The Scientific Temper (2025) Vol. 16 (2): 3758-3773
E-ISSN: 2231-6396, ISSN: 0976-8653
Doi: 10.58414/SCIENTIFICTEMPER.2025.16.2.08
https://scientifictemper.com/

RESEARCH ARTICLE

Advanced hybrid feature selection techniques for analyzing the relationship between 25-OHD and TSH

P. Vinnarasi, K. Menaka*

Abstract

The process of selecting essential features from high-dimensional datasets is a crucial task while handling biological data. It is not just about choosing the right features but also ensuring that the selected features consistently perform well across different datasets or under varying conditions. The selection of features is crucial for the development of a machine learning algorithm, as it influences the model's characteristics and its connection to physiological processes, which is essential in the healthcare sector for identifying illness states with minimal data. The present study is primarily concerned with addressing the challenge of finding the features that are highly correlated with the thyroid stimulating hormone (TSH) from the thyroid datasets since TSH has been regarded as a primary cause of many ailments. It also tries to find the impact of Vitamin D (25OHD) - 25-Hydroxyvitamin D on TSH. The performance and accuracy were greatly improved after the reduction of data dimension with feature selection techniques. The use of feature selection algorithms for healthcare problems could help reduce diagnosis costs and thereby enhance the ability of the healthcare system for accurate and prompt identification of diseases. This work developed two hybrid feature selection techniques (FST - CorrRecursive Feature Selection (CRFS) and RanChi Ensemble Selection (RCES)), combining the specialties of filter and wrapper methods for identifying the influence of vitamin D and other features on the thyroid. Other existing feature selection methods have also been attempted. The findings demonstrated that, when compared to other approaches, our proposed CRFS and RCES techniques produced superior outcomes.

Keywords: Vitamin D & thyroid, Feature selection techniques, Filter and wrapper methods, Hybrid feature selection techniques.

Introduction

Nowadays, a significant portion of the population grapples with various health issues. It is imperative for healthcare entities to tackle the challenging task of investigating and identifying these problems. Hospitals possess a wealth of clinical information, but without an intelligent system, retrieving medical data from these organizations can be quite challenging. Despite the continuous advancements in technology, machine learning (ML) algorithms have proven to be highly effective in disease diagnosis (Thomas Davenport., 2019). Utilizing machine learning techniques, various prediction models could be created for diagnosing various diseases, including but not limited to breast cancer, brain tumors, ovarian cancer, lung cancer, and heart issues. Recent studies highlight vitamin D insufficiency and thyroid disease as the two most commonly encountered conditions on a global scale.

In recent years, there has been a notable increase in the growth of high-dimensional data. Uncovering patterns and features in the patient data system posed significant challenges in identifying thyroid patients. Given the necessity for high-dimensional data to discern infected patients, machine learning plays a pivotal role in disease prediction. Disease detection typically involves numerous tests and clinical examinations. However, the application of ML algorithms employed in the classification of healthcare data has an opportunity to improve the testing procedure, therefore lowering the number of tests needed. This streamlined test suite contributes significantly to both performance and time efficiency. The extraction and selection of features in healthcare data hold paramount importance as they enable clinicians to pinpoint the most

Department of Computer Science, Urumu Dhanalakshmi College [Affiliated to Bharathidasan University], Tiruchirappalli, Tamil Nadu, India.

***Corresponding Author:** K. Menaka, Department of Computer Science, Urumu Dhanalakshmi College [Affiliated to Bharathidasan University], Tiruchirappalli, Tamil Nadu, India., E-Mail: k.menaka@udc.ac.in

How to cite this article: Vinnarasi, P., Menaka, K. (2025). Advanced hybrid feature selection techniques for analyzing the relationship between 25-OHD and TSH. The Scientific Temper, **16**(2):3758-3773.

Doi: 10.58414/SCIENTIFICTEMPER.2025.16.2.08

Source of support: Nil

Conflict of interest: None.

Received: 19/11/2024 **Accepted:** 13/01/2025 **Published** : 20/03/2025

crucial features for diagnosis. This, in turn, facilitates more efficient disease diagnosis and early detection. Feature selection algorithms play a crucial role in enhancing accuracy in classification thereby minimizing the execution time of the classification system (Bassam Abdo Al-Hameli d *et al*., 2021).

The significance of feature selection has grown substantially with the advancements in machine learning. Through exploratory data analysis, redundant and duplicate datasets are eliminated, leading to the identification of key features influencing target prediction via feature selection. Utilizing appropriate approaches for feature selection based on ordinal and categorical data establishes the pivotal features. Overall, the model achieves higher accuracy and faster execution, as these techniques demand fewer computations (Raid Alzubi *et al*., 2017). When creating models, such as choosing variables or predictors, there is a need to select a section of essential attributes from datasets with extensive space. This procedure aids in streamlining the machine-learning process by narrowing down the hypothesis space and eliminating uninteresting features from the provided data.

In general, three main techniques are used in the process of selecting features. They are filters, wrappers, and embedding techniques. A specific classification process, choosing the most crucial data or a subset of qualities is not involved in filter strategies. Chi-square, information gain, and relief are three methods for evaluating filter algorithms. In contrast, wrapper strategies consider a classification algorithm to choose a feature subset suitable for that algorithm. Embedded approaches incorporate feature selection into the model fitting process, allowing the model to select the optimum methodology. In both pattern recognition and ML, the procedure of feature extraction or selection holds paramount importance. The implementation of feature selection strategies not only reduces computational costs but also has the potential to enhance classification performance. Effectively representing data from all features creates a considerable dispute in the machine learning process. All initial features will not be helpful for classification or regression problems. Some features within the dataset are redundant, unnecessary, or merely noise, which can adversely impact classification performance. Applying the process of feature selection to these problems becomes essential in improving classification performance and decreasing the computational cost of the classifier (Saba Bashir *et al*., 2022).

The success of a machine learning-based system centers on the implementation of a suitable feature selection strategy. This technique significantly diminishes the dimensionality of the feature space while preserving an accurate representation of the original data. Among its key benefits are enhanced classification performance, accelerated learning rates, simplified data interpretation, and more accurate projections. Nevertheless, to choose the optimal feature subsets, these algorithms need to adjust their parameters, thereby amplifying computing complexity (Doppala, B.P.*et al*., 2021). This work utilizes the feature selection methods employed in developing a system for recognizing crucial characteristics. Utilizing highly efficient feature selection strategies, this work has constructed a non-invasive, multi-parametric system for identifying vitamin D deficiency in the thyroid (Y. Saeys.,2007).

The present investigation was carried out by taking data from the open-source dataset of (Danilovic *et al*., 2021). In this proposed system, feature selection methods such as 1) Filter chi-square method, mutual information, information gain, 2) Wrapper-recursive feature elimination (RFE), sequential feature selection (SFS), forward selection, and 3) Ensemble-bagging, random forest, and boosting were employed.

In determining the effect of the 25OHD (Vitamin D) and other characteristics on the thyroid, two hybrid feature selection strategies, namely, CorrRecursive feature selection (CRFS) and RanChi ensemble selection (RCES) were developed in the proposed work. These techniques merge the strengths of filter and wrapper methods. In order to predict the onset of autoimmune illnesses like thyroid for a human, hybrid techniques like CorrRecursive feature selection (CRFS) choose 25OHD as the most crucial feature among all the factors including age, BMI, height, etc. The outlined selection techniques are crucial for the creation of an effective machine-learning model. Given the sizable population affected by this condition, the feature selection algorithms proposed in this work hold potential for application in various medical contexts with extensive datasets.

Literature Survey

Imad R *et al*., 2017, made a study to look at the serum 25-OH level of vitamin D in female hypothyroidism patients. A case-control study with 58 participants in each arm was conducted. Women with hypothyroidism were the cases, whereas women with good health comprised the control group. Each participant's levels of thyroid stimulating hormone (TSH), hormones T3 and T4 and hemoglobin were measured. The level of serum vitamin D (25-OH) was determined using spectral analysis. According to their findings, there was no discernible variation in vitamin D levels between the female hypothyroid subjects and the healthy controls.

Yong Guo. *et al*, 2020, proposed a work in which the lack of vitamin D has been related to issues with the thyroid. The objective of their study was to look into the connection between serum 25(OH)D levels and early childhood thyroid function measures. They also made measurements in identifying the measures using electro-chemiluminescence (ECL) immunoassay. They drew the conclusion that there are

no direct correlations seen between serum 25(OH)D levels and thyroid-related parameters (TSH, FT3, and FT4) though they are related to some extent.

Jie Kuang Zhijian Jin *et al*., 2022 proposed a work in which lower levels of the D vitamin and distorted local metabolism of it have been related to the frequency and aggressiveness of several cancers. In addition, the role of vitamin D metabolism in papillary thyroid cancer (PTC) advancement was examined using a tissue microarray. Based on their research, there might not be a connection between the serum 25(OH)D level and the thyroid nodule risk assessment method.

Another work recommended by Mirjana Babic Leko *et al*., 2023 formulated the connection between thyroid function and vitamin D. This review addressed human studies that explored two specific topics: (1) the relation between vitamin D level and thyroid function and (2) the impacts of adding vitamin D to one's diet on thyroid function. Ultimately, they came to the conclusion that despite significant advancements in our understanding of the relationship between the two, there is a great deal of heterogeneity in the studies, making it challenging to reach a firm judgment.

Tereza Planck *et al*., 2018, proposed a work to discover the relationship between vitamin D in Graves disease. The motto of their work was to find the association between laboratory and clinical indicators in graves disease (GD) and vitamin D levels and to compare GD with those of the general population. Despite the fact that the intensity of 25OHD was small in GD patients when compared with others in the population, they determined this decrease had no impact on the disorder's molecular or clinical characteristics.

(Nino Turashvili *et al*, 2021) proposed a work with a belief that a lack of vitamin D may be a key predictor for the onset of chronic autoimmune thyroiditis. The aim of their research work was to examine the vitamin D levels among people who had persistent autoimmune thyroiditis. They noticed some negative correlations between antithyroid peroxidase and antithyroglobulin with vitamin D.

Sandeep Appunni *et al*, 2021, built an experiment to investigate the relationship between vitamin D and hypothyroidism. They utilized an extensive population-based dataset. They classified the participants into three clinically significant groups: deficient, intermediate, and optimum, based on their status of vitamin D in which weighted multivariable logistic regression analyses were performed to determine the probabilities of hypothyroidism. They finally identified that there exists some connection between low vitamin D levels and autoimmune hypothyroidism. They concluded that long-term healthcare initiatives like routinely screening at-risk groups for vitamin D insufficiency could significantly lower the hypothyroidism risk.

Quan Li *et al*, 2023, performed a study in which the relationship between serum 25-hydroxyvitamin D deficiency and thyroid disease in postmenopausal women with type-2 diabetes mellitus was determined. For the purpose of comparison analysis, chi-square as well as t-test was performed. By using the Pearson correlation method, the interaction among various biomarkers of thyroid function and 25(OH)D was evaluated. They looked at impending risk variables for 25(OH)D insufficiency using multimodal logistical regression analysis. They observed that hyperthyroidism and hypothyroidism were notably associated with the presence of a 25(OH)D deficit in postmenopausal women.

Filip Lebiedzi *et al*, 2023, evaluated the impact of vitamin D on the immunopathology of Hashimoto's thyroiditis (HT). This study suggests how several immunological mechanisms in HT may be impacted by vitamin D. They remarked that the data that was used for the study points to a possible function for vitamin D in the management and prevention of HT, but additional study is required to completely comprehend its systems of action and possible benefits for treatment. They concluded that among other immune-regulating effects on HT, vitamin D may affect Th17/Treg cell ratio balance, decrease thyroid class II HLA gene expression, avoid a surplus of B-cell responses, and affect DC-dependent T-cell activation. Although vitamin D function in hypertension (HT) can be better understood in the testing facility, real-world clinical outcomes such as the connections between thyroid, immune response, and symptom manifestation as well as the development and conveying of the disease must also be examined.

Lutfiye Secil *et al*. 2019 conducted an investigation to find the association between serum vitamin d levels and thyroid. The study entailed the review of the file records of a few individuals who tested positive for anti-TPO and had concomitant thyroid function tests and vitamin D assays. Depending on their serum vitamin D levels, the participants were split into two groups: Group 1 had patients whose levels were ≥20 ng/mL, while group 2 had patients whose levels were <20 ng/mL. No other differences were found in the groups' serum levels of free thyroxine (fT4) or thyroid-stimulating hormone (TSH). Vitamin D had a positive correlation with TSH as per the study.

Fathania Firwan Firdaus *et al*, 2020, conducted a review for predicting heart disease with the aid of feature selection and classification approaches of ML. The purpose of this review is to show how computer technology may detect cardiac sickness and save costs and time. They suggested that the characteristics most strongly associated with heart disease could be found through feature selection and the wrapper, filter, hybrid, and embedding are some of them. As per their view, the filter approach performs reasonably well in terms of computation. The hybrid approach generally makes use of several techniques. They reached the premise that the hybrid approach takes advantage of each feature

selection method by combining several of them to provide optimal outcomes.

Faisal Saeed *et al*, 2022, proposed an approach to enhancing the prediction of Parkinson's disease using the feature selection strategies of ML algorithms. This study seeks to offer a complete methodology with many machine-learning approaches, to improve the prediction of Parkinson's disease.

Usha *et al*, 2021, presented a work for predicting heart disease using feature selection techniques based on a data-driven approach. The focus of this study was to provide a feature selection technique-based approach for the prediction of heart disease. The forecast aids doctors in making precise decisions about the well-being of their patients. The suggested model was trained using techniques for data transformation, pre-processing, and collection. In order to improve the prediction of heart disease categorization, this model made use of feature selection approaches, specifically filter and wrapper with classification algorithms. Ada Boost, random forest, decision tree, and logistic regression are practical methods for assessing performance indicators. Accuracy, F1-score, precision, and sensitivity are among the performance measures; specificity indicated an improvement in the prediction's results. This work provided an efficient machine learning-based method for heart disease analysis. With the aid of machine learning classifiers, the task was constructed. Ultimately, they draw the conclusion that the application of feature selection methods to identify the relevant features, which improved classification accuracy while also lowering the computation of the diagnosis process, was another innovative aspect of this research.

Achina *et al*., 2022, devised a model for sentiment classification using hybrid feature selection and ensemble classifier. They combined some of the currently utilized feature selection techniques, such as information gain (IG), Chi-square (CHI), and GINI index (GINI), with a genetic algorithm. As previously indicated, they first collected features from the three distinct selection methods and subsequently used the UNION SET operation to extract the smaller feature set. The feature set is then further optimized by applying the genetic algorithm. Additionally, an ensemble approach based on the error rates found in various domain datasets was presented in this study. They have taken into consideration four support vector machine (SVM) classifier versions in order to test their suggested hybrid feature selection and ensemble classification technique. The experimental findings demonstrated that in all three domain datasets, the suggested strategy outperformed the others. Additionally, they provided the T-test for statistical significance comparing classifiers, and model execution duration, precision, recall, F1-score, and AUC were also compared.

Materials and Methods

System Framework

The aim of the present work is to improve the classification accuracy with the minimized number of features within the dataset that is being used. The framework of the present study comprises vital stages, viz. the collection of data, pre-processing of data, strategies of feature selection (FS), training and evaluation of models' performances and outcomes. This work spots the critical role of FS strategies in accurately categorizing vitamin D deficiency and its role in hypothyroidism and other associated disorders.

Data Collection

In this work, two datasets were exploited to understand the effect of different factors on thyroid health. The first dataset, referenced in (Danilovic *et al*., 2016) comprises 433 data samples encompassing male as well as female subjects. It has nine features, including patient demographics such as weight, height, sex, age, and season of data collection, alongside laboratory measurements such as 25-hydroxyvitamin D (25OHD) levels and thyroid-stimulating hormone (TSH) levels.

The second dataset, hypothyroid, was sourced from the UCI Machine Learning repository, referenced in Thyroid Data, hosted in Kaggle. It comprises 3,772 instances with 28 features, covering demographic information such as age and sex, as well as indicators of medical history and treatment status, such as being on thyroxine medication, query on thyroxine, and query hypothyroid, etc. This dataset provides comprehensive coverage of thyroid-related factors, facilitating in-depth analysis and modeling of hypothyroidism.

The selection of these datasets was based on data availability and volume, aiming for a balance between sufficient data and similarity in feature composition. Prior

Table 1: Description of the serum 25OHD &TSH dataset

Feature name	*Feature type*	*Description*
Patient	numerical	Patient ID or identifier
Weight	numerical	Weight of the patient
Height	numerical	Height of the patient
BMI	numerical	BMI of the patient
sex	nominal	Gender
Age	numerical	Age
season	nominal	Season when the data was recorded
25OHD	numerical	25-hydroxyvitamin D levels
TSH	numerical	Thyroid-stimulating hormone levels
class	nominal	Class label or target variable

Table 2: Description of the hypothyroid dataset

Feature name	*Feature Type*	*Description*
age	numerical	Age of the patient
sex	nominal	Gender of the patient (Male/Female)
on thyroxine	nominal	Thyroxine medication undertaken?
query on thyroxine	nominal	Is there a query about thyroxine medication
on antithyroid medication	nominal	Antithyroid medication undertaken?
sick	nominal	currently sick?
pregnant	nominal	currently pregnant?
thyroid surgery	nominal	Has the patient undergone thyroid surgery?
I131 treatment	nominal	Patient receive I131 treatment? Top of Form Bottom of Form
query hypothyroid	nominal	query about hypothyroidism for the patient?
lithium	nominal	Whether the patient is currently taking lithium medication
goitre	nominal	Patient have a goitre?
tumor	nominal	Patient been diagnosed with a thyroid tumor? Top of Form Bottom of Form er the patient has a thyroid tumor
hypopituitary	nominal	Manifests symptoms consistent with hypopituitarism." Top of Form Bottom of Form
psych	nominal	psychological condition?
TSH measured	nominal	Thyroid-stimulating hormone (TSH) levels measured? Top of Form Bottom of Form
TSH	nominal	Thyroid-stimulating hormone (TSH) levels
T3 measured	nominal	Whether Triiodothyronine (T3) levels were measured
T3	numerical	Triiodothyronine (T3) levels
TT4 measured	nominal	Whether Total thyroxine (TT4) levels were measured
TT4	numerical	Total thyroxine (TT4) levels
T4U measured	nominal	Whether Thyroxine-binding globulin (T4U) levels were measured
T4U	numerical	Thyroxine-binding globulin (T4U) levels
FTI measured	nominal	Free Thyroxine Index (FTI) levels measured?
FTI	numerical	The levels of the Free Thyroxine Index (FTI).
TBG measured	nominal	Are Thyroxine-binding globulin (TBG) levels measured?
referral source	nominal	What is the source of the patient's referral? 4o Top of Form Bottom of Form Top of Form Bottom of Form
binaryClass	numerical	Binary classification label or target variable

to analysis, preprocessing steps were conducted to handle missing values, normalize features, and ensure data integrity. These datasets collectively offer a rich resource for exploring the relationship between various factors of thyroid disease and its association with vitamin D as shown in Tables 1 and 2.

Data augmentation

In this work, data augmentation techniques were employed to enrich the dataset utilized for machine learning tasks. The initial step involved loading the original dataset from the .CSV file, which comprised various attributes

including patient weight, height, BMI, sex, age, and season of data collection, as well as 25OHD levels and TSH levels. Subsequently, synthetic samples were generated to expand the dataset, with a predetermined number of samples for each attribute. Random values were selected for features such as weight, height, BMI, sex, age, season, 25OHD, TSH, and classification status. These synthetic data points were created to mimic the characteristics of the original dataset and introduce diversity.

The synthesized data was then merged with the original dataset, resulting in a larger and more varied dataset suitable for subsequent analysis and model training. This process aimed to augment the robustness and matching capability of ML algorithms by providing additional training examples. Through the application of data augmentation techniques, this research sought to enrich the reliability and effectiveness of ML methodologies in addressing the targeted research objectives (Yixin Liu *et al*., 2019).

Data Preprocessing

This stage plays a significant part in preparing the dataset for analysis and modeling. This section outlines the steps undertaken to preprocess the dataset before feature selection and model training.

Data preprocessing for thyroid dataset

Data Loading

The dataset is loaded into a Pandas DataFrame using the read_csv() function. This allows the dataset to be manipulated and analyzed within the Python environment.

Handling Missing Values

The '?' values in both datasets are replaced with NaN (Not a Number) using the replace() function, standardizing the representation of missing values for further processing.

For both datasets, missing values in numerical columns were addressed with the replacement of the average values of the corresponding column. In the first dataset, numerical columns 'Patient', 'Weight', 'Height', 'BMI', 'Age', '25OHD', and 'TSH' were handled this way, while in the second dataset, numerical columns 'age', 'TSH', 'T3', 'TT4', 'T4U', and 'FTI' underwent the same treatment. This approach ensures that missing values are imputed with values that preserve the statistical properties of each dataset.

For categorical columns such as 'sex', 'season', and 'class' in both datasets and additional categorical columns in the second dataset such as 'on thyroxine', 'query on thyroxine'

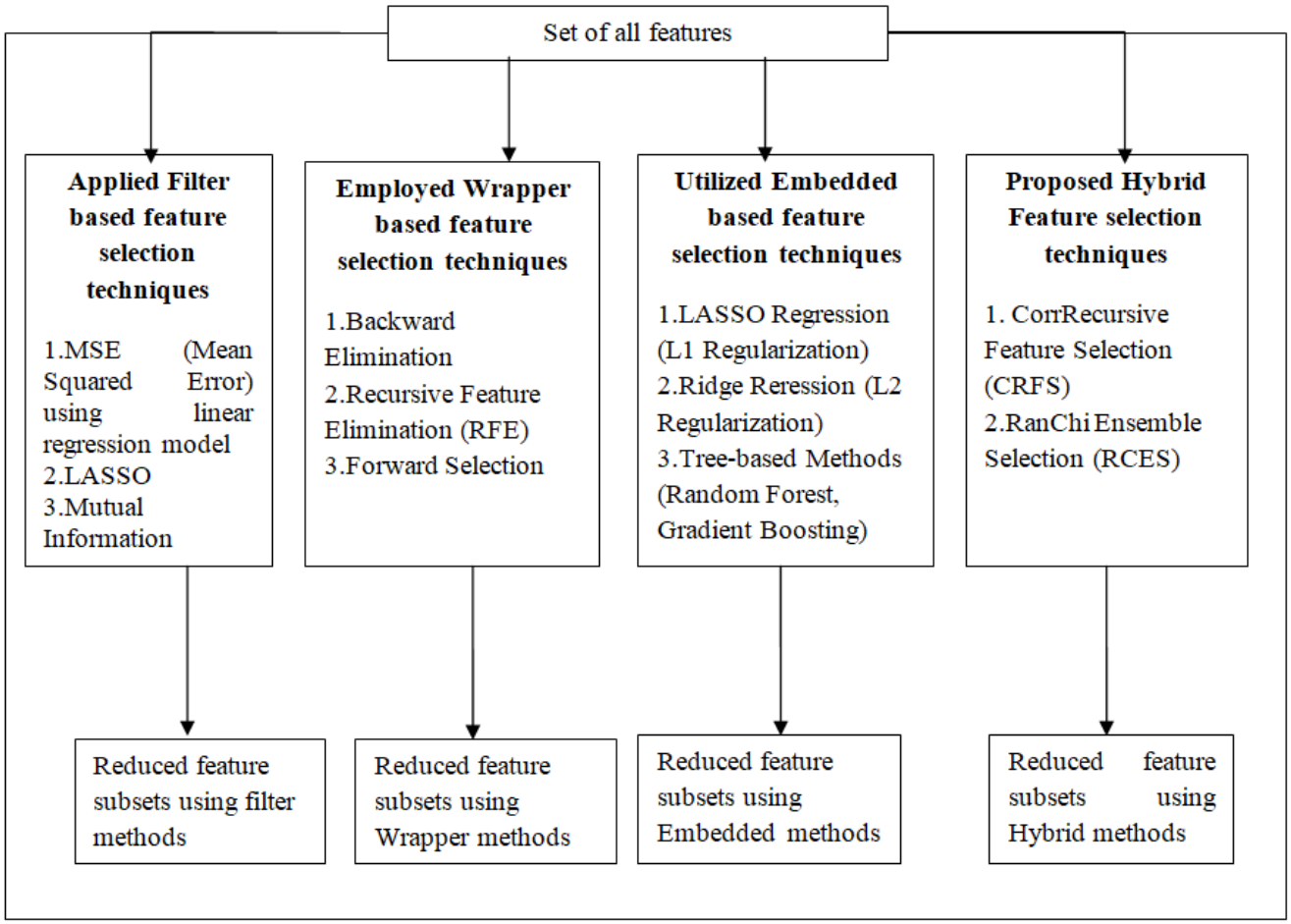

Figure 1: System framework for feature selection

and 'referral source', missing values are replaced with the most frequent value in each column using the mode() function. This strategy ensures that missing categorical values are replaced with values that are representative of the majority within each category.

Data Inspection

Basic information about the dataset is displayed using the info() function. This provides insights into the structure of the dataset, including the number of entries, data types, and presence of missing values.

Feature Encoding

The transformation of categorical features into numerical format is achieved through label encoding. This process entails the application of the fit_transform() function from the LabelEncoder() object to each categorical column within the dataset. By doing so, compatibility with machine learning algorithms that necessitate numerical inputs is ensured.

Splitting of Data

Data Splitting

The *train_test_split()* of the module sklearn.model_ selection splits the dataset into the well-known phases, one for the training phase and another one for the testing phase set. This facilitates the evaluation of model performance on unseen data which partitions the dataset into subsets for training and validation (John Smith *et al*., 2021).

Feature Selection

In this stage, experiments were done to review the effect of feature selection in the analysis of detecting the impact of vitamin D on thyroid disease and the other factors influencing the cause of thyroid disease. It is a well-known fact that this stage of selecting essential features helps construct a more accurate model by eliminating or under-representing less relevant features. This process helps to minimize the training time and thus enhances learning performance. Various feature selection approaches as discussed before were explored in this stage for the present work along with the newly proposed hybrid methods. Techniques from each category were separately applied to the initial datasets, aiming to create optimal feature subsets. The illustration in Figure 1 demonstrates that each of the mentioned techniques (Filter, Wrapper, Embedded, Hybrid) has been applied to the dataset, resulting in reduced subsets of features (Yap Bee Wah *et al*., 2018). These subsets contain the most relevant features according to the respective techniques used. Each technique or combination serves a specific purpose in feature selection, aiming to improve model performance, reduce overfitting, and enhance interpretability.

Selecting appropriate features for the chosen problem is becoming a most challenging task which helps in achieving the optimal outcome in data classification tasks. Though the increase in the number of features is theoretically beneficial for achieving prediction accuracy, the empirical evidence contradicts this notion, as not all features contribute significantly to identifying the data class label. Some attributes are irrelevant to the data label and feature selection removes those attributed from the dataset.

Filtering methods

This method refers to techniques used to preprocess and transform data prior to feeding this into an ML model. These methods are crucial in view of enriching the input data's quality and support for improving the model outcome. Filtering methods are used for feature selection, where inappropriate or superfluous features are removed from the dataset taken. This helps in reducing dimensionality, improving model efficiency, and preventing overfitting (Maryam Khalili *et al*., 2022).

Three particular techniques are used as shown in Figure 1, and they have yielded better results for this study than other filtering techniques. These techniques are:

1. Mean squared error (MSE) using a linear regression model
2. Least absolute shrinkage and selection operator (LASSO)
3. Mutual information

Wrapper Approaches

This method involves strategies in which the training of a prediction model is merged with feature selection. Wrapper methods always choose features depending on a model's prediction performance (Fadi Alharbi *et al*., 2023).

Some of the methods which produced better results for this study are given below:

1. Backward elimination
2. Recursive feature elimination (RFE)
3. Forward selection

Embedded Approaches

Here in this method, feature selection is incorporated into the procedure of training the model itself. This method generally does the process of selecting features in the training process of the model (Souhir Ben Amor *et al*., 2019).

The following embedded methods generated better outcomes than others for this study

1. LASSO regression (L1 Regularization)
2. Ridge regression (L2 Regularization)
3. Tree-based methods (Random Forest, Gradient Boosting)

Proposed Hybrid Methods:

This refers to the task of combining more techniques or models to improve the performance of the classification. For this, the hybrid methods utilize the potentials of the various existing algorithms thereby reducing the limitations of those approaches and trying to produce better results.

Algorithm: Pseudo code of proposed hybrid CRFS algorithm
Input:
- Dataset (df)

Process:
1. Load the dataset.
2. Handle missing values: Replace '?' with NaN and impute missing values.
3. Encode categorical features into numerical format.
4. Divide the dataset into training and testing subsets.
5. Execute Correlation-based Feature Selection (CFS):
   - Calculate the correlation matrix of the training set.
   - Identify highly correlated features and remove them.
6. Execute Recursive Feature Elimination (RFE) using Support Vector Machine (SVM).
   - Initialize SVM model with a linear kernel.
   - Utilize Recursive Feature Elimination (RFE) to select a specified number of features.
   - Apply RFE to select a predefined number of features.
   - Revise the training and testing datasets to contain exclusively the selected features
7. Make predictions on the test set using the trained SVM model.
8. Evaluate the model performance:
   - Compute accuracy, precision, recall, and F1-score.
   - Formulate a confusion matrix.:

Output :selected features and evaluation metrics.

Figure 2: Proposed hybrid CRFS algorithm

In the present work, the following two methods have been proposed for both datasets.

1. CorrRecursive feature selection (CRFS) (Figure 2).
2. RanChi ensemble selection (RCES) (Figure 3).

Results and Discussion

In this comparative analysis, Python has been chosen as the programming language of choice for constructing the analytical model, leveraging Jupyter Notebook within the Anaconda environment. This setup offered several advantages, facilitating dataset exploration and enabling effective pattern identification. Furthermore, the utilization of scikit-feature, enhanced the feature selection process. This open-source library encompasses approximately 40 feature selection algorithms, providing a comprehensive toolkit for model development in Python.

In this study, feature selection has been conducted for two datasets: one related to the serum 25OHD & TSH dataset and another one is the hypothyroid dataset. For identifying the most relevant features related to the study, various feature selection strategies were performed for both datasets. The results observed on the experiments conducted for both datasets are discussed below:

Table 3 shows the comparison of the various performance metrics for all the approaches performed for the serum 25OHD and TSH dataset. In terms of Accuracy, the proposed RCES approach generated the best result with 93% while CRFS generated 90% accuracy. Among the existing approaches, RFE performed well with 89% accuracy. For all the other metrics, the proposed methods yielded good results than the existing methods.

Table 4 shows the comparison of the various performance metrics for all the approaches performed for the hypothyroid dataset. For this data*set al*so, the proposed hybrid methods generated better outcomes than the existing ones and especially the RCES produced better results.

Among all the approaches, the existing LASSO method of the filter approach generated very poor results in terms of all the performance metrics'.

Table 5 shows the best outcomes of all the performance metrics' among all the approaches for the existing and the proposed methods for the serum 25OHD & TSH dataset. It is clearly seen from this table that the proposed hybrid RCES method spawned more good results than the existing ones. This is graphically represented in the following Figure 4.

Table 6 shows the best outcomes of all the performance metrics' among all the approaches for the existing and the proposed methods for the hypothyroid dataset. It is clearly seen from this table that the proposed hybrid RCES method spawned good results than the existing ones. This is graphically represented in the following Figure 5.

In the analysis of two distinct medical datasets, namely the Serum 25OHD & TSH and the hypothyroid dataset,

Algorithm: Pseudo code of proposed hybrid RCES algorithm

Input:

- Dataset (df)

Process:

1. Load the dataset.
2. Handle missing values: Replace '?' with NaN and impute missing values.
3. Encode categorical features into numerical format.
4. Split the data into training and testing sets.
5. Perform Chi-Squared test for feature selection:
   - Select the top k features based on their chi-squared statistics.
6. Train RandomForestClassifier to obtain feature importances:
   - Fit the RandomForestClassifier model on the training data.
   - Obtain feature importances.
7. Select top k features based on importance from the RandomForestClassifier.
8. Identify common features selected by both Chi-Squared test and RandomForestClassifier.
9. Train RandomForestClassifier on the selected features:
   - Fit the RandomForestClassifier model on the training data using the common features.
10. Use the trained RandomForestClassifier to make predictions on the test set.
11. Assess the model performance:
    - determine accuracy, precision, recall, and F1-score.
    - Produce a confusion matrix.

Figure 3: Proposed hybrid RCES algorithm

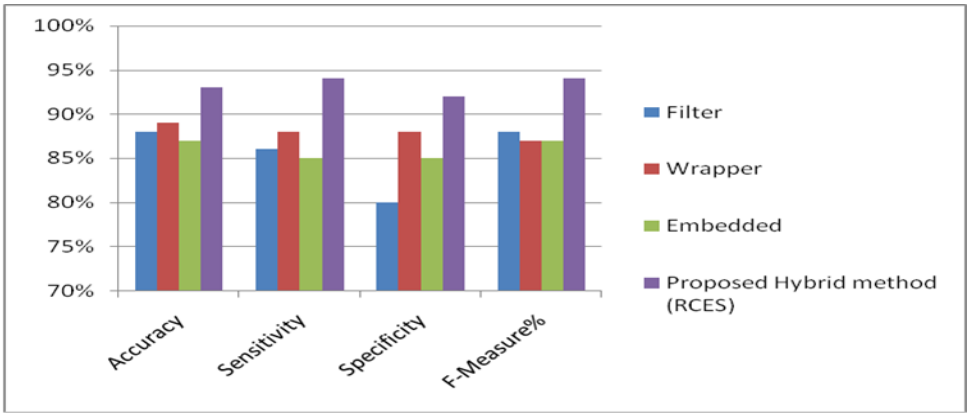

Figure 4: Graphical representation of the performance metrics (in %) obtained by proposed and existing feature selection methods for Serum 25OHD &TSH dataset

Table 3: Comparison of performance metrics of various approaches for the serum 25OHD & TSH dataset

Performance metrics/ methods	*Methods*	*Accuracy*	*Sensitivity*	*Specificity*	*F-Measure%*
Filter	MSE using a linear regression model	0.87	0.81	0.80	0.86
	LASSO	0.51	1.0	0.87	0.67
	Mutual information	0.88	0.86	0.80	0.88
Wrapper	Backward elimination	0.88	0.88	0.88	0.87
	Recursive feature elimination (RFE)	0.89	0.88	0.88	0.87
	Forward selection	0.86	0.87	0.86	0.88
Embedded	LASSO regression (L1 Regularization)	0.86	0.81	0.80	0.85
	Ridge regression (L2 Regularization)	0.87	0.81	0.83	0.86
	Tree-based methods (random forest, gradient Boosting)	0.87	0.85	0.85	0.87
Hybrid	CorrRecursive feature selection (CRFS)	0.90	0.92	0.93	0.92
	RanChi ensemble selection (RCES)	0.93	0.94	0.92	0.94

Table 4: Comparison of performance metrics of various approaches for the Hypothyroid dataset

Performance metrics/ methods	*Methods*	*Accuracy*	*Sensitivity*	*Specificity*	*F-Measure%*
Filter	MSE using linear regression model	0.83	0.78	0.79	0.80
	LASSO	0.49	0.12	0.01	0.12
	Mutual information	0.86	0.84	0.78	0.85
Wrapper	Backward elimination	0.86	0.84	0.86	0.84
	Recursive feature elimination (RFE)	0.87	0.86	0.87	0.84
	Forward Selection	0.84	0.80	0.83	0.84
Embedded	LASSO regression (L1 Regularization)	0.82	0.79	0.76	0.83
	Ridge regression (L2 Regularization)	0.83	0.79	0.80	0.83
	Tree-based methods (Random Forest, Gradient Boosting)	0.83	0.84	0.83	0.84
Hybrid	CorrRecursive feature selection»(CRFS)	0.86	0.84	0.85	0.87
	RanChi ensemble selection» (RCES)	0.88	0.89	0.89	0.86

Table 5: Best outcomes of the serum 25OHD & TSH dataset

Methods	*Accuracy*	*Sensitivity*	*Specificity*	*F-measure%*
Filter	0.88	0.86	0.80	0.88
Wrapper	0.89	0.88	0.88	0.87
Embedded	0.87	0.85	0.85	0.87
Proposed hybrid method (RCES)	0.93	0.94	0.92	0.94

Table 6: Best outcomes of the hypothyroid dataset

Methods	*Accuracy*	*Sensitivity*	*Specificity*	*F-Measure%*
Filter	0.86	0.84	0.78	0.85
Wrapper	0.87	0.86	0.87	0.84
Embedded	0.83	0.84	0.83	0.84
Proposed Hybrid method (RCES)	0.88	0.89	0.89	0.86

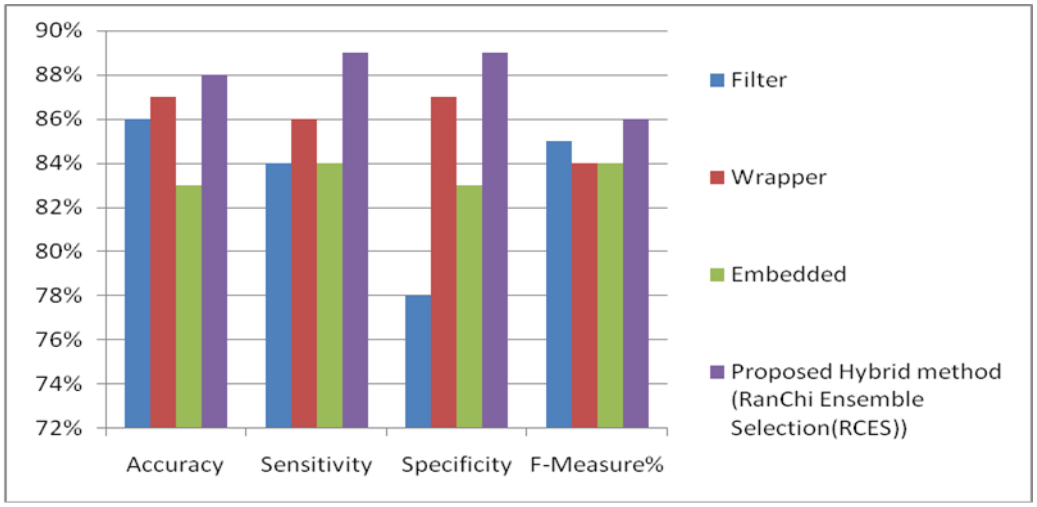

Figure 5: Graphical representation of the performance metrics (in %) obtained by proposed and existing feature selection methods for hypothyroid dataset

several feature selection methods were employed, including filters, wrappers, embedded approaches, and proposed hybrid techniques. For the serum 25OHD & TSH dataset, the top-performing methods under each category were identified and evaluated based on key performance metrics. In the filter category, methods such as MSE using linear regression model and Mutual Information demonstrated notable accuracies, sensitivities, specificities, and F-measures, exhibiting their effectiveness in feature selection while LASSO gave poor results for this study. Similarly, wrapper methods like backward elimination and recursive feature elimination exhibited high performance, as did embedded methods such as LASSO Regression and Tree-based Methods.

The proposed hybrid approaches like CorrRecursive feature selection (CRFS) and RanChi ensemble selection (RCES) delivered very good results than the existing methods by yielding more than 90% for all the performance metrics' thereby standing out first for the Serum 25OHD & TSH dataset**.**

With similar metrics in consideration, various feature selection strategies are applied to the hypothyroid data*set al*so. Despite differences in dataset characteristics, analogous patterns emerged in the top-performing methods across filter, wrapper, embedded, and hybrid categories. Methods such as MSE using linear regression model and mutual information stood out in the filter category, while backward elimination and recursive feature elimination produced good results among wrapper methods. Embedded techniques like LASSO regression and ridge regression generated competitive performance, along with tree-based methods.

The proposed hybrid methods like CorrRecursive feature selection (CRFS) and RanChi ensemble selection (RCES) also demonstrated their efficacy in optimizing feature selection for the hypothyroid dataset and produced more than 88% for all the performance metrics.

Evaluation Metrics

Confusion Matrix

The research evaluates the prediction performance of the algorithm using four performance measures derived from the confusion matrix. The following Figures 6 and 7 illustrate the confusion matrix structure for binary classification involves combining distinct predicted and true values into four cases: True positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) (Jian Yang *et al*., 2022).

- *True Positive (TP)*

Correctly identifies the presence of thyroid disease in a patient.

- *True Negative (TN)*

Correctly identifies the absence of thyroid disease in a patient.

- *False Negative (FN)*

Incorrectly identifies that a patient does not have thyroid disease.

- *False Positive (FP)*

Incorrectly identifies that a patient has thyroid disease.

$$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)}$$

$$Precision = \frac{(TP)}{(TP+FP)}$$

$$Sensitivity = \frac{(TP)}{(TP+FN)}$$

$$Specificity = \frac{(TN)}{(TP+FP)}$$

$$F1-Score = \frac{(2*Precision*Sensitivity)}{(Precision+Sensitivity)}$$

As seen in Figure 6, the confusion matrix offers an ample representation of the model's classification performance. The model proved its ability by correctly classifying 175 examples as negative (True Positives). But in 14 cases, the genuine status was incorrectly reported as positive (False Positives). On the other hand, the model properly classified 231 negative examples (True Negatives) but misclassified 13 negative ones (False Negatives). The given confusion matrix shows that the model performs well overall, as seen by the provided confusion matrix, suggesting that it can distinguish between the two classes with an adequate level of accuracy.

The model performed remarkably well in classifying cases into hypothyroid and non-hypothyroid groups, as shown by the confusion matrix in Figure 7. Specifically, the algorithm was able to accurately forecast 683 cases as hypothyroid (True Positives) indicating that people

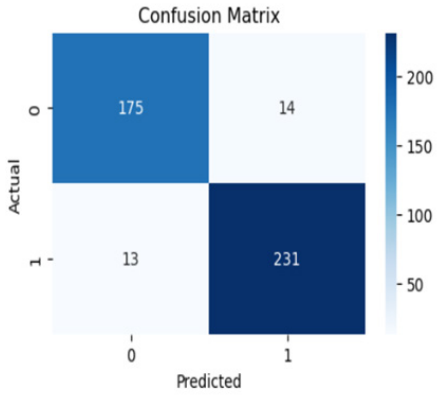

Figure 6: Confusion matrix for serum 25OHD and TSH dataset

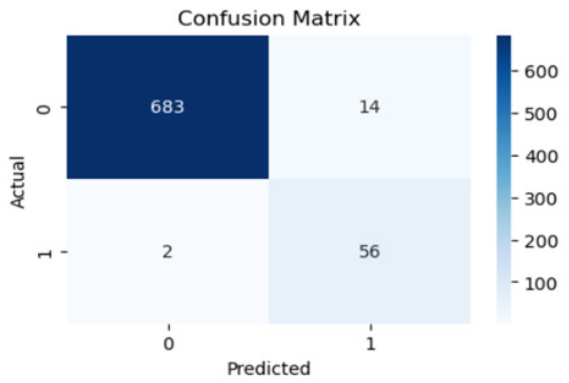

Figure 7: Confusion matrix for hypothyroid dataset

always have hypothyroidism. But 14 cases were incorrectly identified as hypothyroid when they weren't (False Positives). In contrast, the model properly recognized 56 non-hypothyroid cases (True Negatives) but wrongly classified 2 cases as non-hypothyroid (False Negatives). Notwithstanding these small differences, the model performs admirably overall, demonstrating its high degree of accuracy in distinguishing between hypothyroid and non-hypothyroid cases, as the confusion matrix illustrates.

Heatmap

Heatmap is a data visualization technique used to represent numerical data in table form, in which the entity values are symbolized in colors. Typically, heatmaps are presented in a grid format, with rows and columns representing variables or categories of data. Each cell in the grid is filled with a color intensity corresponding to the value of the data point it represents.

In this work, Pearson correlation coefficients are employed to estimate the correlations among feature elements and a heat map is utilized to visualize the correlation level among them. In Figures 8 and 9, each row and column represents the correlation coefficient between the related features. This suggests that the chosen features exert independent effects on the prediction variable (Rohit Bharti *et al*., 2021).

The following Figures 8 and 9 show the correlation heatmaps for the serum 25OHD & TSH dataset and the hypothyroid dataset.

Figures 8 and 9 show the heatmap representation for the serum 25OHD & TSH dataset and the hypothyroid dataset. This shows how the variables in the dataset are related to one another. Stronger positive correlations are shown by darker shades, especially those that veer closer to red, which implies that if one of the variables grows, the other also tends to grow. On the other hand, darker shades that lean towards blue, show stronger negative associations, meaning that when one variable rises, the other tends to fall. The diagonal line shows that every variable has a perfect association with every other variable (Niloy Biswas *et al*., 2023). While lighter hues that are almost white suggest lesser or negligible correlations, clusters of deeper shades may indicate groups of variables that are strongly associated with one another.

Experimental Evaluation

A comprehensive overview of the different techniques used on the serum 25OHD & TSH dataset can be found in Table 3. Feature evaluation is done independently of the model by filter approaches such as MSE utilizing linear regression and LASSO, where LASSO exhibits a trade-off between specificity

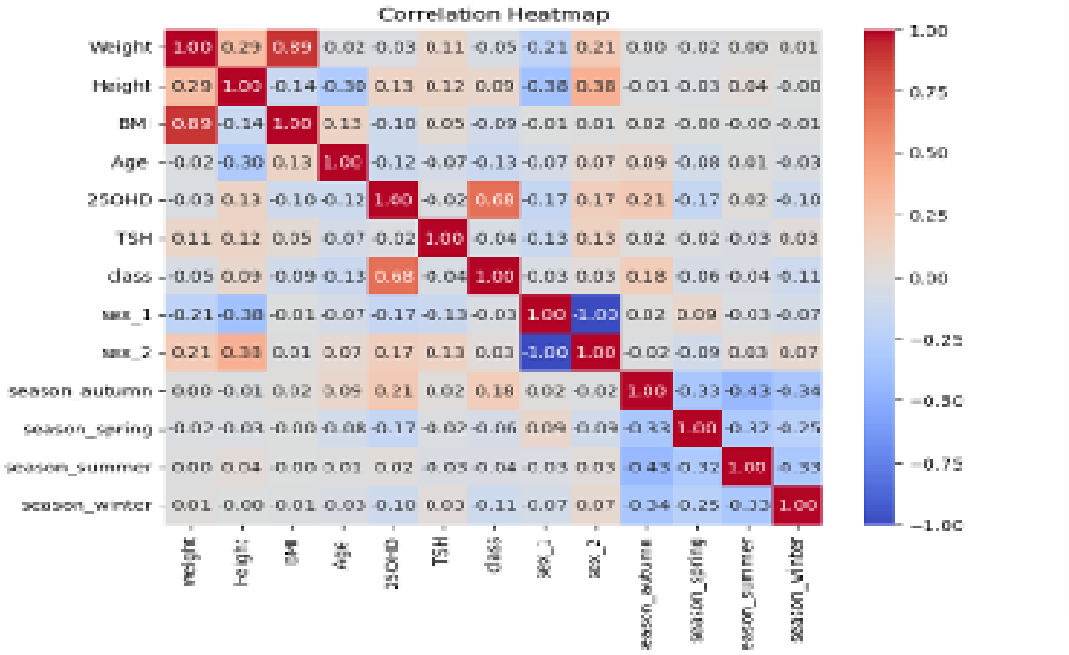

Figure 8: Correlation heatmap for serum 25OHD and TSH dataset

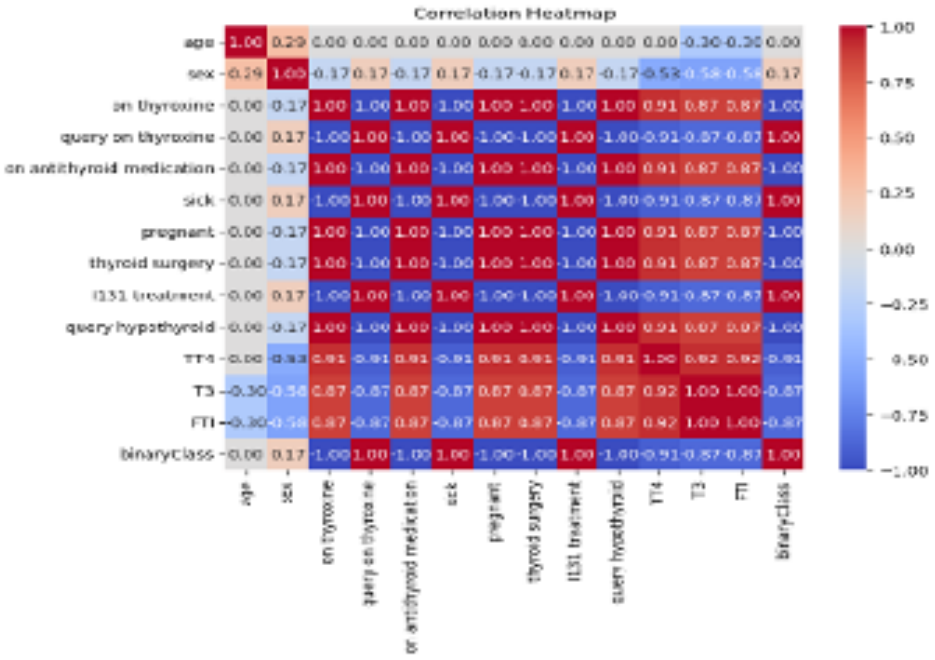

Figure 9: Correlation heatmap for hypothyroid dataset

and sensitivity. Robust feature selection performance was demonstrated using wrapper techniques such as recursive feature elimination (RFE) and backward elimination. Forward selection showed good sensitivity and specificity but was marginally less accurate. Results from embedded techniques, such as ridge regression and LASSO, were balanced, while tree-based approaches were competitive. Hybrid approaches, including CRFS and RCES, outperformed single-feature approaches due to the integration of various feature selection algorithms.

Similarly, Table 4 presents an analysis of techniques specifically designed for the hypothyroid dataset. Reasonable results were obtained with filter approaches such as LASSO and MSE utilizing linear regression, while mutual information provided comparable results. Wrapper techniques, especially RFE and backward elimination, demonstrated dependable feature selection performance. Among embedding methods, both LASSO and ridge regression produced balanced results. Hybrid approaches, exemplified by CRFS and RCES, exhibited superior performance, highlighting the advantages of integrating feature selection algorithms.

Tables 5 and 6 showcase the best outcomes for the Serum 25OHD & TSH and hypothyroid datasets, respectively. While filter, wrapper, and embedded approaches all showed balanced performance in the serum 25OHD & TSH dataset, the suggested hybrid technique (RCES) achieved the best accuracy and F-measure. For the hypothyroid dataset, the filter and wrapper methods demonstrated balanced sensitivity, specificity, and F-measure, with the suggested Hybrid strategy (RCES) outperforming other approaches, underscoring the value of combining different feature selection techniques.

Table 7 presents the results of an existing feature selection algorithm using several feature selection techniques on the serum 25OHD & TSH dataset. These techniques are divided into three categories as mentioned. The table shows the top k attributes that were chosen for each technique. Specifically, several combinations of features like patient, weight, and height were emphasized by filter methods such as mean squared error (MSE), LASSO, and mutual information. Season, 25OHD, and TSH were among the features that wrapper techniques like backward elimination and recursive feature elimination (RFE) found to be relevant. Important variables including height, sex, age, and 25OHD were revealed by embedded approaches like LASSO regression, ridge regression, and tree-based methods (Random Forest, Gradient Boosting). This thorough overview sheds light on the salient characteristics determined by various feature selection techniques. Facilitating further analysis or predictive modeling tasks for the serum 25OHD & TSH dataset.

Table 8 lists the features that were chosen using an established feature selection approach and applied to the hypothyroid dataset using various methods. The techniques, which each reveal the top k selected features, include filter, wrapper, and embedded methods. Significantly, filter techniques including mutual information, LASSO, and MSE revealed unique feature combinations like age, T4U, and hypo-pituitary as critical. Important features like T4U

Table 7: selected features for the existing feature selection algorithm for Serum 25OHD & TSH dataset

Existing feature selection methods		*Top k Selected features*
		Serum 25OHD & TSH dataset
Filter	MSE (Mean Squared Error) using linear regression model	Patient, Weight, Height
	LASSO	Sex, Age , season
	Mutual Information	Height, BMI, 25OHD
Wrapper	Backward Elimination	season, 25OHD,TSH
	Recursive Feature Elimination (RFE)	Age , 25OHD, TSH
	Forward Selection	Patient, Weight, Height
Embedded	LASSO Regression (L1 Regularization)	Height,Sex,25OHD
	Ridge Regression (L2 Regularization)	Patient, Weight, Height
	Tree-based Methods (Random Forest, Gradient Boosting)	Sex, Age ,TSH
Proposed hybrid Method1	CorrRecursive Feature Selection (CRFS)	Height, sex,25OHD
Proposed hybrid Method2	RanChi Ensemble Selection (RCES)	25OHD

Table 8: Selected features for the existing feature selection algorithm for hypothyroid dataset

Existing Feature Selection Methods		*Top k Selected features*
		Hypothyroid dataset
Filter	MSE (Mean Squared Error) using linear regression model	hypopituitary, T4U, FTI measured, T4U measured, TSH measured
	LASSO	sex, on thyroxine, query hypothyroid, TSH measured, TSH.
	Mutual Information	age, on thyroxine, thyroid surgery, TSH, T3
Wrapper	Backward Elimination	T4U measured, FTI measured, FTI, TBG measured, referral source
	Recursive Feature Elimination (RFE)	on thyroxine, TSH, T3, TT4, FTI
	Forward Selection	T3, age, sex, on thyroxine, TSH.
Embedded	LASSO Regression (L1 Regularization)	age, sex, on thyroxine, query on thyroxine,on antithyroid medication
	Ridge Regression (L2 Regularization)	sick, pregnant, thyroid surgery, I131 treatment, query hypothyroid.
	Tree-based Methods (Random Forest, Gradient Boosting)	age, sex, on thyroxine, query on thyroxine, on antithyroid medication
Proposed hybrid Method1	CorrRecursive Feature Selection (CRFS)	on antithyroid medication, thyroid surgery, query hypothyroid, tumor, TSH measured
Proposed hybrid Method2	RanChi Ensemble Selection (RCES)	FTI, on thyroxine, T3, TSH, TT4

measured, FTI, and thyroxin were found using wrapper techniques including backward elimination and RFE. Relevant features including age, sex, and thyroxin were found using embedded methodologies such as LASSO regression, ridge regression, and tree-based methods (Random Forest, Gradient Boosting). This thorough investigation sheds light on the key characteristics found through a variety of feature selection techniques, enabling additional investigation or predictive modeling for the hypothyroid dataset.

The features that have been chosen for a hybrid algorithm that has been applied to the serum 25OHD & TSH dataset as well as the Hypothyroid dataset.. Two methods are combined in the hybrid algorithm: RanChi ensemble selection (RCES) and CorrRecursive feature selection (CRFS). Height, sex, and serum 25OHD were determined by CRFS to be relevant features for the serum 25OHD & TSH dataset. On the other hand, the hypothyroid dataset was selected based on ant thyroid medication, thyroid surgery, tumor, and TSH measurement. However, for the serum 25OHD & TSH dataset, RCES identified 25OHD as the primary characteristic, but for the hypothyroid dataset, it chose FTI, on thyroxin, T3, TSH, and TT4. This table offers significant information for further analysis or research by illuminating the features chosen by the suggested hybrid algorithm for both datasets.

Thus, the study underscores the significance of using suitable feature selection techniques tailored to specific datasets and suggests that combining multiple feature selection algorithms into hybrid approaches presents a viable strategy for enhancing classification accuracy.

Conclusion

In conclusion, the analysis of feature selection methods applied to the serum 25OHD & TSH and the hypothyroid dataset revealed significant insights into optimizing model performance for medical data analysis. Across both datasets, a range of feature selection techniques, including filters, wrappers, embedded approaches, and proposed hybrid methods, were systematically evaluated and compared based on key performance metrics. The analysis of the serum 25OHD & TSH and hypothyroid datasets reveals that, in the case of the serum 25OHD & TSH dataset, the levels of vitamin D play a key role in identifying the underlying cause of thyroid disease. On the other hand, in the analyses performed on the hypothyroid dataset, where the 25OHD & TSH feature is missing, it was found that TSH serves as the primary driver. Thus, the experimental results suggest that vitamin D deficiency may play a significant role in thyroid disease, highlighting the potential importance of early diagnosis of vitamin D deficiency to prevent thyroid and related disorders in the future. The outcome of the experiments explores that by employing a better feature selection strategy, with a minimum number of features and with less time, accurate classification of the expected result could be achieved. This is because only the most essential features related to the selected problem got selected which reduces the algorithmic intricacy and improves the model's prediction accuracy. This work could be further improved by enhancing the accuracy of the prediction with an immense permutation with ML models.

Acknowledgment

The research scholar of this work, Ms.P.Vinnarasi, thanked the research supervisor, Dr. K. Menaka, and the management of Urumu Dhanalakshmi College for providing support and necessary resources for carrying out this research.

References

Ashank Bains, Taha Mur, Nathan Wallace and Jacob Pieter Noordz.(2021). The Role of Vitamin D as a Prognostic Marker in Papillary Thyroid Cancer, Cancers, doi: https://doi.org/10.3390/cancers13143516, Pages 1-9.

Bassam Abdo Al-Hameli, Abdul Rahman A. Alsewari, Abdulaziz Saleh Alraddadi, Arafat Aldhaqm (2021).Classification Algorithms and Feature Selection Techniques for a Hybrid Diabetes, Detection System, International Journal of Computational Intelligence in Control, ISSN: 0974-8571,Pages 81-91.

Danilovic DLS, Ferraz-de-Souza B, Fabri AW,Santana NO, Kulcsar MA, Cernea CR.(2016).25-Hydroxyvitamin D and TSH as Risk Factors or Prognostic Markers in Thyroid Carcinoma, PLOS ONE,,doi:10.1371/journal.pone.0164550, Pages 1-12.

Doppala, B.P., Bhattacharyya, D., Chakkravarthy, M., Kim, T.H.(2021). A hybrid machine learning approach to identify coronary diseases using feature selection mechanism on heart disease dataset, Springer, doi: https://doi.org/10.1007/s10619-021-07329-y, Pages 1-14.

Fadi Alharbi and Aleksandar Vakanski.(2023),Machine Learning Methods for Cancer Classification Using Gene Expression Data: A Review, Bioengineering ,doi: https://doi.org/10.3390/bioengineering10020173, Pages 1-26.

Faisal Saeed, Mohammad Al-Sarem, Muhannad Al-Mohaimeed, Abdelhamid Emara.(2022). Enhancing Parkinson's Disease Prediction Using Machine Learning And Feature Selection Methods", Computers, Materials & Continua, doi: 10.32604/Cmc.2022.023124,Pages.5640-5657.

Fathania Firwan Firdaus, Hanung Adi Nugroho, Indah Soesanti. (2020). A Review of Feature Selection And Classification Approaches For Heart Disease Prediction, Ijitee, doi: Https://Doi.Org/10.22146/Ijitee.59193,Pages.75-82.

Filip Lebiedzi ´ Nski And Katarzyna Aleksandra Lisowska.(2023). Impact of Vitamin D On Immunopathology Of Hashimoto's, Thyroiditis: From Theory To Practice, Nutrients, doi: https://Doi.Org/10.3390/Nu15143174, Pages 1-18.

Hyperlink for accessing the support information https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0164550#sec014. Downloadable data set: https://doi.org/10.1371/journal.pone .0164550.s001

Imad R. Musa, Gasim I. Gasim, Sajjad Khan, Ibrahim A. Ibrahim, Hamdi Abo-Alazm, And Ishag Adam.(2017). No Association Between 25 (Oh) Vitamin D Level And Hypothyroidism Among Females, Open Access Maced J Med Sci , doi: 10.3889/Oamjms.2017.029,Pages 126–130.

Jain, Achina Jain, Vanitab.(2022). Sentiment classification using hybrid feature selection and ensemble classifier, Journal of Intelligent & Fuzzy Systems, doi: 10.3233/JIFS-189738, Pages. 659-668.

Jian Yang and Jinhan Guan, A Heart Disease Prediction Model Based on Feature Optimization and Smote-Xgboost Algorithm, MDPI, Information, doi: https://doi.org/10.3390/info13100475 ,Pages 1-15.

Jie Kuang Zhijian Jin , Lingxie Chen , Qiwu Zhao ,Haiyan Huang , Zhuoran Liu , Weiping Yang , Haoran Feng , Zheyu Yang , Juan J Díez , Marc Pusztaszeri, Jung Min Kim , Elena Bonati , Xi Cheng , Jiqi Yan ·Weihua Qiu(2022). Serum 25-Hydroxyvitamin D Level Is Unreliable As A Risk Factor And Prognostic Marker In Papillary Thyroid Cancer, Ann Transl Med, doi: 10.21037/Atm-22-10, Pages 1-12.

John Smith, Sarah Johnson, Michael Davis.(2021). Comparative Study of Feature Selection Methods for Medical Data Classification Using Machine Learning Techniques, IEEE Transactions on Biomedical Engineering, doi: 10.1109/TBME.2021, Pages 150-165.

Lutfiye Secil, Deniz Blayen.(2019). The Relationship Between Serum Vitamin D Levels And Thyroid Function Tests In Euthyroid And Hypothyroid Patients With Elevated Anti-Tpo, doi:10.5505/Kjms.2019.54037,Pages 158-161.

Maryam Khalili, Ali Reza Manzoori.(2022). Enhanced Feature Selection Techniques for Heart Disease Diagnosis Using Machine Learning Ensemble Model, Expert Systems with Applications, doi: 10.1016/j.eswa. Pages 150-165.

Mirjana Babic Leko, Iva Jureško, Iva Rozić, Nikolina Pleić, Ivana Gunjača, And Tatijana Zemunik.(2023) Vitamin D And The Thyroid: A Critical Review Of The Current Evidence, Int J Mol Sci, doi: 10.3390/Ijms24043586,Pages 1-17.

Niloy Biswas, ,Md Mamun Ali,Md Abdur Rahaman,Minhajul Islam. (2023). Machine Learning-Based Model to Predict Heart

Disease in Early Stage Employing Different Feature Selection Techniques, Hindawi BioMed Research International , doi: https://doi.org/10.1155/2023/6864343, Pages 1-15.

Nino Turashvili , Lali Javashvili , Elene Giorgadze .(2021). Vitamin D Deficiency Is More Common In Women with Autoimmune Thyroiditis: A Retrospective Study, Int J Endocrino, doi: 10.1155/2021/4465563,Pages 1-6.

Quan Li, Rui Bi, Relationship Between Serum 25-Hydroxyvitamin D Deficiency And Thyroid Disease In Postmenopausal Women With Type 2 Diabetes Mellitus.(2023). Diabetes, Metabolic Syndrome And Obesity Dove press, doi: https://Doi.Org/10.2147/Dmso.S404172 ,Pages 1407-1414.

Raid Alzubi, Naeem Ramzan, Hadeel Alzoubi,Abbes Amira.(2017). A Hybrid Feature Selection Method for Complex Diseases SNPs, IEEE Access , doi: 10.1109/ACCESS.2017.2778268, Pages 1292-1301.

Rohit Bharti, Aditya Khamparia, Mohammad Shabaz, Gaurav Dhiman, Sagar Pande, Parneet Singh.(2021),Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning, Hindawi Computational Intelligence and Neuroscience, doi: https://doi.org/10.1155/2021/8387680, Pages 1-11.

S.Usha , Dr.S.Kanchana.(2021). Predicting Heart Disease Using Feature Selection Techniques Based On Data Driven Approach, Webology, Issn: 1735-188x, Pages 97-108.

Saba Bashir , Irfan Ullah Khattak, Aihab Khan, Farhan Hassan Khan, Abdullah Gani , Muhammad Shiraz.(2022). A Novel Feature Selection Method for Classification of Medical Data Using Filters, Wrappers, and Embedded Approaches, Hindawi, doi: 10.1155/2022/8190814, Pages 1-12.

Sandeep Appunni, Muni Rubens, Venkataraghavan Ramamoorthy, Anshul Saxena, Raees Tonse, Emir Veledar & Peter Mcgranaghan.(2021). Association Between Vitamin D Deficiency And Hypothyroidism: Results From The National Health And Nutrition Examination Survey (Nhanes) Bmc Endocrine Disorders , doi: Https://Doi.Org/10.1186/S12902-021-00897-1, Pages 2-9.

Souhir Ben Amor, Amira Barhoumi, Mouna Baklouti, Raja Ghozi, and Adel M. Alimi.(2019). Feature Selection Methods in Machine Learning: A Review and a Comparative Analysis, International Conference on Machine Learning and Computing (ICMLC),doi: 10.1145/3316782.3316831.

Tereza Planck, Bushra Shahida, Johan Malm, And Jonas Manjer. (2018). Vitamin D In Graves Disease: Levels, Correlation With Laboratory And Clinical Parameters, And Genetics, Eur Thyroid doi: 10.1159/000484521.Pages 27–33.

Thomas Davenport A and Ravi Kalakota B.(2019),The potential for artificial intelligence in healthcare, Future Healthcare Journal, doi:10.7861/futurehosp.6-2-94, Pages 94–98.

Thyroid Data - https://archive.ics.uci.edu/ml/datasets/thyroid+disease

Y. Saeys, I. Inza, and P. Larrañaga.(2007). A review of feature selection techniques in bioinformatics, Bioinformatics, doi: https://doi.org/10.1093/bioinformatics/btm344, Pages 2507–2517.

Yap Bee Wah1, Nurain Ibrahim, Hamzah Abdul Hamid, Shuzlina Abdul-Rahman and Simon Fong.(2018). Feature selection methods: Case of filter and wrapper approaches for maximizing classification accuracy, Pertanika, SCIENCE & TECHNOLOGY, doi: https://www.researchgate.net/publication/322920304, Pages.329 - 340.

Yixin Liu, Wee Sun Lee.(2019). A survey on data augmentation for imbalanced classification, ACM Computing Surveys (CSUR), doi:10.1145/3344685,Pages 1-34.

Yong Guo, Chun-Yan Wu, Yu-Hong Deng, And Jie-Ling Wu.(2020). Associations between Serum 25-Hydroxyvitamin D Levels And Thyroid Function Parameters In Previously Healthy Children Aged 6 To 24 Months, Risk Manag Healthc Policy. doi: 10.2147/Rmhp.S269640, Pages 1647–1653.