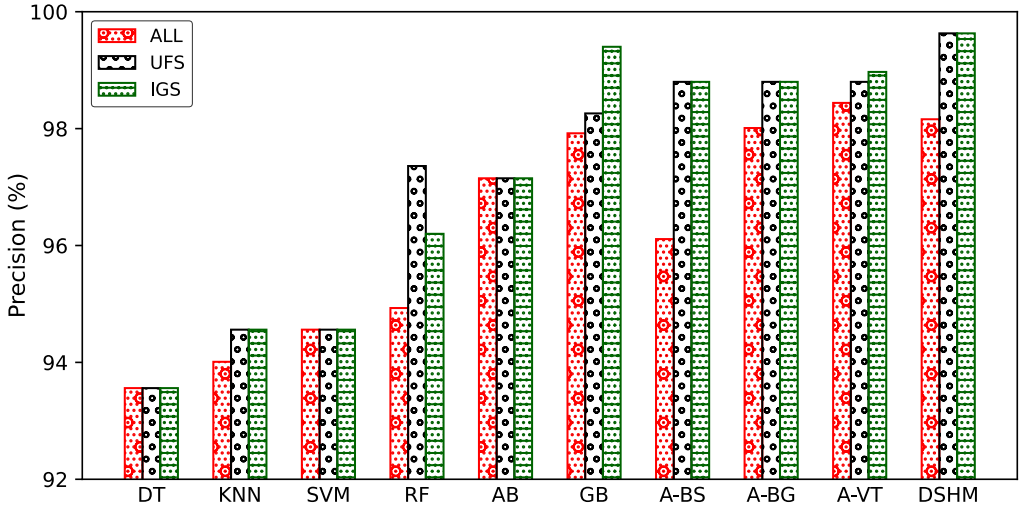

FIGURE 3. Precision measured for the conventional classifiers, static hybrid models, and proposed DSHM on ALL, UFS, and IGS features.

TABLE 5. Analysis for selecting the appropriate ensemble method based on their effectiveness.

BS		BG		VT		ST	
UFS	IGS	UFS	IGS	UFS	IGS	UFS	IGS
98.15	98.40	97.98	98.40	98.49	98.49	98.82	98.99

which outperformed with 99.33% accuracy with the IGS features set. In terms of UFS and ALL feature sets, DSHM also gains an impressive accuracy of 99.20% and 98.52%, respectively. Whereases, DT achieves the lowest accuracy of 97.01% for all different types of feature sets. In terms of the static hybrid model, A-VT gained 99.20% accuracy for both UFS and IGS-selected features.

Our second assessment metric is precision, a measure of the model's accuracy in positive predictions. Precision evaluates the trustworthiness of positive predictions and helps identify instances of false positives, allowing for necessary adjustments in performance. In Figure 3, when examining the UFS and IGS features set, the DSHM achieves the highest precision scores of 99.63% for both sets. In terms of ALL sets, it produced a 98.16% precision score. From the conventional classifiers, GB gained an extraordinary precision score of 99.40% for IGS-based selected features.

Then recall gauges a model's ability to accurately identify the number of positive instances among the total positive samples. Figure 4 illustrates the obtained recall scores for all the employed models on different feature sets. The proposed DSHM achieved recall scores of 97.67%, 98.37%, and 98.81% for ALL, UFS, and IGS feature sets, respectively. When considering the static models, A-BS and A-VT performed a generalized score of 97.59% with UFS-selected features. From the traditional models, the highest recall score is attained with AB, which gained a 98.15% recall score.

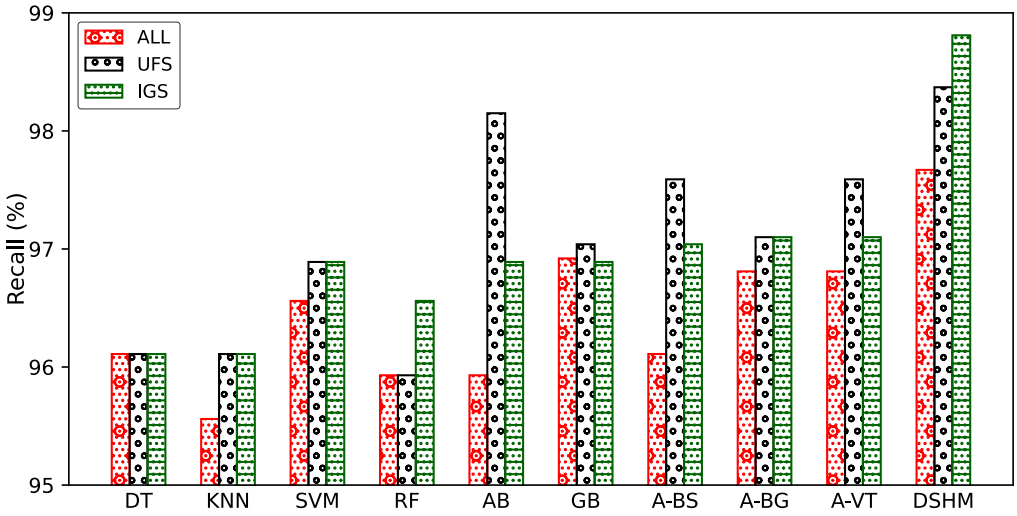

FIGURE 4. Recall measured for the conventional classifiers, static hybrid models, and proposed DSHM on ALL, UFS, and IGS features.

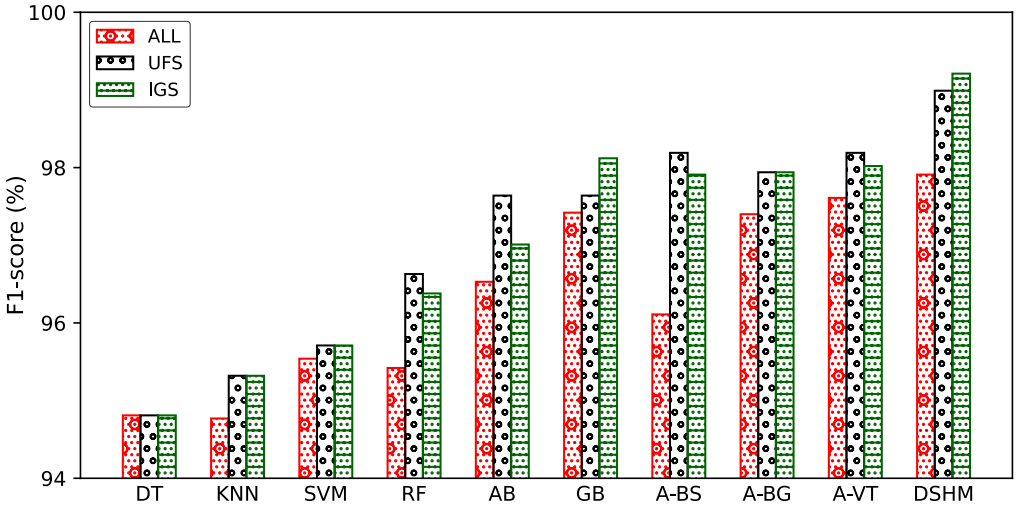

FIGURE 5. F1-score measured for the conventional classifiers, static hybrid models, and proposed DSHM on ALL, UFS, and IGS features.

Our final classification metric is the F1-score, which incorporates precision and recall scores, and evaluates how efficiently our proposed DSHM can predict outcomes. Figure 5 depicts the F1-score of the trained models on different feature sets. In comparison to other models, the DSHM achieved enhanced results on the selected feature sets. Specifically, the proposed model achieved the highest score of 99.21% with IGS features. Concerning the conventional models, GB scores exceed 98.12% with IGS-based selected features.

Moreover, to gain insight into the cost efficiency and deployment speed of our employed models, we have evaluated their compilation time on three different feature sets. Table 6 deprived their compilation times in milliseconds (MS), where DT has taken the overall lowest time for its execution, only 20.3, 18.7, and 18.2 MS for ALL, UFS, and IGS feature sets, respectively, whereas our proposed DSHM demands 451, 382, and 374 MS. However, the static hybrid models have taken the overall highest time, as they need to execute six different conventional classifiers.

D. OUTCOME EXPLANABILITY

In this section, we integrate explainable AI (XAI) in our proposed ML-based healthcare diagnosis models to provide transparency, accountability, validation, patient understanding, and clinical adoption. By incorporating XAI techniques, we can develop trustworthy and effective AI-driven solutions that improve patient outcomes while

TABLE 6. Reported the compilation time for all the employed classifiers on different feature subsets.

Features Set	DT	SVM	KNN	RF	AB	GB	A-BS	A-BG	A-VT	DSHM
ALL	20.3	325	129	62.3	68.7	219	726	897	982	451
UFS	18.7	284	122	63.1	53.6	188	704	813	834	382
IGS	18.2	243	130	56.8	54.2	173	684	842	781	374

TABLE 7. Outcome explanations of DSHM generation by LIME for a random positive and negative case.

Probability of positive prediction (100%)			Probability of negative prediction (99%)		
Negative reasons	Positive reasons	Actual value	Negative reasons	Positive reasons	Actual value
(-)	0.35 < T3 <=1.2	0.9	1.5 < T3 <= 2.3	(-)	2.1
(-)	Age > 52	57	TT4 > 116	(-)	118
(-)	FTI > 131	137	(-)	0 < Psych <= 1	1
(-)	0 < Sex <= 1	1	Age <= 40	(-)	40
(-)	0 < Psych <= 1	1	TSH <= 0.75	(-)	0.52
On Thyroxine <= 0	(-)	0	72.8 < FTI <= 124	(-)	109
(-)	0 < T3 Measured <= 1	1	(-)	0 < Pregnant <= 1	1
(-)	82.1 < TT4 <= 109	101	Sex <= 0	(-)	0
(-)	0 < Pregnant <= 1	1	TT4 Measured <= 0	(-)	0
TSH <= 0.75	(-)	0.7	T3 Measured <= 0	(-)	0
(-)	0 < TT4 Measured <= 1	1	0 < On Thyroxine <= 1	(-)	1
(-)	0 < Tumor <= 1	1	(-)	0 < Tumor <= 1	1

ensuring safety and fairness. To make such an implication, we have utilized an XAI technique named Local Interpretable Model-agnostic Explanations (LIME), which make explanations for individual predictions made by complex ML models. LIME selects an instance, generates perturbed samples around it, and learns a local interpretable model based on the black-box model's predictions on these samples. It then generates explanations by analyzing the interpretable model's coefficients or rules, providing insights into how each feature contributes to the black-box model's prediction for the selected instance. Finally, it's important to evaluate the quality and trustworthiness of the explanations generated by LIME through validation and testing.

Table 7 presents the LIME-generated prediction probabilities and explanations for two randomly selected data samples (one positive and one negative) from the IGS-based selected feature sets (as this feature set outperforms others). The real values of each characteristic are displayed in the ''actual value'' column, whilst the LIME-generated values in the ''Negative reasons'' and ''Positive reasons'' sections indicate whether a feature has a positive or negative impact on prediction probabilities. For example, suppose a feature negatively affects a sample. In that case, its name and recommended value ranges are entered into the ''Negative reasons'' field, and a positive influence is stated in the ''Positive reasons'' field. In the case of a random positive sample, our proposed DSHM model predicts a 100% probability of having thyroid disease. The feature ''T3'' makes the most important contribution to this positive forecast, with its actual value falling somewhere between 0.35 and 1.2, for example, 0.9. Other characteristics, such as ''Age,'' ''FTI,'' ''Sex,'' and so on, also play an important role in good prediction. In the case of the Negative prediction, DSHM forecasts a 99% chance of not having thyroid disease. Again, ''T3'' emerges as the most relevant factor in forecasting, with an actual value of 2.1 falling within the recommended range of 1.5-2.3. Additionally, other feature values such as ''TT4,'' ''Psych,'' ''Age,'' ''TSH,'' ''FTI,'' ''Sex,'' ''TT4 Measured,'' ''T3 Measured,'' and ''On Thyroxine'' also contribute to the negative prediction.

Subsequently, we also consider evaluating the global behavior of the IGS-based features set, understanding the underlying structure of data, validating model performance, and informing decision-making processes. For that, we used Shapley Additive exPlanations (SHAP) on this feature set. SHAP is an XAI technique that provides insights into the importance and contributions of each feature to the output of an ML model. SHAP values are based on Shapley values from cooperative game theory and offer a unified framework for understanding the impact of features on model predictions. Figure 6 illustrates the global behavior of the IGS-based selected features set, showcasing how factors affect predictions on a global scale. The plot displays higher-contributing features at the top, whereas blue, purple, and red indicate low, moderate, and high feature values. Higher feature values (red or purple dots) indicate a lower likelihood of thyroid disease, with particularly negative SHAP values. Blue dots with lower feature values typically imply higher illness risk, as evidenced by positive SHAP values. This figure demonstrates how T3, FTI, TSH, TT4, Age, On Thyroxine, Sex, T3 Measured, and Psych are the most influencing characteristics of this thyroid diagnostic model and significantly affect the dataset's global behaviors, similar to local explanations. However, the features of TSH Measured and hypopituitary do not have such a significant influence on these XAI models. Hence

TABLE 8. Evaluate the statistical significance between the two models using the Mann-Whitney U Statistical Test.

Models	Models (α = 0.05)									
	DT	KNN	SVM	RF	AB	GB	A-BS	A-BG	A-VT	DSHM
DT	(-)	1.5, 0.19	**0, 0.04**	0, 0.06	0, 0.06	0, 0.06	0, 0.06	0, 0.06	**0, 0.04**	**0, 0.04**
KNN	1.5, 0.19	(-)	2, 0.37	1.5, 0.26	1, 0.2	1, 0.2	0.5, 0.12	0.5, 0.12	0, 0.07	**0, 0.04**
SVM	0, 0.04	2, 0.37	(-)	2.5, 0.5	1.5, 0.26	1.5, 0.26	1, 0.18	1, 0.18	1, 0.17	**0.5, 0.05**
RF	0, 0.06	1.5, 0.26	2.5, 0.5	(-)	3, 0.65	2.5, 0.5	1.5, 0.26	1.5, 0.26	1, 0.18	0.5, 0.12
AB	0, 0.06	1, 0.2	1.5, 0.26	3, 0.65	(-)	4.5, 1	3.5, 0.82	2.5, 0.5	2, 0.37	1.5, 0.26
GB	0, 0.06	1, 0.2	1.5, 0.26	2.5, 0.5	4.5, 1	(-)	2.5, 0.5	2, 0.4	2, 0.37	1.5, 0.26
A-BS	0, 0.06	0.5, 0.12	1, 0.18	1.5, 0.26	3.5, 0.82	2.5, 0.5	(-)	3, 0.65	2, 0.37	2, 0.4
A-BG	0, 0.06	0.5, 0.12	1, 0.18	1.5, 0.26	2.5, 0.5	2, 0.4	3, 0.65	(-)	3, 0.64	2.5, 0.5
A-VT	**0, 0.04**	0, 0.07	1, 0.17	1, 0.18	2, 0.37	2, 0.37	2, 0.37	3, 0.64	(-)	3, 0.64
DSHM	**0, 0.04**	**0, 0.04**	**0.5, 0.05**	0.5, 0.12	1.5, 0.26	1.5, 0.26	2, 0.4	2.5, 0.5	3, 0.64	(-)

TABLE 9. Evaluate the generalization of our proposed model using an independent dataset.

ALL features				UVS features				IGS features			
Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
98.64	98.88	98.37	98.62	98.93	99.17	98.77	98.96	99.15	98.98	99.33	99.15

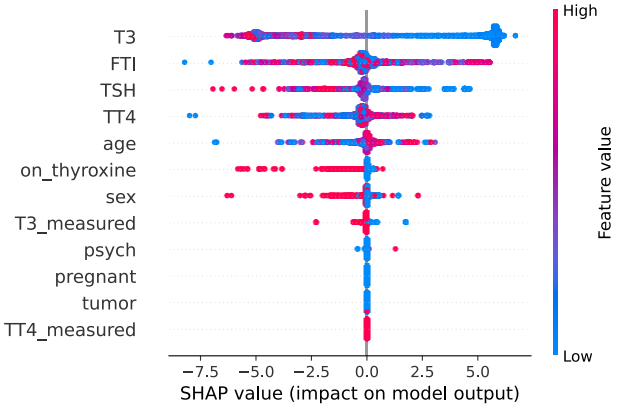

FIGURE 6. Display the global behavior of the high contributory features set.

we discarded these features to the LIME and SHAP-based interpretation.

E. PAIR-WISE STATISTICAL TEST OF EMPLOYED CLASSIFIERS

In this subsection, we perform a statistical test to analyze the significance of our proposed method. For this evaluation, a Mann-Whitney U statistic [41] is applied to the accuracy of different feature sets, determining whether there is a difference between two employed models based on their performance. The total ranks for one of the groups in the comparison are represented by the U-Test. A lower U-Test value suggests that the first group's value distribution is typically higher than the second group's value distribution. The incredibly small p-value points to a highly significant outcome, suggesting that the null hypothesis—that there is no difference between the two groups—is strongly supported by the evidence. The significance level α, which denotes the likelihood of rejecting a true null hypothesis, or Type I error, is set to 0.05 in this case. Stated differently, the significance level denotes the cutoff point utilized to ascertain the statistical significance of the observed outcomes. Table 8 shows the P-value and U-Test; the first value indicates the U-Test, while the second value indicates the P-value. This table shows that for DSHM with the pair of DT, KNN, and A-VT, the total rank is 0. It usually means that the two samples are perfectly separated from one another. Since the p-value is nearly zero, there is substantial evidence to refute the null hypothesis. The incredibly small P-values of 0.05 and 0.04 in various observations point to a highly significant outcome, providing compelling evidence to reject the null hypothesis that the two models' executed outcomes show a statistically significant difference. For clear observations, we have highlighted these pairs which are statistically significant to each other.

F. DISCUSSION

Our proposed DSHM proves its significance in early-stage disease prediction by achieving superior scores across various classification metrics. Nevertheless, the model's generalizability could be affected by unforeseen data circumstances, leading to challenges such as over-fitting and under-fitting in classification models. Over-fitting occurs when a function closely matches a limited number of data points, while under-fitting arises when the model struggles to accurately map inputs and outputs during training, resulting in significant training errors. Indications of under-fitting manifest as high bias and low variance during the training process. To address these challenges from our proposed DSHM, we have considered various mechanisms before training the model. First, we implemented various preprocessing techniques to clean the dataset. Additionally, we tackled class imbalance using the BOO-ST method to achieve a balanced distribution and selected relevant features to enhance model performance. These measures aim to reduce the risk of over-fitting and improve the generalizability of the model [22]. Subsequently, we developed the proposed

TABLE 10. Comparative analysis between existing and our study in terms of core methodologies and predictive performance.

Ref and year	Data source	Data balancing method	Best Performing classifier	Hybrid selection method	Perform accuracy
Sultana [6] 2023	University of California Irvine	SMOTE	RF	(-)	99%
Haneet [7] 2023	University of California Irvine	SMOTE	LDA	Static	85.45%
Saima [8] 2022	University of California Irvine	SMOTE	ANN	(-)	95.87%
Ege [9] 2022	University of California Irvine	(-)	ANN	(-)	98%
Olatunji [10] 2021	King Fahad Specialist Hospital	(-)	RF	(-)	90.91%
Tahir [11] 2022	University of California Irvine	SMOTE	RF	(-)	94.80%
Diganta [12] 2023	University of California Irvine	(-)	RF	(-)	99.14%
Rajasekhar [13] 2022	University of California Irvine	(-)	RF	(-)	99%
Ramazan [14] 2020	University of California Irvine	(-)	Hybrid (AB and GB)	Static	99.08%
Yufei [15] 2023	University of California Irvine	(-)	ST	Static	92.3%
Bhavna [16] 2020	University of California Irvine	(-)	Hybrid (C4.5 and RF)	Static	96%
Yadav [18] 2020	Chandan Diagonosis Center, Jaunpur	(-)	ST	Static	99%
Ananda [27] 2023	Kaggle	SMOTE	3SHC	Static	99.2%
Our Study	Kaggle	BOO-ST	DSHM	Dynamic	99.33%

model by integrating multiple preferable baseline classifiers using an effective ensemble method. This integration helps mitigate individual biases and captures diverse perspectives, thereby alleviating over-fitting [24]. Moreover, the baseline classifiers within the ensemble were trained with a fine-tuned set of parameters to effectively control the learning process. These approaches render our proposed model less susceptible to over-fitting, ensuring they produce more generalized results [27]. Furthermore, we have used one more dataset to evaluate the generalization of our proposed model, which is publicly available in [42]. Table 9 represents the performed accuracy, precision, recall, and f1-score by our proposed DSHM model using this dataset. Where our proposed model obtained a robust accuracy of 99.15% with an IGS-based selected features set. The outcomes are also superior in terms of other performance indicator metrics. This table demonstrates the generalization of our proposed model across an independent dataset. Finally, a comparison is shown in Table 10 between our model and existing thyroid predictive models, where our models surpass the outcome of existing models. These findings lend support to our assertion and improve the credibility of the proposed strategy.

V. CONCLUSION

As the prevalence of thyroid disease continues to increase globally, predicting the illness and categorizing patients has become a more challenging task for practitioners. In response to this growing impact, we have introduced a Machine Learning-based disease prediction system leveraging the most essential features. The utilization of the BOO-ST method has been crucial in preparing a balanced dataset for our experiments. Experimental analysis underscores the pivotal role of the proposed DSHM, particularly when utilizing IGS-based selected features. However, it is essential to note that the proposed DSHM demands higher computational resources and incurs increased computational costs compared to traditional single classifiers. Hence in future studies, we plan to investigate addressing this challenge, one potential solution is the exploration of a distributed learning mechanism. Additionally, we will explore the different ways to select the more appropriate baseline models in a dynamic environment. Moreover, we envision integrating our approach into a blockchain network, aiming to enhance the security and accessibility of information across various healthcare settings, including hospitals and clinics.

REFERENCES

[1] G. Bereda, "Definition, causes, pathophysiology, and management of hypothyroidism," *Mathews J. Pharmaceutical Sci.*, vol. 7, no. 1, pp. 1–5, Jan. 2023.
[2] K. Guleria, S. Sharma, S. Kumar, and S. Tiwari, "Early prediction of hypothyroidism and multiclass classification using predictive machine learning and deep learning," *Meas., Sensors*, vol. 24, Dec. 2022, Art. no. 100482.
[3] P. V. Voulgari, A. I. Venetsanopoulou, N. Kalpourtzi, M. Gavana, A. Vantarakis, C. Hadjichristodoulou, G. Chlouverakis, G. Trypsianis, Y. Alamanos, and G. Touloumi, "Thyroid dysfunction in greece: Results from the national health examination survey EMENO," *PLoS ONE*, vol. 17, no. 3, Mar. 2022, Art. no. e0264388.

[4] N. K. Singh, N. Hage, S. Prabhala, B. Ramamourthy, S. Nagaraju, and K. M. Kappagantu, "Probable impact of environmental radiation on thyroid swellings in areas of Eastern Hyderabad and Nalgonda," *Egyptian J. Otolaryngol.*, vol. 39, no. 1, p. 150, Sep. 2023.

[5] M. H. Alshayeji, "Early thyroid risk prediction by data mining and ensemble classifiers," *Mach. Learn. Knowl. Extraction*, vol. 5, no. 3, pp. 1195–1213, Sep. 2023.

[6] A. Sultana and R. Islam, "Machine learning framework with feature selection approaches for thyroid disease classification and associated risk factors identification," *J. Electr. Syst. Inf. Technol.*, vol. 10, no. 1, pp. 1–23, Jun. 2023.

[7] H. Kour, B. Singh, N. Gupta, J. Manhas, and V. Sharma, "Bagged based ensemble model to predict thyroid disorder using linear discriminant analysis with SMOTE," *Res. Biomed. Eng.*, vol. 39, no. 3, pp. 733–746, Aug. 2023.

[8] S. S. Islam, M. S. Haque, M. S. U. Miah, T. B. Sarwar, and R. Nugraha, "Application of machine learning algorithms to predict the thyroid disease risk: An experimental comparative study," *PeerJ Comput. Sci.*, vol. 8, p. e898, Mar. 2022.

[9] E. Savci and F. Nuriyeva, "Diagnosis of thyroid disease using machine learning techniques," *J. Modern Technol. Eng.*, vol. 7, no. 2, pp. 134–145, 2022.

[10] S. O. Olatunji, S. Alotaibi, E. Almutairi, Z. Alrabae, Y. Almajid, R. Altabee, M. Altassan, M. I. B. Ahmed, M. Farooqui, and J. Alhiyafi, "Early diagnosis of thyroid cancer diseases using computational intelligence techniques: A case study of a Saudi Arabian dataset," *Comput. Biol. Med.*, vol. 131, Apr. 2021, Art. no. 104267.

[11] T. Alyas, M. Hamid, K. Alissa, T. Faiz, N. Tabassum, and A. Ahmad, "Empirical method for thyroid disease classification using a machine learning approach," *BioMed Res. Int.*, vol. 2022, pp. 1–10, Jun. 2022.

[12] D. Sengupta, S. Mondal, A. Raj, and A. Anand, "Binary classification of thyroid using comprehensive set of machine learning algorithm," in *Frontiers of ICT in Healthcare*, Singapore. Springer, 2023, pp. 265–276.

[13] R. Chaganti, F. Rustam, I. De La Torre Díez, J. L. V. Mazón, C. L. Rodríguez, and I. Ashraf, "Thyroid disease prediction using selective features and machine learning techniques," *Cancers*, vol. 14, no. 16, p. 3914, Aug. 2022.

[14] R. Solmaz, A. Alkan, and M. Gunay, "Mobile diagnosis of thyroid based on ensemble classifier," *Dicle Univ. J. Eng.*, vol. 11, pp. 915–924, Sep. 2020.

[15] Y. Xie, W. Yu, S. Song, W. Wang, W. Gao, Y. Jia, S. Wen, C. Wang, and S. Wang, "Thyroid disease diagnosis based on feature interpolation and dynamic weighting ensemble model," Tech. Rep., 2023.

[16] B. Dharamkar, P. Saurabh, R. Prasad, and P. Mewada, "An ensemble approach for classification of thyroid using machine learning," in *Progress in Computing, Analytics and Networking*. Singapore: Springer, 2020, pp. 13–22.

[17] T. Akhtar, S. O. Gilani, Z. Mushtaq, S. Arif, M. Jamil, Y. Ayaz, S. I. Butt, and A. Waris, "Effective voting ensemble of homogenous ensembling with multiple attribute-selection approaches for improved identification of thyroid disorder," *Electronics*, vol. 10, no. 23, p. 3026, Dec. 2021.

[18] D. C. Yadav and S. Pal, "Thyroid prediction using ensemble data mining techniques," *Int. J. Inf. Technol.*, vol. 14, no. 3, pp. 1273–1283, May 2022.

[19] Z. Jiang, T. Pan, C. Zhang, and J. Yang, "A new oversampling method based on the classification contribution degree," *Symmetry*, vol. 13, no. 2, p. 194, Jan. 2021.

[20] S. Lu, H. Chai, A. Sahoo, and B. T. Phung, "Condition monitoring based on partial discharge diagnostics (don't short) using machine learning methods: A comprehensive state-of-the-art review," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 27, no. 6, pp. 1861–1888, Dec. 2020.

[21] R. M. O. Cruz, R. Sabourin, and G. D. C. Cavalcanti, "META-Des.Oracle: Meta-learning and feature selection for dynamic ensemble selection," *Inf. Fusion*, vol. 38, pp. 84–103, Nov. 2017.

[22] A. Sutradhar, M. Al Rafi, F. M. J. M. Shamrat, P. Ghosh, S. Das, M. A. Islam, K. Ahmed, X. Zhou, A. K. M. Azad, S. A. Alyami, and M. A. Moni, "BOO-ST and CBCEC: Two novel hybrid machine learning methods aim to reduce the mortality of heart failure patients," *Sci. Rep.*, vol. 13, no. 1, p. 22874, Dec. 2023.

[23] H. Kaur, H. S. Pannu, and A. K. Malhi, "A systematic review on imbalanced data challenges in machine learning: Applications and solutions," *ACM Comput. Surv.*, vol. 52, no. 4, pp. 1–36, Jul. 2020.

[24] A. Sutradhar, M. Al Rafi, M. J. Alam, and S. Islam, "An early warning system of heart failure mortality with combined machine learning methods," *Indonesian J. Electr. Eng. Comput. Sci.*, vol. 32, no. 2, pp. 1115–1122, Nov. 2023.

[25] *Thyroid Dataset*. Accessed: 2023. [Online]. Available: https://www.kaggle.com/datasets/bidemiayinde/thyroid-sickness-determination

[26] A. Raghuvanshi, U. K. Singh, G. S. Sajja, H. Pallathadka, E. Asenso, M. Kamal, A. Singh, and K. Phasinam, "Intrusion detection using machine learning for risk mitigation in IoT-enabled smart irrigation in smart farming," *J. Food Qual.*, vol. 2022, pp. 1–8, Feb. 2022.

[27] A. Sutradhar, M. Al Rafi, P. Ghosh, F. M. J. M. Shamrat, M. Moniruzzaman, K. Ahmed, A. Azad, F. M. Bui, L. Chen, and M. A. Moni, "An intelligent thyroid diagnosis system utilising multiple ensemble and explainable algorithms with medical supported attributes," *IEEE Trans. Artif. Intell.*

[28] L. Aversano, M. L. Bernardi, M. Cimitile, M. Iammarino, P. E. Macchia, I. C. Nettore, and C. Verdone, "Thyroid disease treatment prediction with machine learning approaches," *Proc. Comput. Sci.*, vol. 192, pp. 1031–1040, Jan. 2021.

[29] F. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves, "Data imbalance in classification: Experimental evaluation," *Inf. Sci.*, vol. 513, pp. 429–441, Mar. 2020.

[30] Z. Wang, C. Wu, K. Zheng, X. Niu, and X. Wang, "SMOTETomek-based resampling for personality recognition," *IEEE Access*, vol. 7, pp. 129678–129689, 2019.

[31] E.-S. M. El-Kenawy and M. Eid, "Hybrid gray wolf and particle swarm optimization for feature selection," *Int. J. Innov. Comput. Inf. Control*, vol. 16, no. 3, pp. 831–844, 2020.

[32] M. Alirezanejad, R. Enayatifar, H. Motameni, and H. Nematzadeh, "Heuristic filter feature selection methods for medical datasets," *Genomics*, vol. 112, no. 2, pp. 1173–1181, Mar. 2020.

[33] S. B. Kotsiantis, "Decision trees: A recent overview," *Artif. Intell. Rev.*, vol. 39, no. 4, pp. 261–283, Apr. 2013.

[34] X. Ding, J. Liu, F. Yang, and J. Cao, "Random radial basis function kernel-based support vector machine," *J. Franklin Inst.*, vol. 358, no. 18, pp. 10121–10140, Dec. 2021.

[35] G. Muhammad, S. Naveed, L. Nadeem, T. Mahmood, A. R. Khan, Y. Amin, and S. A. O. Bahaj, "Enhancing prognosis accuracy for ischemic cardiovascular disease using K nearest neighbor algorithm: A robust approach," *IEEE Access*, vol. 11, pp. 97879–97895, 2023.

[36] P. Ghosh, S. Azam, M. Jonkman, A. Karim, F. M. J. M. Shamrat, E. Ignatious, S. Shultana, A. R. Beeravolu, and F. De Boer, "Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques," *IEEE Access*, vol. 9, pp. 19304–19326, 2021.

[37] W. Hu, W. Hu, and S. Maybank, "AdaBoost-based algorithm for network intrusion detection," *IEEE Trans. Syst., Man, Cybern., B*, vol. 38, no. 2, pp. 577–583, Apr. 2008.

[38] B. Ma, F. Meng, G. Yan, H. Yan, B. Chai, and F. Song, "Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data," *Comput. Biol. Med.*, vol. 121, Jun. 2020, Art. no. 103761.

[39] A. Sutradhar, S. Tajmen, A.-A. Dhaly, F. M. J. M. Shamrat, M. S. R. Talukder, and A. Khater, "Skin cancer classification and early detection on cell images using multiple convolution neural network architectures," in *Proc. 3rd Int. Conf. Smart Electron. Commun. (ICOSEC)*, Oct. 2022, pp. 1089–1094.

[40] H. Kaneko, "Cross-validated permutation feature importance considering correlation between features," *Anal. Sci. Adv.*, vol. 3, nos. 9–10, pp. 278–287, Oct. 2022.

[41] T. W. MacFarland, J. M. Yates, and T. W. MacFarland, "Mann-whitney U test," in *Introduction to Nonparametric Statistics for the Biological Sciences Using R*, 2016, pp. 103–132.

[42] *Thyroid Dataset*. Accessed: 2023. [Online]. Available: https://archive.ics.uci.edu/dataset/102/thyroid+dataset

[43] S. Akbar, A. Raza, T. A. Shloul, A. Ahmad, A. Saeed, Y. Y. Ghadi, O. Mamyrbayev, and E. Tag-Eldin, "PAtbP-EnC: Identifying anti-tubercular peptides using multi-feature representation and genetic algorithm-based deep ensemble model," *IEEE Access*, vol. 11, pp. 137099–137114, 2023.

[44] S. Akbar, M. Hayat, M. Tahir, S. Khan, and F. K. Alarfaj, "CACP-DeepGram: Classification of anticancer peptides via deep neural network and skip-gram-based word embedding model," *Artif. Intell. Med.*, vol. 131, Sep. 2022, Art. no. 102349.
[45] F. Ali, S. Akbar, A. Ghulam, Z. A. Maher, A. Unar, and D. B. Talpur, "AFP-CMBPred: Computational identification of antifreeze proteins by extending consensus sequences into multi-blocks evolutionary information," *Comput. Biol. Med.*, vol. 139, Dec. 2021, Art. no. 105006.
[46] A. Raza, J. Uddin, A. Almuhaimeed, S. Akbar, Q. Zou, and A. Ahmad, "AIPs-SnTCN: Predicting anti-inflammatory peptides using fastText and transformer encoder-based hybrid word embedding with self-normalized temporal convolutional networks," *J. Chem. Inf. Model.*, vol. 63, no. 21, pp. 6537–6554, Nov. 2023.
[47] S. Akbar, A. Raza, and Q. Zou, "Deepstacked-AVPs: Predicting antiviral peptides using tri-segment evolutionary profile and word embedding based multi-perspective features with deep stacking model," *BMC Bioinf.*, vol. 25, no. 1, p. 102, Mar. 2024.
[48] S. Akbar, Q. Zou, A. Raza, and F. K. Alarfaj, "IAFPs-Mv-BiTCN: Predicting antifungal peptides using self-attention transformer embedding and transform evolutionary based multi-view features with bidirectional temporal convolutional networks," *Artif. Intell. Med.*, vol. 151, May 2024, Art. no. 102860.
[49] Z. Liu, P. Wei, J. Jiang, W. Cao, J. Bian, and Y. Chang, "MESA: Boost ensemble imbalanced learning with meta-sampler," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 14463–14474.
[50] J. Gómez-Ramírez, M. Ávila-Villanueva, and M. Á. Fernández-Blázquez, "Selecting the most important self-assessed features for predicting conversion to mild cognitive impairment with random forest and permutation-based methods," *Sci. Rep.*, vol. 10, no. 1, p. 20630, Nov. 2020.
[51] F. Li, W. Lu, J. W. Keung, X. Yu, L. Gong, and J. Li, "The impact of feature selection techniques on effort-aware defect prediction: An empirical study," *IET Softw.*, vol. 17, no. 2, pp. 168–193, Apr. 2023.
[52] W.-H. Hou, X.-K. Wang, H.-Y. Zhang, J.-Q. Wang, and L. Li, "A novel dynamic ensemble selection classifier for an imbalanced data set: An application for credit risk assessment," *Knowl.-Based Syst.*, vol. 208, Nov. 2020, Art. no. 106462.

MAJDI KHALID received the Ph.D. degree from Colorado State University, Fort Collins, CO, USA, in 2019. He is currently an Associate Professor with the Department of Computer Science and Artificial Intelligence, College of Computing, Umm Al-Qura University, Makkah, Saudi Arabia. His research interests include machine learning, deep learning, and computer vision.

F. M. JAVED MEHEDI SHAMRAT (Member, IEEE) received the B.Sc. degree from the Department of Software Engineering, Daffodil International University (DIU). He is currently pursuing the Master of Computer Science degree with the Department of Computer System and Technology, Universiti Malaya, Federal Territory of Kuala Lumpur, Malaysia. He was a Lecturer with the Department of Computer Science and Engineering, European University of Bangladesh. He is a Formal Research Associate with DIU. He has more than 65 publications in IEEE, Springer, Elsevier, and PubMed-indexed journals. His primary research interests include the intersection of the Internet of Things, deep learning, data science, image processing, neural networks, artificial intelligence, bioinformatics, and machine learning. He is an Associate Member of Bangladesh Computer Society. He has achieved the Best Student Research Award for the Department of Software Engineering from DIU. Also, he has enlisted among the researchers in Bangladesh on the AD Science Index.

HANAN ALSHANBARI received the Ph.D. degree from Coventry University, U.K., in 2018. She is currently an Associate Professor with the Department of Computer Science and Artificial Intelligence, College of Computing, Umm Al-Qura University, Makkah, Saudi Arabia. Her research interests include computer vision, with a particular emphasis on medical imaging.

MAJED FARRASH received the Ph.D. degree from the University of East Anglia, U.K., in 2016. He is currently an Assistant Professor with the Department of Computer Science and Artificial Intelligence, College of Computing, Umm Al-Qura University, Makkah, Saudi Arabia. His research interest includes various applications of artificial intelligence.

THAMIR M. QADAH (Member, IEEE) received the Ph.D. degree from Purdue University, West Lafayette, IN, USA, in 2021. He is currently an Assistant Professor with the Computer and Network Engineering Department, College of Computing, Umm Al-Qura University, Makkah, Saudi Arabia. His research interests include designing and implementing secure, dependable, and high-performance software systems that exploit modern hardware technologies and cloud infrastructures. Moreover, he served as a Committee Member for the Artifact Evaluation Committee of ASPLOS, OSDI, and SOSP. His research on queue-oriented transaction processing was recognized with the Best Paper Award in Middleware'18. Since 2015, he has been serving the research community as a Reviewer for top-tier conferences, such as SIGMOD, VLDB, ICDE, ICDCS, ATC, EDBT, Middleware, and CIKM, and a Reviewer for IEEE Access.