An Intelligent Thyroid Diagnosis System Utilizing Multiple Ensemble and Explainable Algorithms With Medical Supported Attributes

Ananda Sutradhar, Mustahsin Al Rafi, Pronab Ghosh, F M Javed Mehedi Shamrat, *Member, IEEE*, Md. Moniruzzaman, Kawsar Ahmed, *Member, IEEE*, AKM Azad, Francis M. Bui, *Member, IEEE*, Li Chen, *Member, IEEE*, and Mohammad Ali Moni

***Abstract*—The widespread impact of thyroid disease and its diagnosis is a challenging task for healthcare experts. The conventional technique for predicting such a vital disease is complex and time-consuming. A data-driven approach may offer predictive solutions, but it relies on all relevant attributes, which are computationally expensive. Hence, we propose a novel machine learning (ML) based disease prediction system that could potentially predict it by considering three crucial steps. First, to reduce the dimension of the dataset, three feature selection techniques were employed, including feature importance (FIS), information gain selections (IGS), and least absolute shrinkage and selection operator (LAS). Moreover, recommended medical references were considered while developing a feature set having the identical attributes as high-risk factors (HRF). Second, the models, including the three stage hybrid classifier (*3SHC*) and the three stage hybrid artificial neural network (*3SHANN*), are used as classifiers on the training data set. Third, a local interpretable model-agnostic explanations (LIME) to the *3SHC* with the HRF samples was applied to individually explain the predictions. Then, the overall behaviors of both gender and age categories were explored with the help of a partial dependence plot (PDP). Finally, the proposed system is validated with extensive experiments, where the *3SHC* achieves an accuracy (ACC) of 99.29%, which can play a crucial role in preventing thyroid disease and alleviating stress in the healthcare sector.**

***Impact Statement*—Artificial intelligence plays a crucial role in the healthcare system. Hence, machine learning algorithms could efficiently detect thyroid diseases early and help save lives. In this work, the proposed Three Stage Hybrid Classifier (3SHC) and Three Stage Hybrid Artificial Neural Network (3SHANN) significantly reduce the overfitting and underfitting issues due to the functionality of training models. However, the proposed 3SHC method achieves an accuracy of 99.29%, which outperforms the state-of-the-art models and shows that aged female people are more vulnerable to thyroid disease, significantly proving the existing literature. Also, the proposed method can be performed on a single CPU, efficiently solving the computational power limitation. Moreover, Local Interpretable Model-agnostic Explanations (LIME) are fitted with the classifier and features to explain the predicted outcomes and generate individual explanations. Thus, this method could be integrated into the Blockchain network in the healthcare system to preserve and exchange patient data through hospitals, diagnostic laboratories, pharmacy firms, and physicians. Furthermore, this will encourage other AI researchers to explore different methods for disease detection.**

***Index Terms*—Local interpretable model-agnostic explanations (LIME), high-risk factor, random forest (RF), three stage hybrid artificial neural network (3SHANN), three stage hybrid classifier (3SHC).**

I. Introduction

THYROID is one of the most prevalent diseases across the globe that occurs when the thyroid gland fails to produce a substantial amount of the relevant hormone. People with this disease experience specific symptoms that include but are not limited to laziness, weight gain, quickened heart rate, dry skin, and hair loss [1]. The prevalence of it within the general population is in the range of 0.3% to 3.7% in the USA and 0.2% to 3.5% in Europe, which gradually increases yearly [2]. According to the American Cancer Society, in 2023, an estimated 2120 new deaths have already occurred in the USA due to the abnormal growth of cells in the thyroid gland.

Preventing this disease by diagnosing or predicting it at an early stage is one of the ways to avoid the severe consequences it may have on one's life. However, diagnosing this disease from

Manuscript received 10 April 2023; revised 25 June 2023 and 31 August 2023; accepted 18 October 2023. Date of publication 27 October 2023; date of current version 21 June 2024. This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC). This paper was recommended for publication by Associate Editor Marco S. Nobile upon evaluation of the reviewers' comments. *(Corresponding authors: Mohammad Ali Moni; Kawsar Ahmed.)*

Ananda Sutradhar and Mustahsin Al Rafi are with the Department of Computer Science and Engineering, Daffodil International University, Dhaka 1216, Bangladesh (e-mail: ananda15-2404@diu.edu.bd; mustahsin15-2415@diu.edu.bd).

Pronab Ghosh is with the Department of Computer Science, Lakehead University, Thunder Bay, ON P7B 5E1, Canada (e-mail: pghosh1@lakeheadu.ca).

F M Javed Mehedi Shamrat is with the Department of Computer System and Technology, University of Malaya, Kuala Lumpur 50603, Malaysia (e-mail: javedmehedicom@gmail.com).

Md. Moniruzzaman is with the Department of Software Engineering, Lakehead University, Thunder Bay, ON P7B 5E1, Canada (e-mail: monirimmi@gmail.com).

Kawsar Ahmed, Francis M. Bui, and Li Chen are with the Department of Electrical and Computer Engineering, University of Saskatchewan, Saskatoon SK S7N 5A9, Canada (e-mail: k.ahmed@usask.ca; francis.bui@usask.ca; lic900@usask.ca).

AKM Azad is with the Department of Mathematics and Statistics, Faculty of Science, Imam Mohammad Ibn Saud Islamic University, Riyadh 13318, Saudi Arabia (e-mail: savilbd@gmail.com).

Mohammad Ali Moni is with the AI and Cyber Futures Institute, Charles Stuart University, Bathurst, NSW 2795, Australia (e-mail: mmoni@csu.edu.au).

Digital Object Identifier 10.1109/TAI.2023.3327981

the conventional laboratory test is very complex and requires extensive knowledge and experience. Also, this manual process is time-consuming and can produce inaccurate results. To address such issues, the rapid advancement of AI in detecting diseases could be a potential solution. However, developing such a system for detecting the disease has several key problems and challenges as outlined below.

A. *Problem Statement*

The state-of-the-art thyroid disease detection systems face the following key challenges.

1) Lack of efficient techniques to deal with the relevant and redundant features from the dataset.
2) Over-reliance on ablation techniques for feature selection while ignoring the roles of expert knowledge-driven medical references.
3) The usage of highly imbalanced and prone-to-bias datasets leads to overfitted or underfitted models.
4) Considering a random sampling approach to train and validate their model while using a highly imbalanced dataset, yielding biased results.
5) The lack of generalization capability is due to the failure of hybrid or multiple ensemble methods.

B. *Motivation*

Earlier studies to detect and diagnose the disease were primarily focused on a comparative analysis of various *ML* methods [3], [4], [5], [6]. In particular, their main focus is based on early detection systems [7], [8], [9], [10], [11], [12], [13], [14], developing ensemble [15], [16], [17], artificial neural networks (*ANN*) [18], [19], [20], and explainable models [21], [22], [23]. However, these studies have raised some issues, which can be improved by employing cutting-edge technologies. For instance, Sengupta et al. [3] overlooked data imbalance issues, leading to a tendency to predict the majority classes and encountering overfitting. In addition, in the study [4], a significant number of instances (6371) were randomly discarded from a specific class to achieve data balancing, which can result in the loss of valuable information and an amplified bias. Therefore, we recommend generating a balanced dataset employing an efficient method before incorporating it into the predictive framework. Next, the studies [5], [6], [7] did not attempt any steps to reduce the data dimensionality, resulting in computational costs were increased. Also, Salman and Sonuc [8] dropped three features without considering their impact on predicting results, demonstrating a lack of robustness. To tackle these challenges, Das et al. [9] and Kumar et al. [10] adopted a feature selection approach, choosing merely seven and four features out of thirty and eighteen initial attributes, respectively. However, obtaining consistent results from a single feature set remains challenging, since the selected feature (SF) set lacks alignment with recommended medical references. Hence, a potential solution to enhance the model's robustness is combining feature sets from various feature selection techniques with the guidance of recommended medical references. Subsequently, Sultana and Islam [11], Naeem et al. [12], Olatunji et al. [13], and Alnaggar et al. [14] utilized a random sampling approach to validate the effectiveness of their model, which resulted in biases as the distribution of samples across classes did not properly reflect the underlying population [24]. We introduced a stratified five fold cross-validation approach to solve the issue, which keeps a similar class ratio across the defined number of folds.

The conventional ensemble techniques are trained individually [15], [16], [17], where any incorrect selection by the predicted models can result in lower accuracies. In particular, the studies in [15] and [16] have the potential to introduce bias and overfitting due to their excessive reliance on the Voting (*VT*) ensemble method [25]. Thereby, the study [17] explored a distinct ensemble method named Stacking (*ST*). Despite that, the single ensemble classifier has two primary issues, including limited diversity and overfitting [26], [27]. In response to the aforementioned concerns, we introduce a three stage hybrid classifier (*3SHC*) that employs Bagging (*BG)* to alleviate the overfitting issues during *VT* aggregation [28], and endeavors to amplify performance by synergistically integrating a base and meta-level classifier [29]. The *3SHC* combines the strengths of multiple ensemble methods and thus achieves a more comprehensive understanding of the data. On the other hand, utilizing a standalone *ANN* model [18], [19] uses deep neural networks with many parameters, which makes it challenging to prevent overfitting while generating results on complex neural nets during testing [30]. Simultaneously, Li et al. [20] proposed a hybrid *ANN* model and showed that the model's performance may be reduced if the input data is noisy or biased. To overcome these problems, using multiple *ANNs* allow us to leverage their individual strengths, where each *ANN* may excel in learning various features or patterns and capture a comprehensive representation of the data [31]. Therefore, combining the predictions of three individual *ANN* models, we aimed to develop another hybrid model called the three stage hybrid artificial neural network (*3SHANN*) that employed a majority voting approach. In this case, the final prediction is determined by the class label that receives the majority of votes from the individual models. This approach provides a more stable and diverse set of predictions while lowering the risk of overfitting, thus making it a versatile tool for a wide range of ML tasks.

Moreover, the studies in [21] and [22] aimed to assess the impact of the outcome by employing an explainable model called shapley additive explanation (SHAP). However, the SHAP is computationally expensive and sensitive to feature correlation [32]. Due to the utilization of LIME, the study in [22] may deviate from these concerns. Nevertheless, the sensitive nature of the medical domain makes it crucial to ensure a higher level of confidence when developing a diagnostic model. One way to achieve enhanced interpretability and trust in such models' outcomes is by employing multiple explainable AI methods. Therefore, we introduced permutation feature importance (PFI) as an initial step to justify our preferred features. Then, we performed LIME, which helps interpret the model by shedding light on how it reaches its predictions. Furthermore, we evaluated the PDP to explore the approximate risk ranges and most susceptible groups from different subsamples. These additions of PFI and PDP provide valuable insights and instill confidence in our proposed study.

C. Novelty and Contribution

Feature selection based on the guidance of medical references and a combination of multiple ensemble and explainable approaches is a relatively unexplored area in thyroid disease detection. The key contributions are as follows.

1) We select feature sets by multiple feature selection techniques in combination with recommended medical references to reduce unnecessary features or dimensionality.
2) We then propose two intelligent ML-based novel hybrid classifiers (e.g., *3SHC* and *3SHANN*), where the *3SHC* is observed as the best performer.
3) Then, to enhance the trust-ability and confidence of the diagnosis application, *3SHC* incorporates the LIME to clarify which characteristics are more responsible for the prognosis of an individual.
4) By developing the preferred feature set into various subsamples based on gender and age groups, we aim to identify the most susceptible group to thyroid disease and explore the approximate ranges of feature values associated with higher risks.
5) Furthermore, the *3SHC* achieves an ACC of 99.29%, which outperforms the state-of-the-art models and incorporating with PDP, it concludes that aged and female individuals are more vulnerable to the disease, which significantly aligned with the findings in literature [2].

The remainder of this article is organized as follows. Section II presents an overview of the related work. A detailed explanation of the research methodology is described in Section III. Section IV displays the experimental outcomes of the proposed work with a suitable application and explanatory models. Section V discusses the superiority of our model and compares the findings with current work. Finally, Section VI concludes this article.

II. Related Work

Recently, a substantial number of researchers have directed their attention toward thyroid disease detection. Consequently, upon reviewing the existing studies on the disease, it becomes apparent that a majority of researchers have centered their efforts on conducting thorough comparisons, early diagnosis, introducing hybrid classifiers, and constructing interpretable models. For example, Sengupta et al. [3] showed a comparison of ML classifiers, where the Random Forest (*RF*) achieved the highest ACC of 99.14%. Chaganti et al. [4] using multiple ML and DL classifiers and showed that Extra Tree (*ET*) based SFs yield the highest results (99% ACC) with the *RF* classifier. The authors claimed that the ML classifiers are more essential than DL in terms of ACC and computational complexity. Likewise, Alyas et al. [5] created a comparative analysis of four ML classifiers, where the *RF* method produced an improved ACC of 94.8%. Aversanoa et al. [6] applied several preprocessing techniques and ML classifiers [i.e., *RF*, Decision Tree (*DT*), Ada Boost (*AB*), Gradient Boost (*GB*), and *ET*], where the *ET* obtained the highest ACC of 84%.

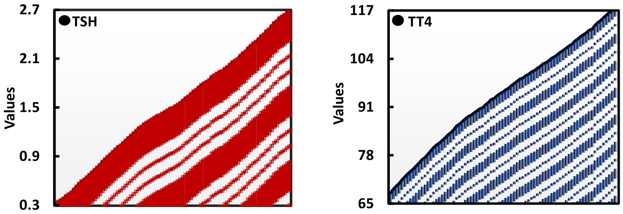
Hu et al. [7] demonstrated the potential of ML classifiers for detecting thyroid disease. Four classifiers, including *GB* and *ANN*, were used to perform the classification task. The *GB* classifier showed the highest area under the curve (AUC) score of 93.8%. Salman and Sonuc [8] built an ML-based system with real patient data, where the *RF* generated the highest ACC of 98.93% on the modified feature set compared. Das et al. [9] have proposed a multiclass classification of hypothyroidism utilizing relevant features and supervised ML classifiers. For the proposed study, they only chose seven relevant features out of thirty based on the correlation scores. Kumar et al. [10] suggested an optimization-based feature selection method, namely differential evaluation with butterfly optimization algorithm (*DE-BOA*) to enhance the outcomes of their study. To categorize the thyroid disorder, the fuzzy C-means (*FCM*) algorithm has achieved 94.3% ACC with *DE-BOA*-based SFs. Sultana and Islam [11] and Naeem et al. [12] examined various ML classifiers for diagnosing the disease. Their experimental section reveals that the ACC improved by 99% and 84.72% for the *RF* and *SVM* classifier, respectively. Olatunji et al. [13] introduced an ML-based tool with the aid of a Saudi Arabian dataset. Where the *RF* was observed as the best performer with a 90.91% ACC. Subsequently, Alnaggar et al. [14] utilized *XGboost* classifier that achieved 99% ACC.

Moreover, a few researchers introduced some ensemble classifiers employing *VT*, and *ST*. Likewise, Solmaz et al. [15] have proposed an android-based thyroid diagnosis application with an ensemble classifier, where a 99.08% ACC was recorded with the aid of the ensemble classifier. Xie et al. [17] presented an *ST*-based ensemble classifier by considering *DT* and *RF*, *AB*, *ET*, and *XGboost* as base classifiers. Based on the experimental section, their presented classifier showed an ACC score of 92.3%. Similarly, Dharamkar et al. [16] presented a suitable ensemble method using the *VT* scheme with *C4.5* and the *RF* classifiers (*CCTML*), where the *CCTML* model produced an acceptable ACC of 96%. Numerous studies showed a great interest in developing *ANN* methods. For instance, Islam et al. [18] and Savcı and Nuriyeva [19] examined various different classifiers, where the *ANN* produced the highest ACC of 95.87% and 98% compared to others. Li et al. [20] have proposed a diagnostic framework for thyroid disease using association rule (*AR*) mining and *ANN*. They used a back-propagation *ANN* to classify the thyroid as sick or negative. Finally, their proposed *AR-ANN* method has produced 95.58% ACC for the sick dataset. Furthermore, Arjaria [21] and Lu et al. [22] utilized a SHAP explainable method to explore the model predictions and impact on input features, respectively. Subsequently, Hossain et al. [23] employed both SHAP and LIME methods to explore the behaviors more precisely.

Despite using multiple random feature selection techniques and traditional ML classifiers, the ACC to predict thyroid diseases was limited. Also, the methodologies did not show any verification based on the recommended medical references, which raises concerns about their reliability. Unlike prior existing work, we proposed an explainable AI-based intelligent thyroid prediction system utilizing two-hybrid classifiers, namely *3SHC* and *3SHANN*. Through extensive experiments, we obtained an

TABLE I
METHODOLOGICAL COMPARISON OF KEY METHODS OR FINDINGS BETWEEN THE EXISTING STUDIES AND OUR PROPOSED SYSTEM

Key methods or findings	**[3]**	**[4]**	**[5]**	**[6]**	**[7]**	**[8]**	**[9]**	**[10]**	**[11]**	**[12]**	**[13]**	**[14]**	**[15]**	**[16]**	**[17]**	**[18]**	**[19]**	**[20]**	**[21]**	**[22]**	**[23]**	**Our study**
Both over and under-sampling	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-	✓	-	-	-	-	-	✓
A novel HRF Technique	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓
The proposed classifier (3SHC)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓
The proposed classifier (3SHANN)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓
Justification of the preferred features	-	-	-	-	-		-	- -	-	-	-	-	-	-	-	-	-	-	-	-	-	✓
Reasons behind the prediction	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-	✓	✓
A decision-making thyroid application	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓
In-depth analysis with sub-samples	-	-	-	-	-	-	-		-	- -	-	-	-	-	-	-	-	-	-	-	-	✓



(a) (b)

Fig. 1. Assessing linearity before performing interpolation on features with missing values: examples using (a) TSH and (b) TT4.

improved ACC of 99.29%, surpassing the outcome of prior existing methods. Then, we considered the PFI and LIME into account to justify the preferred features and explore the outcome explanations, respectively. In addition, we evaluated the PDP to generate approximate risk ranges for different subsamples to further enhance the interpretability. This augmentation of PFI, LIME, and PDP provides valuable insights and instills confidence in our proposed study. The novelty of our study compared to the current studies is shown in Table I, where (✓) indicates the existing research that already considers the mentioned method, and (-) means the current study has not been considered yet. This table demonstrates the novel contribution of our study in the field of *in silico* diagnosis of thyroid disease.

III. RESEARCH METHODOLOGY

In this section, we describe the overall methodological framework of our study, including a predictive model, suitable application, and explanatory model, as schematized in Fig. 2. Moreover, the used notations are listed in Table II.

TABLE II
NOTATIONS USED IN THE STUDY

Notation	**Definition**	**Notation**	**Definition**
OSD	over-sampled dataset	USD	under-sampled dataset
FIS	Feature Importance of Extra Tree Classifier	IGS	Information Gain Feature Selection
LAS	Least Absolute Shrinkage and Selection Operator	HRF	High-Risk Factor Analysis
1SHC	First Stage Hybrid Classifier	2SHC	Second Stage Hybrid Classifier
3SHC	Third Stage Hybrid Classifier	3SHANN	Third Stage Hybrid Artificial Neural Network
$\mathrm{X} = \{\mathrm{x}_1, x_2, \mathrm{x}_3, \ldots x_n\}$	Input data	$\mathrm{C} = \{\mathrm{C}_1$ to $\mathrm{C}_4\}$	Number of base classifiers
$\mathrm{Y} = \{y_1, \mathrm{y}_2, y_3, \ldots y_n\}$	Output data	$D_{tr} = \sum_{i=1}^{m} (x_i,\ y_i)$	Training data
M	Set of the minority classes in SMOTE	$x \in M$	Determine k nearest neighbors in SMOTE
S_R	Sampling rate in SMOTE	M_1	Random S_R samples
I	Majority classes in NM	O	Minority classes in NM
R	Reduced features in LAS operation	$l(\beta,\ y_j)$	Log-likelihood function of LAS
λ	Tune parameter in LAS	β_j	Coefficient in LAS
E	Entropy of DT	G	Information Gain of DT
$P+$	Positive samples in DT	$P-$	Negative samples in DT
l,L	Lower, upper limit of RF	$F(x)$	Optimal function of GB
F	Frequency of the training instances in AB	h_e	Output of the weak learners for AB
e	Weak learners of AB	a_e	Weight of e for AB
$\{r_1$ to $r_n\}$	Residuals in GB	$hn(x)$	Weak learners of GB
C_h	Meta classifier	D_h	Input for C_h
S	Bootstrap samples in 2SHC	$1/N$	Standard deviation
k	Number of classes	$\{Y_1,Y_2,Y_3\}$	Scaled individual ANN
$\{w_1, w_2, w_3\}$	The weight matrix of 3SHANN	$\{b_1, b_2, b_3\}$	The bias vector of 3SHANN
K	Logits in 3HANN	Y_F	3SHANN final outcome

TABLE III
REPRESENTATIVE FEATURES, DATA TYPES, AND NUMBER OF IDENTICAL VALUES (NIV) OF THE DATASET

Name	Type	NIV	Name	Type	NIV	Name	Type	NIV
Age	IN	94	Sex	BO	2	OT	BO	2
QOT	BO	2	OAM	BO	2	Sick	BO	2
PRG	BO	2	TS	BO	2	I1T	BO	2
QHO	BO	2	QHE	BO	2	LI	BO	2
GO	BO	2	TU	BO	2	HY	BO	2
PS	BO	2	TSHM	BO	2	TSH	FL	288
T3M	BO	2	TT4M	BO	2	TT4	FL	242
T4UM	BO	2	T4U	FL	147	FTIM	BO	2
FTI	FL	235	TBGM	CO	1	T3	FL	70
TBG	Null	-	RS	ST	5	TF	BO	2

A. Data Description

A real-world thyroid dataset is collected from a well-known data repository, namely, Kaggle [33], where 30 features are presented with 3772 different case records of the thyroid. These attributes include Age, Sex, On Thyroxine (OT), Query On Thyroxine (QOT), On Antithyroid Medication (OAM), Sick, Pregnant (PRG), Thyroid Surgery (TS), I131 treatment (I1T), Query Hypothyroid (QHO), Query Hyperthyroid (QHE), Lithium (LI), Goiter (GO), Tumor (TU), Hypopituitary (HY), Psych (PS), TSH Measured (TSHM), TSH, T3 Measured (T3M), TT4 Measured (TT4M), TT4, T4U Measured (T4UM), T4U, FTI Measured (FTIM), FTI, TBG Measured (TBGM), T3, TBG, Referral Source (RS), and Target Feature (TF). Table III listed their corresponding data types (e.g., integer (IN), string (ST), Boolean (BO), constant (CO), and float (FL)) and the number of identical values (NIV). The features including TSH, T3, TT4, FTI, and T4U assess the various function related to the thyroid gland and measures its level in blood from lab test. Several clinical attributes like Sick, Pregnant, TS, GO, TU, PS, etc. are also included in the employed dataset. In the test report analysis, 3541 negative and 231 positive thyroid cases are found in the TF, rendering a highly imbalanced dataset.

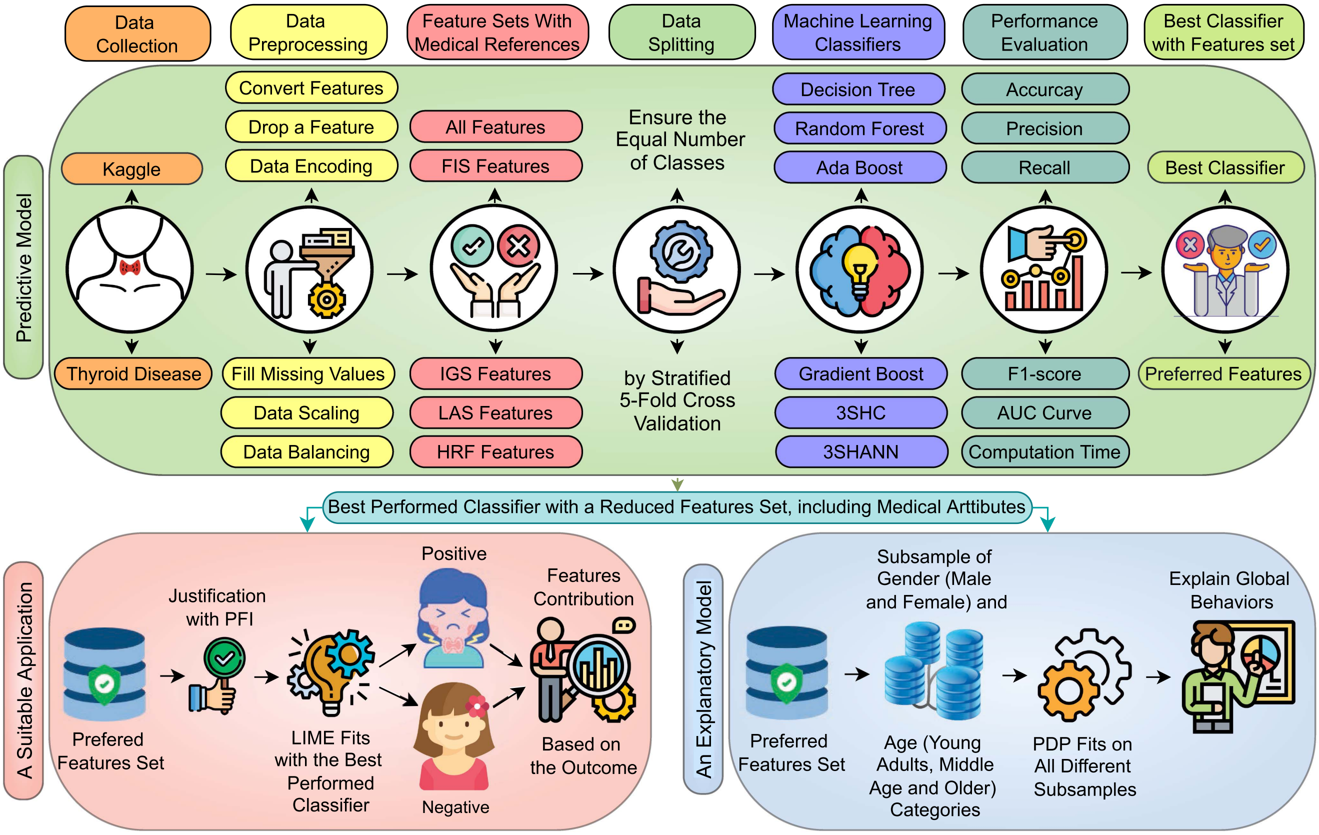


Fig. 2. Flow diagram of our system discussed with three subfolds (Predictive Model, A Suitable Application, and An Explanatory Model).

B. Data Preprocessing

Data preprocessing is a crucial step in training any ML model for extracting meaningful insights from the original data. Fundamentally, the processing techniques transform the original data into an understandable or readable format. In our experiment, most of the features are categorical, which needed to be converted into numeric vectors before feeding into the neural network [34]. The data are then encoded using a Level encoder without affecting its dimension. After that, we observed a significant number of features containing missing values, namely TSH, TT4, T4U, FTI, Age, T3, Sex, and TBG. Any feature with missingness beyond 50% [6] was eventually dropped from our further analyses, otherwise, an imputation technique, namely, Linear interpolation, was used to fill the missing values. Fundamentally, Linear interpolation is an imputation method in which the missing and nonmissing values are assumed to have a linear relationship. To check the linearity, we observed the relationships between the variables that have missing values and the variables that were subjected to be imputed. Results suggested a linear pattern between two features named TSH and TT4, as shown in Fig. 1.

Furthermore, to ensure that all missing values have been filled in, we recheck the data for missing values. In this way, we greatly improve the data utilization and reduce the impact of the missingness of independent variables on the filling error of dependent variables [35]. Moreover, a Min–max scaler technique is used to normalize the data between 0 and 1. Next, we describe the preprocessing approaches adopted for alleviating data imbalance issues.

1) Synthetic Minority Oversampling Technique: After performing the abovementioned preprocessing techniques, an oversampling technique Synthetic Minority Oversampling Technique (SMOTE) is employed to overcome the data imbalance issue, named an over-sampled dataset (OSD). In the primary steps, it sets the minority class of M, and the k-nearest neighbors (*KNN*) of x are determined for each $x \in M$ by evaluating the Euclidean distance between x and each sample in M. According to the unbalanced percentage, the sampling rate S_R is determined. The set of M_1 is created by randomly selecting S_R samples (x_1 to x_n) from each $x \in M$ KNN. For each case $x_i \in M_1$ ($i = 1$ to S_R), a new example is created using (1), where $rand(0, 1)$ represents a random number between 0 and 1. Finally, it yields 7082 distinct samples as the final dataset, with 3541 samples for each class

$$x'' = x + rand(0, 1) \times [x - x_i]. \tag{1}$$

2) Near-Miss: Furthermore, the Near Miss (NM) undersampling technique is utilized to balance the class distribution by eliminating the majority of class samples, namely, the USD. To simplify this process, it begins with calculating the distances between all instances in the majority and minority classes. After that, it chooses the I instances of the majority class that is closest to the minority class. Then, the closest method will produce $(O \times I)$ instances of the larger class if the minority class contains O instances. In order for both classes to have an equal number of records, the number of the majority class has been decreased to match the total number of the minority class. After ending this process, this strategy produced 231 samples for both classes. These preprocessing steps are followed by 28 columns that were taken as input, and the class column serves as the model's output.

C. Feature Selection Based on Expert Knowledge-Driven Features

Feature selection is a process that chooses a subset of relevant features to use as input while developing a predictive model. In this study, three well-known feature selection techniques, i.e., Feature Importance, Information Gain, and Least Absolute Shrinkage and Selection Operator are adopted. In addition to those, we have developed another feature set, namely, the HRF. The following explanations provide the specifications of the abovementioned feature selection techniques:

1) Feature Importance of Extra Tree Classifier (FIS): The term "feature importance" refers to a strategy that rates input features based on their ability to predict a given target variable. It comes with tree tree-based classifier and we have used the *ET* classifier as a learner, which computes the importance score by aggregating the impurity reduction of each feature.

2) Information Gain Feature Selection (IGS): Information gain is a prominent technique and entropy-based feature selection for selecting the most relevant features. This method computes the depletion in entropy or amazes from transforming a dataset, and mainly works by calculating the gain of each variable in the context of the target variable.

3) Least Absolute Shrinkage and Selection Operator (LAS): To identify the relevant features, LAS regression will attempt to reduce the coefficient of the less important feature to 0, as the simultaneous presence of two features with linear correlation would increase the value of the cost function. Equation (2) stated the definition of the LAS penalized regression method

$$LAS = argmin\left[-l(\beta,\ y_j) + \lambda\sum_{j=1}^{R}|\beta_j|\right]. \tag{2}$$

where R is the number of features that have been reduced, λ is the tuning parameter that determines how strong the $L1$ penalty will be, β_j is the coefficient of jth feature, and $l(\beta,\ y_j)$ is the log-likelihood function.

4) High-Risk Factor Analysis (HRF): To achieve a reliable feature set with accurate findings, the most contributing feature set, HRF, is created based on the identical features among FIS, IGS, and LAS approaches. The HRF feature is obtained by (3), which is as follows:

$$\begin{aligned} \text{HRF} &= (\text{FIS}) \cap (\text{IGS}) \cap (\text{LAS}) \\ \forall\{\text{FIS}\}, &\{\text{IGS}\}, \{\text{LAS}\} \subset \text{Processed dataset} \\ \text{FIS IGS} &= \{1, .., 12\}\ \{1, .., 12\} \in\ \text{OSD and USD} \\ \text{LAS} &= \{1, .., 14\} \in\ \text{OSD and}\ \{1, .., 15\} \in\ \text{USD}. \end{aligned} \tag{3}$$

Both FIS and IGS techniques select 12 significant features based on the importance of each feature rank in both OSD and USD. In contrast, the LAS picks 14 attributes for OSD and 15 attributes for USD. After receiving a feature set for each, we consider picking the HRF-based attributes among all the different feature sets. The number of initial features chosen by HRF for OSD and USD reaches 9 and 7, respectively.

Moreover, to enhance the clinical relevance we have gathered the 10 thyroid-based attributes from the recommended medical literature and references [36], [37], [38], [39]. These features play a vital role in diagnosing the disease, we consider these features as expert knowledge-driven medical reference features. To ensure that not one of the SF sets is missing from the medical reference features throughout the classification process, these 10 attributes are compared with the acquired features of FIS, IGS, LAS, and HRF. If any medical attributes are missing in the initially SF sets, we add missing attributes to update the final selected feature (FSF) sets. Table IV presents SF from each technique, as well as the FSF that are used in the classification process. Note, the (✓) mark represents that the mentioned techniques selected the particular feature, and the (-) mark represents not selected. As some of the features with the names QOT, OAM, LI, HY, FTIM, TBGM, and RS are not selected by the processes described, we have omitted them from the table. Finally, we received the identical FSF for OSD and USD in the case of HRF. Table V holds the Ordinary Least Squares Regression (OLS) results for these. Where it is observed that the fundamental standard error of the coefficient estimate of all features is close to 0. In addition, the coefficient is statistically significant when the p-value is low, which means it is unlikely to have happened by chance alone. Hence, the obtained p-value of this table indicates that the relationships between the features are statistically significant. Then, the stratified cross-validation (SCV) was implemented on the processed dataset and FSF. The SCV can enhance the accuracy and generalizability of the model by ensuring that each fold contains a representative sample for training and testing in an equal number of classes. For both OSD and USD, we utilized five folds of SCV to train and evaluate the models. The models are executed five times, with each fold acting as the test set once and the remaining four folds being utilized for training. We can get a more accurate and reliable assessment of the model's performance on unseen data by averaging the performance across multiple folds. In addition, we enabled the shuffle parameter to remove any potential biases in the data during the process.

D. Simulation Assumptions

During the simulation phase, we assumed the following properties as shown in Table VI.

1) First, we assumed that the data used in the simulation are Independent and Identically Distributed (IID), meaning that each sample in the dataset is unrelated to other samples and follows the same underlying distribution. Assuming IID data can simplify the analysis and modeling process. However, in medical diagnosis, patients' data often exhibit

TABLE IV
LIST OF FEATURE SETS SELECTED FROM DIFFERENT FEATURE SELECTION TECHNIQUES FROM OSD AND USD

Techniques	Age	Sex	OT	Sick	PRG	TS	I1T	QHO	QHE	GO	TU	PS	TSHM	TSH	T3M	TT4	T4UM	T4U	FTI	T3	TT4M
SF-FIS (OSD / USD)	✓/ ✓	✓/ ✓	✓/ ✓	✓/ -	- / -	- / -	- / -	- / ✓	✓/ -	- / -	- / -	✓/ -	- / ✓	✓/ ✓	✓/ ✓	✓/ ✓	- / -	✓/ ✓	✓/ ✓	✓/ ✓	- / ✓
SF-IGS (OSD / USD)	✓/ ✓	✓/ -	✓/ ✓	✓/ -	- / -	- / -	✓/ ✓	- / -	- / -	- / -	- / ✓	- / -	✓/ ✓	✓/ ✓	- / ✓	✓/ ✓	- / -	✓/ ✓	✓/ ✓	✓/ ✓	- / ✓
SF-LAS (OSD / USD)	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	- / -	✓/ -	- / ✓	✓/ ✓	- / ✓	- / ✓	- / -	✓/ ✓	- / -	✓/ ✓	✓/ -	✓/ ✓	✓/ -	✓/ ✓	✓/ ✓	✓/ ✓	- / ✓
SF-HRF (OSD / USD)	✓/ ✓	✓/ -	✓/ ✓	✓/ -	- / -	- / -	- / -	- / -	- / -	- / -	- / -	- / -	- / -	✓/ ✓	- / -	✓/ ✓	- / -	✓/ ✓	✓/ ✓	✓/ ✓	- / -
Medical attributes	✓	✓	✓	✓	✓	-	-	-	-	-	-	-	-	✓	-	✓	-	✓	✓	✓	-
FSF-FIS (OSD / USD)	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	- / -	- / -	- / ✓	✓/ -	- / -	- / -	✓/ -	- / ✓	✓/ ✓	✓/ ✓	✓/ ✓	- / -	✓/ ✓	✓/ ✓	✓/ ✓	- / ✓
FSF-IGS (OSD / USD)	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	- / -	- / ✓	- / -	✓/ -	- / -	- / ✓	✓/ -	- / ✓	✓/ ✓	✓/ ✓	✓/ ✓	- / -	✓/ ✓	✓/ ✓	✓/ ✓	- / ✓
FSF-LAS (OSD / USD)	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	- / -	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	- / -	✓/ ✓	- / -	✓/ ✓	- / -	✓/ ✓	- / -	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓
FSF-HRF (OSD / USD)	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	- / -	- / -	- / -	- / -	- / -	- / -	- / -	- / -	✓/ ✓	- / -	✓/ ✓	- / -	✓/ ✓	✓/ ✓	✓/ ✓	- / -

TABLE V
RESULTS OF ORDINARY LEAST SQUARES REGRESSION (OLS) FOR HIGH-RISK FACTOR (HRF) FEATURES IN OSD AND USD

	over-sampled dataset (OSD)					under-sampled dataset (USD)				
Features	Coefficient	Std. error	*t*-Statistic	*p* >*t*	[0.025 0.975]	Coefficient	Std. error	*t*-Statistic	*p* >*t*	[0.025 0.975]
Age	0.0023	0	9.807	0	0.002 0.003	0.0012	0	1.165	0.245	-0.001 0.003
Sex	-0.0229	0.009	-2.417	0.016	-0.042 -0.004	-0.0726	0.0039	-1.847	0.065	-0.150 0.005
OT	-0.2747	0.016	-16.704	0	-0.307 -0.242	-0.2433	0.063	-3.873	0	-0.367 -0.120
Sick	-0.1436	0.025	-5.758	0	-0.192 -0.095	0.1651	0.069	2.392	0.017	0.029 0.301
PRG	0.1129	0.05	2.256	0.024	0.015 0.211	0.1251	0.274	0.457	0.648	-0.413 0.663
TSH	-0.0013	0	-5.797	0	-0.002 -0.001	-0.0013	0	-1.171	0.242	-0.003 0.001
T3	-0.4002	0.006	-65.795	0	-0.412 -0.388	-0.4050	0.02	-15.56	0	-0.456 -0.354
TT4	-0.0014	0	-5.413	0	-0.002 -0.001	0.0049	0	2.942	0	0.002 0.008
T4U	0.3609	0.035	10.236	0	0.292 0.430	-0.2275	0.206	-1.104	0.27	-0.632 0.177
FTI	0.0038	0	15.912	0	0.003 0.004	-0.0019	0	-1.227	0.22	-0.005 0.001

TABLE VI
ASSUMPTIONS OVER THE SIMULATION PROCESS

Data	Drop features	Relations between missing and nonmissing feature values	Feature Set
(IID)	Missing values >50%	Linear	Relevant and Complete

dependencies and correlations. Ignoring these dependencies can lead to inaccurate results, which cannot handle the complexity of real-world diagnostic scenarios.

2) Second, any feature with missing data beyond 50% [6] was eventually dropped from our model for further analysis. While dropping features with a high number of missing values can help mitigate issues related to incomplete datasets, it can also lead to potential information loss. The decision to exclude features with high missingness should be carefully justified, as important diagnostic information might be discarded.
3) Third, we hypothesized that the missing and nonmissing values have a linear relationship, where we observed the relationships between the variables in our experiment that have missing values and the variables that were subjected to be imputed. Assuming a linear relationship between missing and nonmissing values can oversimplify the underlying patterns and dependencies. Missing values can arise due to specific characteristics of the individuals or cases or systematic biases. These reasons can lead to more complex relationships that go beyond simple linearity.
4) Last, we hypothesized that all relevant features were included in the analysis based on the recommended medical references. It is important to consider all crucial features that could impact the accuracy and effectiveness of the diagnosis system. Failing to account for important features can introduce biases and affect the system's performance.

E. Classifiers Description

We introduce four well-known traditional classifiers, i.e., Decision Tree, RF, Ada Boost, Gradient Boost, and two hybrid classifiers (*3SHC* and *3SHANN*) in our study. A detailed explanation of each classifier is described as follows.

1) Decision Tree: A decision tree (*DT*) is built by an algorithmic approach that identifies ways to split a dataset based on different conditions. The "splitting" is used to construct the tree from the root node upward by picking the "Best Feature" from the available features. To choose the "Best Feature," Entropy (*E*) and Information Gain (*G*) are calculated. The computation of *E* and *G* is done using (4) and 5, where (D_{tr}) is the training dataset, $(P+)$ and $(P-)$ refer to positive and negative samples, respectively. This process is repeated for every subtree rooted at the new node. Particularly, the *DT* classifier could be the right choice when the dataset contains too much logical, discrete, or categorical data.

$$E(D_{tr}) = -\,(P+)\log_2{(P+)} - (P-)\log_2{(P+)} \quad (4)$$

$$G(Attribute\ x) = E(Decision\ Attribute\ y) - E(x, y). \quad (5)$$

2) Random Forest (RF): RF classifier is built using a group of *DTs*, and each tree in the ensemble is collected from a sample drawn from a training set with replacement, called the bootstrap sample. A majority voting approach is considered for the classification task, where the most frequent variable is selected as a predicted class. We consider the given data as $X = \{x_1, x_2, x_3, \ldots., x_n\}$ with responses $Y = \{y_1, y_2, y_3, \ldots., y_n\}$, where l has a lower limit of 1 and an upper limit of L. As seen in (6), it illustrates how to conduct prediction for samples by averaging the predictions for x'

provided by each unique tree for x, where $\sum_{l=1}^{L} fl(x')$

$$j = \frac{1}{L}\sum_{l=1} fl(x') \tag{6}$$

3) Ada Boost: Adaboost (*AB*) is one of the ensemble boosting classifiers. *AB* constructs a strong classifier by integrating multiple low-performing classifiers to get highly accurate and robust classifiers. The fundamental concept of *AB* classifiers is to set the data weights and train the data sample to be weighted with a starting weight as $1/F$, where F is the frequency of training instances. After getting the outcome, the error is calculated as follows:

$$Error = \frac{(correct - F)}{F}. \tag{7}$$

Then, the final classification is measured as

$$H_e = +/-\sum_{e=1}^{e}(a_e h_e(p)). \tag{8}$$

Here, $e =$ the total number of weak learners, $h_e(p) =$ the output of the weak learners, $a_e =$ the weight of learners e.

4) Gradient Boosting (GB): *GB* is an ensemble method that combines multiple weak learning models, which can make a robust predictive model with high-dimensional data and reduce the loss function through optimization. Letus say we have a training dataset $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, the goal is to learn an optimal function $F(x)$ that predicts the dependent variables y for new data points. At begin initialize the model with a constant value $f_0(x)$ to calculate residuals $\{(r_1 = y_1 - f_0(x_1))\}$,...,$\{(r_n = y_n - f_0(x_n))\}$. Then, update our model by fitting a weak learner $h_1(x)$ to predict the residuals for the present model. Until the residuals can no longer be significantly decreased, this procedure will continue by fitting new weak learners $h_n(x)$ to the residuals. The final procedure is derived as follows:

$$F(x) = f_0(x) + h_1(x) + h_2(x) +, \ldots, + h_n(x). \tag{9}$$

5) Three-Stage Hybrid Classifier: "Unity is strength," this proverb reasonably captures the fundamental principle behind machine learning's (ML) extremely potent ensemble approaches. The proposed *3SHC* is developed by combining traditional classifiers with Voting (*VT*), Bagging (*BG*), and Stacking (*ST*) ensemble classifiers, where *VT* mainly works to train the various base classifiers and combines the predictions from each classifier. The *BG* ensemble classifier creates some bootstrap data samples and fits the base classifiers on each sample, then aggregating their individual predictions. In *ST*, a meta classifier receives the output of base classifiers as input for the final classification. First, we set four conventional algorithms *(DT, RF, AB, GB)* as base classifiers on the training data (D_{tr}) using the *VT* approach. This method is appropriate when we use more than two base classifiers for making predictions [40]. We set the type of voting parameter as *HARD*. Since it works with all classifiers and every classifier votes for a class, the combination of each vote provides a reliable outcome with reduced errors [41]. The process of this type of voting is more straightforward, interpretable, and robust against outliers. The *VT* approach can be evaluated as

$$1SHC = mode\{DT(D_{tr}), RF(D_{tr}), AB(D_{tr}), GB(D_{tr})\}. \tag{10}$$

Second, as the *BG* method cannot fit fully independent models because it would require too much data, we divide the training data set (D_{tr}) into S number of data samples $D_{tr1}, \ldots, D_{trS}$ using a bootstrap method. We then parallely fit the *1SEC* classifier on these multiple datasets and achieve multiple predictions $\{1SHC(D_{tr1}), \ldots, 1SHC(D_{trS})\}$. After that, the *BG* method takes an average of the predictions of various classifiers to obtain an ensemble model with lower variance. This method can be written as

$$2SHC = \sum_{i=0}^{S}\{1SHC(D_{tri}), \ldots, 1SHC(D_{trS})\}/S \tag{11}$$

Third, the *ST* method is applied to produce a strong model and significantly improve the final outcome by learning data from both base and meta-level classifiers [29]. To construct a dataset for training meta classifier, *1SHC* and *2SHC* are used as base classifiers on the training data $(1SHC(D_{tr}), 2SHC(D_{tr}))$ and to get a base prediction result. Based on the predictions of training data, construct a new training set (D_h) to train a second-level. A *LG* algorithm (C_h) is trained as a meta classifier on D_h, which is significantly used in binary classification and the final prediction is received. The final stage is defined as, $3SHC = C_h(D_h)$, Algorithm 1 holds the overall procedure of the proposed classifier.

6) Three-Stage Hybrid Artificial Neural Network: We build upon the idea of the Three *3SHANN* for performing classification tasks of the thyroid. We exploit the multiple *ANNs* with completely diverse architectural designs and take advantage of their complementing cues through ensembles. We also use a Sigmoid activation layer for individual classifiers during the training of binary classification. Fig. 3 depicts the full workflow of this classifier, which is comprised of *Input, Inference, Feature Transformation, Activation function, and Ensembling*. The input data are provided for all *ANN* models. Each *ANN* has different interface layers in our network architecture (*3SHANN*); for example, ANN_1 has approximately 1 057 281 and 267 521, ANN_2 has around 266 497 and 68 097, and ANN_3 has about 67 713 and 17 665 trainable parameters.

In feature transformation, it helps with feature fusion by embedding higher-dimensional *ANN* features to lower-dimensional data using three hidden layers, each of which is linked to a single *ANN* classifier. A Rectified Linear Unit (*ReLU*) activation function is used to rectify the hidden layers' inputs. To improve the performance of the fusion, an input neuron x is updated using the formula $f(x) = \max(x, 0)$ in a neurological manner. Moreover, the single-hidden layer linked with the specific *ANN* feature space X_i decreases the rectified feature dimensionality *N* to *k* as

$$Y_{i,K} = \sum_{n=1}^{N}\omega_{i,k,n} \times X_{i,n} + b_{i,k}. \tag{12}$$

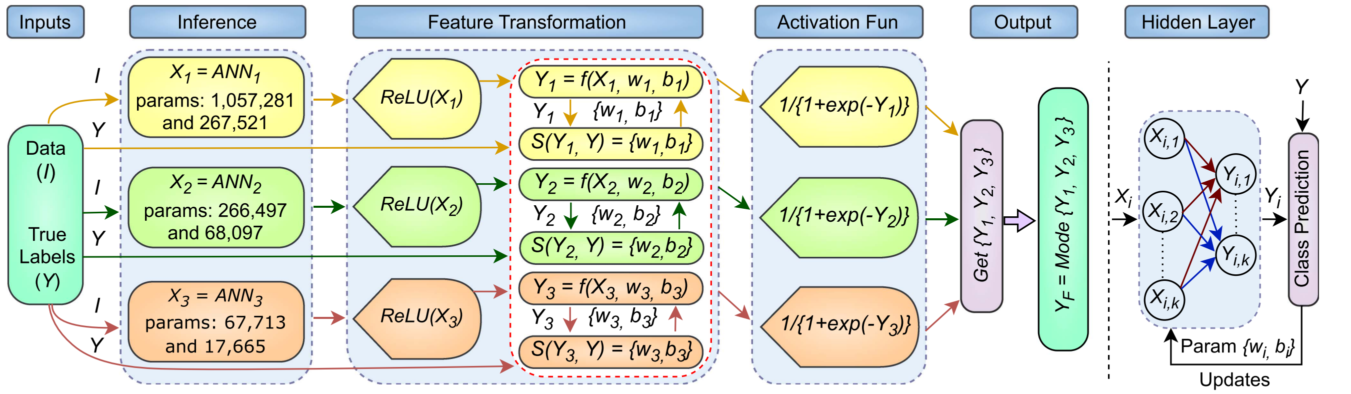


Fig. 3. Architecture of the proposed *3SHANN* classifier by combining multiple *ANNs* with different trainable parameters based on a majority voting.

Algorithm 1: The Three-Stage Hybrid Classifier (*3SHC*): A Combination of Multiple Ensemble Classifiers for Disease Classification.

```
Inputs: Number of base classifiers, $C = C_1$ to $C_4$, Training Data, $D_{tr} = \sum_{i=1}^{m}(x_i, y_i)$, Number of bootstrap samples = $S$, Meta classifier = $C_h$.
Outputs: Classify either negative (0) or sick (1).
First-Stage: Train base classifiers as $DT$, $RF$, $AB$ and $GB$ on the training instances $D_{tr}$.
for $i = 1; i \leq 4;$ i ++ do
    $1SHC_{Base} \leftarrow (C_1(D_{tr}), \cdots, C_4(D_{tr}))$
end for
$1SHC \leftarrow mode\ (1SHC_{Base})$
Second-Stage: Generate $S$ number of bootstrap samples (e.g., $D_1, \cdots, D_S$) from the training instances $D_{tr}$, and then do training with the $1SHC$.
for $j = 1; j <= S; j++$ do
    $D_{trj}, \cdots, D_{trS} \leftarrow (Bootstrap\ (D_{tr}))$
end for
for $b = 1; b <= S; b++$ do
    $2SHC \leftarrow \sum_{b=1}^{S}(1SHC\ (D_{trb}), \cdots, 1SHC\ (D_{trS}))/S$
end for
Third-Stage: Train the $1SHC$ and $2SHC$ as base estimators, and construct a new data set of predictions for a meta-classifier $C_h$.
$Base_p \leftarrow (1SHC(D_{tr}), 2SHC(D_{tr}))$
for $s = 1; s < m; s++$ do
    Apply $Base_p$ to classify training instances $x_s$
    $x_s \leftarrow Base_p\ (x_s)$
    $D_h \leftarrow ({x_s}^{\delta}, y_s)$, where $x_s^{\delta} \leftarrow (x_{1\ s}, \ldots, x_{ms})$
end for
$3SHC \leftarrow (C_h(D_h))$
```

TABLE VII
UTILIZED THE SET OF PARAMETERS DURING THE CLASSIFICATION PROCESS

Classifiers	**Used Parameters**
DT	criterion = "entropy," max depth = 3, random state = 10
RF	n estimators = 6, max depth = 3, random state = 10
AB	n estimators = 6, learning rate = 2.0, random state = 10
GB	max features = 1, max depth = 3, random state = 10
3SHC(VT)	estimators = (DT, RF, GB, AB), type = hard
3SHC(BG)	estimators = (1SHC), n estimators = 6, random state = 10
3SHC(ST)	estimators = (1SHC, 2SHC), use probas = True, meta classifier = LG
3SHANN	loss = binary cross., optimizer = Adam, dropout = 0.03, lr = 0.0005

The weight matrix $\boldsymbol{\omega}$ and the bias vector b is initialized to a zero-mean Gaussian distribution with a standard deviation of $1/N$ and a small value of 0.0005 to prevent divergence in the learning process. Here, N is the feature dimension of the ith *ANN* feature X_i and k is the number of classes. As the output vector Y_i has K number of logits or preactivated values, the matrix size of the parameters is $K \times N$. Each class in the target is represented by a single neuron in Y_i, which has not yet been scaled by an activation function. The binary cross-entropy loss function is utilized here since the target class of our dataset is in binary classes. An adaptive moment estimation (*ADAM*)-based optimizer is used to update the parameters in order to minimize the loss function, which is computationally less expensive. Here, the number of epochs was set as *50* in conjunction with the early stopping method. In addition, as shown on the right side of Fig. 3, each individual logit is employed independently to predict the classes, followed by a loss computation to update the corresponding parameters w_i and b_i of the chosen hidden layer for the particular feature. The training dataset is then used to fine-tune the weight matrices (w_1, w_2, w_3) and bias vectors (b_1, b_2, b_3). Then Y_1, Y_2, and Y_3 are individually scaled with the Sigmoid activation function, that is mostly used in binary classification and occupies the range between 0 and 1. Finally, the majority voting is computed as follows: Y_F = mode$\{Y_1, Y_2, Y_3\}$, which can significantly mitigate errors in final results [41]. Table VII includes a list of the parameters used during classifications.

F. Model's Explainability

Though ML algorithms can play an essential role in data processing, the stakeholders struggle in decision-making due to the lack of transparency regarding the models' outputs. Hence, explainable AI is utilized to solve such a demand and enhance the system's transparency with trust in decision-making. Here, we introduce three explainable approaches named PFI, LIME, and PDP, where the PFI is a model-independent global explanation

TABLE VIII
ACCURACY (ACC), PRECISION (PRE), RECALL (REC), F1-SCORE (F1), AND COMPILATION TIME (CT) MEASURED FOR ALL THE CLASSIFIERS IMPLEMENTED ON VARIOUS FEATURE SETS IN BOTH OVER-SAMPLED (OSD) AND UNDER-SAMPLED DATASETS (USD)

	over-sampled dataset (OSD)						under-sampled dataset (USD)					
Feature sets	**DT**	**RF**	**AB**	**GB**	**3SHC**	**3SHANN**	**DT**	**RF**	**AB**	**GB**	**3SHC**	**3SHANN**
ALL_ACC(%)	93.15	93.57	94.07	94.91	97.17	97.24	93.54	94.62	96.77	95.69	97.84	97.84
ALL_PRE(%)	93.29	93.87	94.11	94.99	97.17	96.29	94.11	95	96.76	95.83	97.56	97.77
ALL_REC(%)	93.18	93.61	94.08	94.9	97.17	98.11	93.75	94.79	96.8	95.62	98.14	97.77
ALL_F1(%)	93.15	93.57	94.07	94.91	97.17	98.11	93.54	94.62	96.77	95.68	97.8	97.77
ALL_CT(ms)	18.2	114	61.7	288	4300	23830	20.8ms	33.2	30.1	70.8	2250	20900
FIS_ACC(%)	94.35	95.62	94.14	96.75	98.58	98.87	97.84	95.69	96.77	96.77	97.84	97.84
FIS_PRE(%)	94.37	95.63	94.19	96.78	98.59	98.85	97.87	95.91	96.76	97.87	97.87	97.77
FIS_REC(%)	94.34	95.63	94.15	96.74	98.58	98.85	97.91	95.83	96.8	97.91	97.91	97.77
FIS_F1(%)	94.35	95.62	94.14	96.75	98.58	98.85	97.84	95.69	96.77	97.84	97.84	97.77
FIS_CT(ms)	16	45.1	42	209	2620	13770	22	20.6	21.4	134	1490	15100
IGS_ACC(%)	94.35	95.05	94.14	96.11	98.3	98.16	96.77	94.62	96.77	96.77	98.92	97.84
IGS_PRE(%)	94.37	95.18	94.19	96.11	98.32	98	97.11	94.72	97.05	96.78	98.95	100
IGS_REC(%)	94.34	95.08	94.15	96.12	98.29	98.28	96.59	94.72	96.66	96.78	98.91	95.83
IGS_F1(%)	94.35	95.05	94.14	96.11	98.3	98.14	96.75	94.62	96.76	96.77	98.92	97.87
IGS_CT(ms)	18.5	58.5	41.3	187	2270	12650	34.5	23.5	63.2	61.7	2560	12340
LAS_ACC(%)	94.49	96.61	93.93	95.9	98.09	98.87	97.84	96.77	96.77	94.62	98.92	98.92
LAS_PRE(%)	94.61	96.61	93.98	95.91	98.09	98.86	97.87	96.76	96.76	94.61	98.91	97.77
LAS_REC(%)	94.47	96.61	93.94	95.91	98.09	98.86	97.91	96.8	96.8	94.65	98.95	100
LAS_F1(%)	94.48	96.61	93.93	95.9	98.09	98.86	97.84	96.77	96.77	94.62	98.92	98.87
LAS_CT(ms)	24.5	31.8	43	456	4330	15705	23.9	52.7	29.6	169	1540	14010
HRF_ACC(%)	96.18	97.45	97.52	97.66	99.29	99.08	96.77	97.84	97.84	97.84	98.92	98.92
HRF_PRE(%)	96.22	97.45	97.53	97.6	99.3	99.03	96.42	97.56	97.87	97.56	98.75	100
HRF_REC(%)	96.17	97.46	97.52	97.6	99.29	99.18	97.22	98.14	97.91	98.14	99.07	97.77
HRF_F1(%)	96.18	97.45	97.5	97.6	99.29	99.05	96.71	97.8	97.84	97.8	98.89	98.87
HRF_CT(ms)	15.9	45.8	40.9	188	2160	13740	20.1	34.7	47.5	47.5	2030	11090

method used to ranks the relevant features based on how each attribute affects the predictions of the trained ML models. Next, the LIME is applied to explain the features individually for a single sample and then evaluates the most responsible factor for the outcomes. Finally, the PDP is used for explaining a model's global behavior by illustrating the relationship between each predictor's marginal effect on the target variable.

IV. Experiments, Results, and Findings

In this article, we have presented ML-based thyroid prediction with several classifiers on different feature sets. To show the robustness of our research in terms of classification, some performance matrices are measured for each classifier by taking an average of stratified fivefold cross-validation. We also compute the running time to demonstrate how efficient our proposed classifiers are in detecting the models' performances. After examining all outcomes, an application to initially detect the disease utilizing a feature set and the best-performing classifier are shown. Finally, a deep explanatory study is presented with the sub-samples of age and gender groups.

A. *Predictive Model*

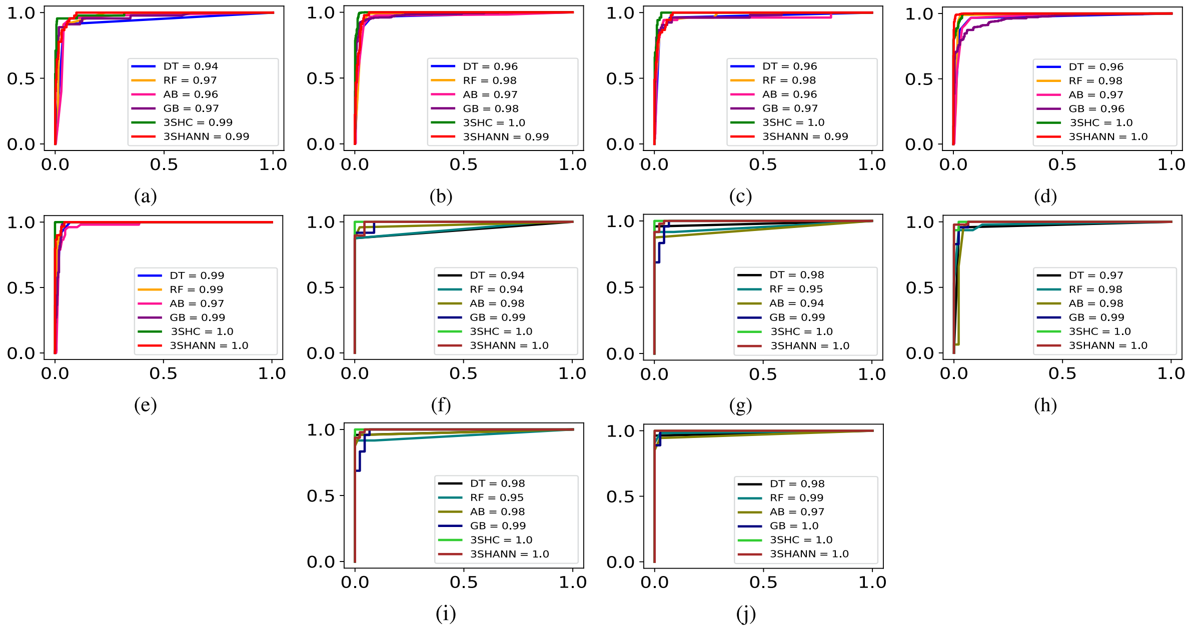
We describe the performance comparison of six ML classifiers on different feature sets. The obtained FSF, including all (28), FIS (13), IGS (13), LAS (15), and HRF (10), are generated for the OSD. In the case of USD, these numbers are 28, 14, 15, 16, and 10, respectively.

1) Accuracy: The most crucial metric in ML is accuracy (ACC), which assesses how well a classifier prediction generalizes to new or previously unexplored data. According to Table VIII, we observed that the *3SHC* and *3SHANN* generate the most reliable ACC for all different feature sets. While considering the OSD, the *3SHC* generates the highest ACC of 99.29% with the HRF-based feature set. In contrast, the *3SHANN* achieves a comparatively lower ACC of 99.08%. In terms of the USD, the *3SHC* discovered an ACC of 98.92% based on IGS, LAS, and HRF features. These ACC indicate that the *3SHC* correctly classifies approximately 91 test samples out of 92.

2) Precision: Precision (PRE) assesses the reliability of the classifier's positive predictions. It enables us to identify false positive instances and modifies their performance as necessary. While considering the OSD, the *3SHC,* and *3SHANN* obtain the highest PRE scores of 99.3% and 99.03%, respectively, with the HRF-based feature set. In terms of the USD, the *3SHC* and *3SHANN* also earn the highest PRE scores from the HRF and IGS attributes.

3) Recall: Recall (REC) determines how precisely a model can identify the number of positive instances to the total number of positive samples. The proposed *3SHC* and *3SHANN* generate REC of 99.29% and 99.18%, respectively, for the HRF-based feature set. In terms of all and FIS features, the *3SHANN* outperforms *3SHC* (see Table VIII). The best REC score for the USD is achieved on the HRF, and LAS features, where the *3SHC* and *3SHANN* obtain results between 97.77% and 100%.

4) F1-Score: F1- score (F1), comprised of PRE and REC scores, calculates how effectively a classifier can predict the result. The *3SHC* and *3SHANN* obtain improved results on the OSD compared to others. Particularly, the *3SHC* performs the highest scores of 99.29% with HRF features (see Table VIII).



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j)

Fig. 4. AUC measured for all classifiers on both over-sampled (OSD) and under-sampled datasets (USD) implemented on (a) OSD- ALL, (b) OSD-FIS, (c) OSD-IGS, (d) OSD-LAS, (e) OSD-HRF, (f) USD-ALL, (g) USD-FIS, (h) USD-IGS, (i) USD-LAS, (j) USD-HRF feature sets.

Regarding the USD, the *3SHC* and *3SHANN* produce scores over 98.87% with both LAS and HRF features.

5) Compilation Time: The compilation time (CT) refers to how fast a classifier can predict the samples and provide a promising solution. According to Table VIII, the *3SHANN* takes the highest time with an overall of 11000–23000 ms (ms) for all different feature sets. In contrast, the *3SHC* takes a comparatively lower time (around 2000–4050 ms).

6) Area Under the Curve: AUC is one of the most crucial evaluation matrices, indicating how efficiently a classifier is in determining the presence or absence of the disease. In Fig. 4, we present the performed AUC scores with different color combinations for OSD (a)–(e) and USD (f)–(g), where the x and y-axis represent the false positive and true positive rates, respectively. Drawing from the outcomes of the OSD evaluation, our proposed *3SHC* and *3SHANN* exhibited remarkable performance, nearing 100% AUC scores for LAS and HRF features. However, the DT classifier displayed comparatively lower scores, reaching approximately 94% for all features. Conversely, our proposed classifiers showcased approximately uniform scores (100%) across five diverse feature sets in USD. Furthermore, the baseline classifiers demonstrated commendable performance in both OSD and USD scenarios.

B. Suitable Decision Making Application

A suitable application can help to deal with the upcoming challenges in medical areas. It can be used to diagnose patients and improve medical care and medical education. In the earlier section, we compared performance results for all different feature sets. According to this comparison, our proposed *3SHC* achieves the highest ACC for the OSD-based HRF features. In order to provide further justification on it, we measure the PFI for the highest performing features set. Table IX describes the HRF features according to the weight of PFI. The first value in each row indicates how much the model's performance deteriorated as a result of the random shuffling. The number after ± indicates the variation in performance from one reshuffle to the next. Here, it is visible that no characteristics had a negative influence and each feature has a sustainable impact on the prediction.

TABLE IX
PFI RANK OF THE OSD HIGH-RISK FACTOR (HRF) FEATURES

Weight	Feature	Weight	Feature
0.6075 ± 0.2835	T3	0.1203 ± 0.2094	T4U
0.0900 ± 0.0850	TT4	0.0732 ± 0.0586	FTI
0.0533 ± 0.1016	Age	0.0359 ± 0.0297	TSH
0.0162 ± 0.0293	On Thyroxine	0.0032 ± 0.0059	Sex
0.0010 ± 0.0028	Sick	0.0007 ± 0.0028	Pregnant

In order to meet the stakeholder's demand for a sensitive medical domain, we concentrated on coming up with the appropriate justifications in light of the stockholder's outcome and incorporating LIME into the particular HRF data sample. The unscale HRF samples are utilized here for visualizing the appropriate value ranges of features. Table X displays the prediction probability for two randomly selected data samples [positive (pos) and negative (Neg)]. The value field indicates the actual value of each feature. In contrast, the Neg and Pos thyroid fields display the LIME-generated values, which indicates whether the feature has a Neg or Pos influence on prediction probabilities. If a feature negatively influences a sample, it is filled in the Neg thyroid with feature names and recommended value ranges, whereas a Pos influence is filled in the Pos thyroid. For a Pos sample, our proposed *3SHC* generates the probability of having thyroid is 100%. T3 is the most contributing feature to the Pos

TABLE X
OUTCOME EXPLANATIONS GENERATION BY LIME FOR A RANDOM POSITIVE AND NEGATIVE CASE OF THE THYROID

Positive prediction (100%)			Negative prediction (99%)		
Neg thyroid	**Pos thyroid**	**Value**	**Neg thyroid**	**Pos thyroid**	**Value**
-	0.85<T3<=1.4	1.07	1.4<T3<=2.0	-	1.60
Pregnant<=0	-	0	Pregnant<=0	-	0
-	0.81<T4U<=0.9	0.82	-	FTI>124.0	139.0
Sick<=0	-	0	TT4>116.0	-	124.0
-	Age>72.46	78	-	0.81<T4U<=0.9	0.89
-	FTI>124.0	136.6	Sick<=0	-	0
-	98.6<TT4<=116	111.6	-	On Thyroxine >0	1
-	On Thyroxine<=0	0	-	61<Age<=74.46	63
-	0<Sex<=1	1	TSH<=0.65	-	0.03
TSH>2.80	-	2.99	-	0<Sex<=1	1

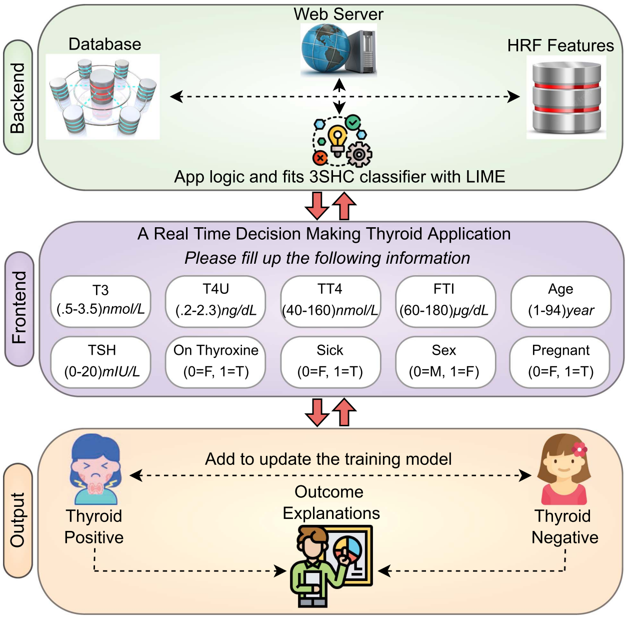


Fig. 5. Development of a novel ML application for thyroid disease detection using the *3SHC* and HRF, as well as integration of LIME for interpretability.

case, and the actual value of T3 (1.07) lies in the affected range between 0.85 and 1.4. Other features like T4U, Age, and FTI, also significantly aid in generating the result. In the case of a Neg sample, the most contribution to the disease is also T3. The value of T3 is 1.60, which is between the recommended ranges of 1.41 and 2.00. Alongside, other feature values, such as Pregnancy, TT4, Sick, TSH, and Sex help to generate a NEG result.

We include the input ranges and units of measurement for each feature presented in Fig. 5. These units are nanomoles per litre (nmol/L), nanograms per deciliter (ng/dL), micrograms of lead per deciliter (μg/dL), milli-international units per litre (mIU/L). This application will be functional by using the following steps:

1) The database updates with the reports.
2) To build input and train the *3SHC* classifier, all attributes are chosen from the HRF-based features.
3) LIME fits with the training model on input features for generating individual explanations.
4) The training model processes a set of attributes.

TABLE XI
ADDITIONAL DETAILS OF ALL SUB-SAMPLES IN TOTAL, NEGATIVE (NEG), AND POSITIVE (POS) SAMPLES (S)

	over-sampled dataset (OSD)			under-sampled dataset (USD)		
Sub-sample	**Total S**	**Neg S**	**Pos S**	**Total S**	**Neg S**	**Pos S**
Male	2516	1260	1256	161	80	81
Female	4566	2281	2285	301	151	150
Children	49	49	0	2	2	0
Young adults	1418	1116	302	74	52	22
Middle age	2714	1471	1243	177	94	83
Older	2901	905	1996	209	83	126

TABLE XII
ACCURACY OF THE *3SHC* DESIGNED ON VARIOUS AGE GROUPS (HRF FEATURES) DEPENDING ON GENDER CATEGORIES FROM BOTH OSD AND USD

	Young adults		Middle age		Older	
	Male	**Female**	**Male**	**Female**	**Male**	**Female**
OSD	98.91%	98.74%	98.73%	99.10%	99.19%	99.52%
USD	96.42%	97.82%	98.55%	99.07%	98.38%	99.31%

5) If the result is thyroid positive, it will suggest the patient to a doctor or undergo additional testing.
6) In order to receive a robust result, data can be uploaded to a database, and the trained model can be updated periodically.

C. Explanatory Model

This section discusses the global behaviors of different data samples from HRF features. We have divided the HRF samples separately into two partitions (i.e., gender and age groups). The gender groups belong to males and females, and the age groups belong to ranges of 1–14 (children), 15–40 (young adults), 41–65 (middle age), and greater than 65 (older) [39], [42]. The representative number of samples of these subsamples is referred to in Table XI.

A Partial Dependence Plot (PDP) is utilized to gain a deeper understanding of the underlying patterns and relationships within these samples. The children group has no positive case in both OSD and USD; hence we do not use this sample as a further explanation. The PDP plots for HRF-based features are depicted in Fig. 6, showcasing various subsamples. To enhance clarity in visualization, the OSD plots (a)–(j) and USD plots (k)–(h) have been generated using distinct color combinations. Specifically, blue, orange, deep-pink, green, purple, and red are used for OSD, while black, teal, olive, navy, lime-green, and brown are used for USD. These plots depict the partial dependence of specific input features on the *y*-axis, while the *x*-axis represents the range of values, with minor ticks indicating distinct characteristic values. As shown in Fig. 6(a) and 6(k), the generated minor ticks are quite condensed, and the partial lines are upperside-arisen when the T3 (triiodothyronine) values are low. Hence, we can infer that a lower value of T3 contributes to the risk of the disease. Next, the thyroid's T4U (thyroxine utilization rate) poses the most significant risk when its values range is comparatively lower. In contrast, higher TT4 values have a minimal impact on the disease risk [see Fig. 6(c) and 6(m)]. Notably, intermediate FTI (free thyroxine index) values were found to pose the most significant threat to the disease. Then, in terms of Fig. 6(e) and 6(o), it is

TABLE XIII
PERFORMANCE COMPARISON BETWEEN OUR APPROACH AND EXISTING METHODS

Author and year	Data size (row, column)	Data source	Num of classes	Normalization, Feature selection	Best performed classifier	Perform ACC	Perform AUC	Performance gain by our findings
Sengupta et al. [3]	3772, 29	University of California Irvine	2	Min-max, -	Random Forest	99.14%	99.80%	0.15%, 0.10%(AUC) increase
Chaganti et al. [4]	9172, 31	University of California Irvine	5	- , Extra tree	Random Forest	99%	-	0.30% increase
Alyas et al. [5]	3163, 29	University of California Irvine	2	-, Feature contribution	Random Forest	94.80%	-	4.73% increase
Aversanoa et al. [6] 2021	2784, 27	AOC Federico II	3	Mean, -	Extra Tee	84%	-	18.2% increase
Hu et al. [7]	176727, NA	Wakayama, Gunma, Hidaka, and Kuma Hospital	2	-, Feature importance	Gradient Boost	-	93.8%	6.50%(AUC) increase
Salman and Sonuc [8]	1250, 17	Iraqi people	3	-, -	Random Forest	98.93%	-	0.36% increase
Kumar et al. [10]	4152, 18	University of California Irvine	3	-, DE-BOA	FCM	94.3%	-	5.29% increase
Sultana and Islam [11]	2800, 28	University of California Irvine	2	-, LAS	Random Forest	99%	-	0.03% increase
Naeem et al. [12]	3371, 33	Kaggle	2	-, -	Support Vector Machine	84.72%	-	17.19% increase
Olatunji et al. [13]	218, 15	King Fahad Specialist Hospital	2	-, Correlation coefficient	Random Forest	90.91%	-	9.21% increase
Alnaggar et al. [14]	7200, 21	University of California Irvine	3	- , Feature importance	XGBoost	99%	-	0.30% increase
Solmaz et al. [15]	7200, 21	University of California Irvine	3	-, -	Ensemble (AB and DT)	99.08%	-	0.21% increase
Dharamkar et al. [16] 2020	7547, 30	University of California Irvine	2	-, -	Voting (C4.5 and RF)	96%	-	3.42% increase
Xie et al. [17]	9172, 26	University of California Irvine	6	-, -	Ensemble (Stacking)	92.3%	99.6%	7.57%, 0.30%(AUC) increase
Islam et al. [18]	3163, 25	University of California Irvine	2	-, Correlation coefficient	ANN	95.87%	-	3.57% increase
Savcı and Nuriyeva [19]	7200, 22	University of California Irvine	3	Min-max, Correlation coefficient	ANN	98%	-	1.31% increase
Li et al. [20] 2020	3163, 25	University of California Irvine	2	-, Associate rule mining	AR-ANN	95.58%	-	3.89% increase
Arjaria [21] 2022	215, 5	University of California Irvine	3	-, -	Logistic Regression	90.77%	-	9.38% increase
Lu et al. [22] 2023	6497, 34	Taipei Medical University Shuang Ho Hospital	2	-, Recursive feature elimination	XGBoost	92.3%	93.4%	7.57%, 6.96%(AUC) increase
Hossain et al. [23] 2023	3221, 30	University of California Irvine	4	-, Feature importance	Random Forest	91.42%	-	8.61% increase
Our Study	**3772, 30**	**Kaggle**	**2**	**Min-max, High-Risk factor**	**3SHC**	**99.29%**	**99.90%**	**-**

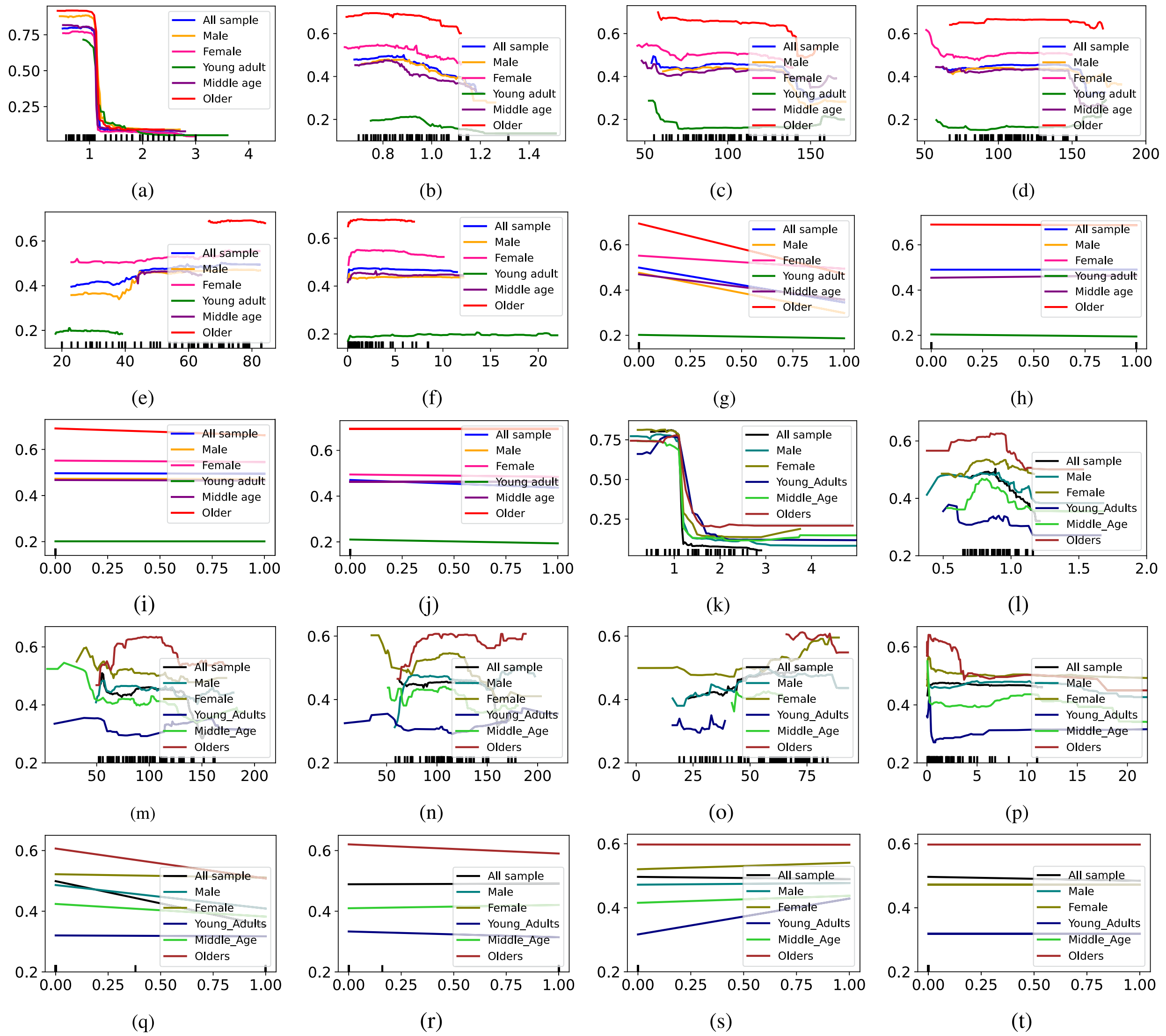
clearly visible that the partial lines of older samples are quite higher than others. Therefore, we can conclude that older folks have been more severely affected by this disease than other age groups. Subsequently, Fig. 6(f) and 6(p) reveals the lower value of TSH (thyroid-stimulating hormone) mostly contributes to high thyroid risks. Categorical features are encoded as 1 for true and 0 for false, and the male is represented as 0 while the female is 1. The Sex plots [see Fig. 6(h) and 6(r)] did not display the PDP lines for male and female samples, similarly, in the Pregnant plots [see Fig. 6(j) and 6(t)], the male data line is absent. This is due to the fact that the mentioned sub-samples consist entirely of a single class, either male or female in the case of the Sex attribute, and only false case in the Pregnant characteristic for male samples. Moreover, plots for On Thyroxine, Sick, and Pregnant assert that both classes (true or false) can be affected by the disease, sometimes true classes are comparatively more affected than false. Furthermore, it is visible in all PDP plots (both OSD and USD) that the plotted line values in older people are much higher than in other subsamples. Regarding gender categories, the female data lines are considerably higher than the male categories. Therefore, we can infer that older persons are at higher risk for this disease than other age groups, and females are more affected than males.

V. DISCUSSION

The timely detection of thyroid disease is paramount, as a delayed diagnosis can have dire consequences, and the associated treatment costs can quickly become exorbitant. In light of these challenges, ML methods have emerged as a vital tool in preventing the disease from rapidly progressing. By enabling early detection, these methods can reduce the cost of thyroid care and treatment. However, the black-box nature of many medical diagnosis models is a cause for concern, especially when human life is at stake. To address this issue, we seek to leverage explainable ML models to provide accurate and transparent predictions that might aid in effectively managing thyroid disease.

We employed three feature selection techniques, including FIS, IGS, and LAS, to generate feature sets depending on the recommended medical references. We then selected the common features across the feature sets to construct a new feature set called the HRF, which could be identical in identifying potential high-risk features. Afterward, hybrid models and traditional ML algorithms were applied to classify the disease. However, the generalizability of the model may be hindered by unforeseen data circumstances, which could result in overfitting and underfitting issues in classification models. Overfitting can occur when a function is too closely matched to a small number of data points. At the same time, underfitting happens when the model cannot accurately map inputs and outputs during training due to a significant training error. Indicators of underfitting include high bias and low variance during the training process [43].

This study undergoes several preprocessing stages to ensure the cleanliness and precision of the training data, followed by proposing the *3SHC* and *3SHANN* by integrating multiple



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k) (l) (m) (n) (o) (p) (q) (r) (s) (t)

Fig. 6. PDP implemented both over-sampled (OSD) and under-sampled datasets (USD) on the High-Risk Factor (HRF) features (a) OSD-T3, (b) OSD-T4U, (c) OSD-TT4, (d) OSD-FTI, (e) OSD-Age, (f) OSD-TSH, (g) OSD-On Thyroxine, (h) OSD-Sex, (i) OSD-Sick, (j) OSD-Pregnant, (k) USD-T3, (l) USD-T4U, (m) USD-TT4, (n) USD-FTI, (o) USD-Age, (p) USD-TSH, (q) USD-On Thyroxine, (r) USD-Sex, (s) USD-Sick, (t) USD-Pregnant.

efficient ML methods. In the initial phase of the *3SHC*, *VT* is employed to aggregate predictions from various traditional classifiers. However, it is acknowledged that this approach is susceptible to biases and over-fitting when the base classifiers are complex [25]. To tackle these issues, we introduced the *BG* technique that reduces the model's tendency to rely on specific patterns [28]. Finally, the *ST* technique further enhances the ensemble's performance by leveraging the diverse predictions of individual classifiers [29]. By employing a meta-classifier on the outputs of individual classifiers, the *ST* learns to weigh and combine their predictions effectively. We argue that by substantially capitalizing on the strengths of diverse ensemble methods, the proposed classifier can enhance the performance significantly. Considering these various efficient approaches, our objective is to tackle the challenges associated with overfitting and high bias. On the other hand, combining the predictions of three individual *ANN* models, we aimed to develop a *3SHANN* classifier, where the final prediction is determined by the class label that receives the majority of votes from the individual models. Using multiple *ANNs* allows us to excel in learning various features or patterns and capture a comprehensive representation of the data [31]. Also, a dropout approach was employed, wherein neurons were randomly removed to prevent excessive co-adaptation during training. Several epochs were executed using an early stopping optimization technique to reduce over-fitting even further, which terminated the training process when no improvements were observed on a validation set [43]. By considering them in our training models, we hypothesize that our proposed framework will yield a highly generalizable model with a stable and diverse set of forecasts.

In the previous experimental section, we assessed the capabilities of our proposed classifiers using various performance indicators. The *3SHC* consistently demonstrated robust results across all performance metrics, including, ACC, PRE, REC, F1, and AUC. Notably, the classifier accurately detected around 1406 out of 1416 test samples for the OSD and 91 out of 92 test samples for the USD, showcasing its significant contribution to disease management based on numerical test reports. We calculate the accuracy in male and female samples from different age groups to get additional insight. Table XII shows that the *3SHC* produces a robust result in each sample. Moreover, a comparison is shown in Table XIII between our study and state-of-the-art models, where we obtained an improved ACC of 99.29% with a 99.90% robust AUC score, surpassing the outcome of prior existing methods. These findings partially support our claim and enhance the credibility of the proposed method.

Furthermore, our proposed application can efficiently diagnose whether a particular person is affected by the disease and show the contributions of each feature value to the predicted outcome. The stakeholders could swiftly determine which features or cases are most affected by the disorder using the proposed system and proceed with treatment accordingly. This application can help to improve diagnosis, treatment, and effective management, and reduce the cost of treatment. Finally, PDP was used to analyze global behaviors of preferred features with some data subsamples. It is apparent from the PDP that aged and female individuals are highly affected by the disease, which notably supports the literature [2].

VI. Conclusion

Since thyroid disease has risen in prevalence worldwide, practitioners find it more challenging to predict the illness by categorizing their patients. As a result, to mitigate the rising impact of the disease, we have proposed an ML-based intelligent disease prediction system with the aid of essential features based on the recommended medical references, followed by oversampling and undersampling methods to prepare a balanced dataset. Experimental analysis reveals that the *3SHC* plays a crucial role in the suggested medical-based HRF features compared to others. Afterward, to explain each feature's contribution to the prediction and clarify, which attributes are more responsible for the prognosis, LIME is used. Next, we use a PDP approach for each subsample, for instance, gender (e.g., male and female) and age categories (e.g., young adults, middle age, and older). Later, our proposed *3SHC* achieves an ACC of 99.29%, which outperforms the state-of-the-art models and shows that elder and females are more affected by thyroid disease. However, the proposed *3SHC* requires higher computational resources and entails increased computational costs compared to single-classifier methods. One possible way to address this problem is to use a distributed learning mechanism, which we aim to explore in our future studies. We also plan to integrate our approach into a blockchain network to enable the security and accessibility of information from any hospital, clinic, or healthcare setting. Furthermore, we are planning to conduct an *in-silico* simulation study to compare the performance of our proposed method with existing approaches under varying sample sizes and degrees of missingness, different data distributions, and levels of complexity factors. By incorporating these conditions, we will aim to gain insights and make meaningful comparisons with previously reported methods in the existing literature.

References

[1] G. Bereda, "Definition, causes, pathophysiology, and management of hypothyroidism," *Mathews J. Pharmaceut. Sci.*, vol. 7, pp. 1–5, 2023.

[2] P. V. Voulgari et al., "Thyroid dysfunction in Greece: Results from the national health examination survey emeno," *PLoS One*, vol. 17, no. 3, 2022, Art. no. e0264388.

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

[4] R. Chaganti, F. Rustam, I. D. L. T. 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, 2022, Art. no. 3914.

[5] 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, 2022.

[6] L. Aversanoa et al., "Thyroid disease treatment prediction with machine learning approaches," in *Procedia Comput. Sci.*, vol. 192, pp. 1031–1040, 2021.

[7] M. Hu et al., "Development and preliminary validation of a machine learning system for thyroid dysfunction diagnosis based on routine laboratory tests," *Commun. Med.*, vol. 2, no. 1, 2022, Art. no. 9.

[8] K. Salman and E. Sonuç, "Thyroid disease classification using machine learning algorithms," in *J. Phys., Conf. Ser.*, vol. 1963, 2021, Art. no. 012140.

[9] R. Das, S. Saraswat, D. Chandel, S. Karan, and J. S. Kirar, "An AI driven approach for multiclass hypothyroidism classification," in *Proc. Adv. Netw. Technol. Intell. Comput.*, Cham: Springer International Publishing, 2022, pp. 319–327.

[10] S. J. K. J. Kumar, P. Parthasarathi, M. Masud, J. F. Al-Amri, and M. Abouhawwash, "Butterfly optimized feature selection with fuzzy c-means classifier for thyroid prediction," *Intell. Automat. Soft Comput.*, vol. 35, pp. 2909–2924, 2023.

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

[12] A. B. Naeem et al., "Hypothyroidism disease diagnosis by using machine learning algorithms," *Int. J. Intell. Syst. Appl. Eng.*, vol. 11, no. 3, pp. 368–373, 2023.

[13] S. O. Olatunji et al., "Early diagnosis of thyroid cancer diseases using computational intelligence techniques: A case study of a Saudi Arabian dataset," *Comput. Biol. Med.*, vol. 131, 2021, Art. no. 104267.

[14] M. Alnaggar, M. Handosa, T. Medhat, and M. Z. Rashad, "Thyroid disease multi-class classification based on optimized gradient boosting model," *Egyptian J. Artif. Intell.*, vol. 2, no. 1, pp. 1–14, 2023.

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

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

[17] Y. Xie et al., "Thyroid disease diagnosis based on feature interpolation and dynamic weighting ensemble model," 2023. Accessed: 2023. [Online]. Available: https://doi.org/10.21203/rs.3.rs-2851005/v1

[18] 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, 2022, Art. no. e898.

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

[20] D. Li, D. Yang, J. Zhang, and X. Zhang, "AR-ANN: Incorporating association rule mining in artificial neural network for thyroid disease knowledge discovery and diagnosis," *IAENG Int. J. Comput. Sci.*, vol. 47, no. 1, pp. 25–36, 2020.

[21] S. K. Arjaria, "Developing an explainable machine learning-based thyroid disease prediction model," *Int. J. Bus. Analytics*, vol. 9, pp. 1–18, 2022.

[22] Y.-T. Lu, H.-J. Chao, Y.-C. Chiang, and H.-Y. Chen, "Explainable machine learning techniques to predict amiodarone-induced thyroid dysfunction risk: Multicenter, retrospective study with external validation," *J. Med. Internet Res.*, vol. 25, 2023, Art. no. e43734.

[23] M. B. Hossain et al., “An explainable artificial intelligence framework for the predictive analysis of hypo and hyper thyroidism using machine learning algorithms,” *Hum.-Centric Intell. Syst.*, vol. 3, pp. 211–231, 2023.

[24] T. Garg, S. Masud, T. Suresh, and T. Chakraborty, “Handling bias in toxic speech detection: A survey,” *ACM Comput. Surv.*, vol. 55, pp. 1–32, 2023.

[25] H. Liu and M. Cocea, “Granular computing-based approach for classification towards reduction of bias in ensemble learning,” *Granular Comput.*, vol. 2, pp. 131–139, 2017.

[26] C. Lin, J. Xu, J. Hou, Y. Liang, and X. Mei, “Ensemble method with heterogeneous models for battery state-of-health estimation,” *IEEE Trans. Ind. Inform.*, vol. 19, no. 10, pp. 10160–10169, Oct. 2023.

[27] M. Aljasim and R. Kashef, “E2DR: A deep learning ensemble-based driver distraction detection with recommendations model,” *Sensors*, vol. 25, no. 5, 2022, Art. no. 1858.

[28] J. H. Eun, K. S. Ho, J. J. Seok, and O. J. Hee, “Visual attributes of thumbnails in predicting youtube brand channel views in the marketing digitalization era,” *IEEE Trans. Computat. Social Syst.*, to be published, doi: 10.1109/TCSS.2023.3289410.

[29] W. Zengshuai, Z. Minhua, and L. P. Xiaoping, “A novel classification method based on stacking ensemble for imbalanced problems,” *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–13, 2023.

[30] M. E. C. Bento, “Physics-guided neural network for load margin assessment of power systems,” in *IEEE Trans. Power Syst.*, to be published, doi: 10.1109/TPWRS.2023.3266236.

[31] L. K. C. et al., “Determination of vehicle loads on bridges by acoustic emission and an improved ensemble artificial neural network,” *Construction Building Mater.*, vol. 364, 2023, Art. no. 129844.

[32] R. I. Hamilton and P. N. Papadopoulos, “Using SHAP values and machine learning to understand trends in the transient stability limit,” in *IEEE Trans. Power Syst.*, to be published, doi: 10.1109/TPWRS.2023.3248941.

[33] *Thyroid dataset*, 2021. Accessed: 2022. [Online]. Available: https://www.kaggle.com/datasets/bidemiayinde/thyroid-sickness-determination

[34] A. Raghuvanshi et al., “Intrusion detection using machine learning for risk mitigation in IoT-enabled smart irrigation in smart farming,” *J. Food Qual.*, vol. 2022, pp. 1–8, 2022.

[35] G. Huang, “Missing data filling method based on linear interpolation and lightgbm,” in *J. Phys., Conf. Ser.*, vol. 1754, 2021, Art. no. 012187.

[36] C. Cecchini and MD, “Thyroid function tests: Interpreting your results,” 2021. Accessed: 2022. [Online]. Available: https://www.goodrx.com/health-topic/diagnostics/thyroid-function-test-interpretation

[37] R. Kerslake, “How to understand thyroid panel test results,” 2022. Accessed: 2022. [Online]. Available: https://www.singlecare.com/blog/normal-thyroid-levels

[38] S. Xiang, J. D. Yuhan Cao, C. Li, J. Qiu, and X. Li, “The association between urinary phthalate metabolites and serum thyroid function in us adolescents,” *Sci. Rep.*, vol. 13, 2023, Art. no. 11601.

[39] P. O. Carrilloa et al., “Definition of reference ranges for free t4, tsh, and thyroglobulin levels in healthy subjects of the jaén health district,” *Endocrinología, Diabetes y Nutrición (English ed.)*, vol. 64, no. 8, pp. 417–423, 2017.

[40] H. K. Tripathy, L. Jena, S. Mishra, P. K. Mallick, and G.-S. Chae, “Stacked KNN with hard voting predictive approach to assist hiring process in it organizations,” *Int. J. Elect. Eng. Educ.*, 2021, Art. no. 0020720921989015.

[41] J. P. Consuegra-Ayala, Y. Gutiérrez, Y. Almeida-Cruz, and M. Palomar, “Intelligent ensembling of auto-ml system outputs for solving classification problems,” *Inf. Sci.*, vol. 609, pp. 766–780, 2022.

[42] J. Kim, N. Prasitlumkum, S. Randhawa, and D. Banerjee, “Association between subclinical hypothyroidism and incident hypertension in women: A systematic review and meta-analysis,” *J. Clin. Med.*, vol. 10, no. 15, 2021, Art. no. 3314.

[43] P. Ghosh et al., “SkinNet-16: A deep learning approach to identify benign and malignant skin lesions,” *Front. Oncol.*, vol. 12, 2022, Art. no. 931141.