Journal of Namibian Studies, 35 S1 (2023): 5012-5027 ISSN: 2197-5523 (online)

Integration Of Feature Reduction Method With Feature Selection In Disease Prediction

Hendra Nusa Putra[1] , Sarjon Defit[2] ,
Gunadi Widi Nurcahyo[3]

[1]Medical Record. Email: nusahendra@gmail.com

[2]Faculty of Computer Science, UPI YPTK Padang, Indonesia.
Email: sarjondefit@upiyptk.ac.id

[3]Faculty of Computer Science, UPI YPTK Padang, Indonesia.
Email: gunadiwidi@yahoo.co.id

Abstract

This research focuses on feature selection and evaluating the accuracy of machine learning models. The objective is to identify the most relevant features and optimize the model's performance while overcoming overfitting. The study begins by meticulously selecting 47 features out of an initial set of 132 features, which demonstrate improved accuracy. Several evaluation steps are conducted, including assessing feature coefficients, utilizing five model estimators, testing with ensemble methods, and evaluating results using the confusion matrix. The findings indicate that the applied technique successfully selects the most appropriate features and effectively mitigates overfitting. By considering feature coefficients, employing multiple model estimators, and leveraging ensemble techniques, the selected features significantly contribute to accurate predictions of the desired target. The evaluation results demonstrate high accuracy and the ability of the model to distinguish between positive and negative classes. Overall, the research showcases the effectiveness of the feature selection process and accuracy evaluation in optimizing machine learning models. The steps taken successfully address overfitting concerns and yield satisfactory outcomes. The study provides valuable insights into feature selection and model optimization for future research.

Keywords: Feature Selection, Attribute Reduction, Disease, Prognosis, Symptom

Introduction

In the current era of information, the healthcare industry is increasingly adopting patterns in processing, interpreting, and utilizing medical big

data. Big data refers to a large or complex collection of data that is difficult to capture, process, manage, and interpret within a reasonable amount of time using current technology[1]. Big data technology has limitations in efficient storage, processing, querying, and analyzing medical data. Technologies such as Deep Learning and Machine Learning simulate human thinking, assisting doctors in diagnosing and treating patients by providing personalized healthcare services and promoting intelligent application processes in healthcare. Data mining techniques are increasingly being used to generate more comprehensive datasets from high volumes of clinical care data that are less controlled, containing fewer quality controls for data error reduction (e.g., duplicate entries), incomplete data, and unstructured doctor's notes [2].

Data mining plays a crucial role in the medical industry for disease prediction. Software systems can be developed to make accurate medical decisions and ensure cost-effective optimal care. Medical data mining has robust capabilities to unveil hidden patterns in medical domain records. The real challenge of medical data mining lies in classifying and predicting medical datasets. The number of studies on medical big data is rapidly and steadily increasing, indicating the immense potential in this field. Keywords like big data, data mining, health, cloud computing, machine learning, and electronic health record system are currently the central issues in medical big data research. Conversely, research involving keywords such as the Internet of Things, e-health, sensors, predictive modeling, quantified self, smart city, wearable devices, and m-health is still in its early stages but may emerge as new mainstream trends in the future and warrant further investigation [1].

Feature selection plays a crucial role in disease prediction models as it helps identify the most relevant and informative variables related to the target disease. [3] By selecting the most significant features, we can reduce the complexity of the prediction model and improve its accuracy. This process involves identifying and removing irrelevant or redundant variables from the dataset, which can have undesirable effects on model performance. [4] Moreover, feature selection helps in addressing the challenges of overfitting and underfitting. Choosing the right subset of features can help reduce overfitting, improve generalization, and enhance the efficiency of the predictive model by reducing training duration and computational complexity. [3] Furthermore, feature selection can also aid in interpreting the results of predictive models by highlighting which variables are most influential in predicting the disease outcome. Furthermore, feature selection can assist in minimizing the "curse of dimensionality" by reducing the number of features and focusing on the most relevant ones. By selecting a subset of features with strong responses to the target variable, the predictive power of the model can be enhanced, and the overall performance accuracy can be significantly improved. Additionally, feature selection methods can help identify the variables that have a strong impact on disease prediction.

The feature reduction method is a process used in data analysis and machine learning to select the most relevant features or variables from a larger set. This method can be divided into two categories: feature extraction and feature selection. Feature extraction is a dimensionality reduction method that aims to transform the input data into a set of features that have been reduced while retaining the most relevant information from the original data. [5] This is achieved by applying mathematical techniques to identify patterns, relationships, or underlying structures in the data. Feature extraction is a crucial step in the feature reduction process as it helps uncover latent patterns and transform high-dimensional data into a lower-dimensional feature space. These reduced features can then be used for various purposes, such as classification, visualization, and compression.

Research Question

This research project seeks to investigate and address the following questions:

1. How can the feature reduction approach be linked with the feature selection method in the context of disease prediction?
2. Does integrating the feature reduction approach with the feature selection method increase disease prediction performance when compared to utilizing the feature selection method alone?
3. What impacts do different modifications of the feature reduction method have on illness prediction performance?

Literature Survey

Research conducted by Ekta Maini et al. [6] is based on the concept of feature selection in machine learning to improve model performance. Feature selection is done to reduce data dimensions and eliminate irrelevant or redundant features. In this study, four feature selection algorithms—Relief, RelieFF, Chi-square test, and MRMR—were applied to the heart disease prediction system. Then, system performance is evaluated using five different classification algorithms. The results showed a significant increase in the accuracy, sensitivity, specificity, and processing speed of the heart disease prediction system. The findings of this study also identify the best feature selection algorithm for each of the classification algorithms used. Overall, this research highlights the importance of feature selection in building an effective machine learning model. This discovery provides valuable insights into the appropriate feature selection algorithm for a particular classification algorithm. By optimizing feature selection, a model that is more accurate and efficient in predicting heart disease can be produced.

The theoretical framework of this study involves the use of feature selection and a hybrid Congruence coefficient Kumar-Hassebrook similarity method to select the best features for heart disease detection.

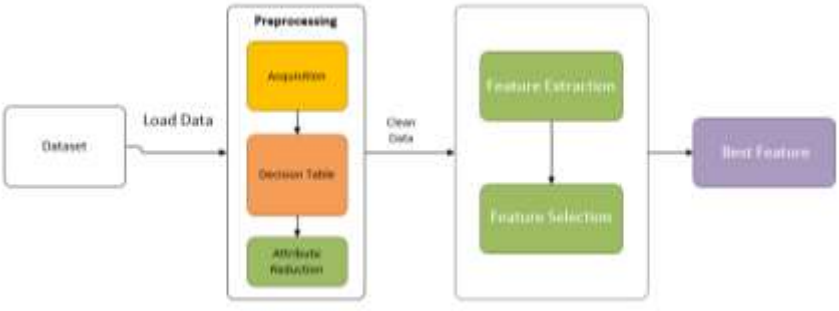
The study also uses the SqueezeNet deep learning model, which is tuned by the dwarf mongoose optimization algorithm (DMOA), to predict heart disease. The study aims to improve the efficacy of heart disease detection by addressing the issue of deficient test data.The findings of the study show that the proposed DMOA-SqueezeNet method achieved a maximum accuracy of 0.925, a sensitivity of 0.926, and a specificity of 0.918. The study demonstrates that the use of feature selection and hybrid similarity methods can improve the accuracy of heart disease detection. The study also shows that the DMOA-SqueezeNet method is an effective approach for heart disease prediction.[7]

The theoretical framework of the research conducted by Ndung'u et al.[3] is to determine the commonly used attributes that influence user preferences in online recommendation systems. This study uses the feature selection method to mine those preferences by selecting the high-impact attributes of the items. In machine learning, feature selection is used as a data pre-processing method, but its use is extended in this work to achieve two goals: redundant and uninformative feature removal and the selection of relevant formative features based on response variables. The final objective is suggested to be to identify and define the frequent and shared features that most online marketplace users will like when they express their preferences. This study explores all clustering of feature selection, filter, wrapper, embedding, and intrinsic algorithms to show how each can generate relevant features to be used to model recommender systems. The results show that different feature selection methods produce different feature scores, and the intrinsic method has the best overall results with a model accuracy of 85%. The features selected are frequently preferred attributes that influence user preference.

Proposed System

These features are being integrated for disease prediction. Using the following framework model, a first design seeks the best features from the extraction and classification processes:

Figure 1. Framework Model



As shown in the graph, the initial process begins with the dataset, followed by preprocessing, which includes the acquisition, decision table, and attribute reduction processes. The next stage is the feature extraction and feature selection steps, from which you will obtain the best features that can be used.

a. Dataset: The data used in machine learning takes the form of symptom and disease data.

b. Acquisition is one of the stages in the data preprocessing process that aims to ensure that the data obtained from its source is correct and relevant before further processing. This technique is performed at the initial stage before the data is processed and analyzed.

c. Preprocessing: After going through the preprocessing stage, data cleaning will be carried out, which involves filtering the dataset, filling in missing values, smoothing noisy data, identifying or removing outliers, and splitting inconsistent data

d. A Decision Table is a technique used to identify business rules related to a dataset. This technique is commonly used in data analysis to obtain useful information for decision-making.

e. Feature Extraction: This process involves transforming complex data representations into simpler and more informative representations. The primary objective of feature extraction is to identify and extract the most relevant or important features from the original data.

Dataset

The researcher utilized data from a reliable and trustworthy source through www.kaggle.com, which has also been verified by Dr. M. Djamil Padang Hospital. This medical data is stored in a suitable format for analysis, specifically in CSV format. The dataset primarily focuses on disease data and symptoms, which are used to predict disease diagnosis. The dataset consists of 133 features/parameters and covers 41 types of diseases/prognoses.

Table 1. Types of Disease

No	Disease	No	Disease	No	Disease
P1	Acne	P15	Fungal infection	P30	Migraine
P2	Allergy	P16	GERD	P31	Malaria
P3	Arthritis	P17	Gastroenteritis	P29	Jaundice
P4	Alcoholic hepatitis	P18	Hepatitis A	P32	Osteoarthristis
P5	AIDS	P19	Hepatitis B	P33	Paralysis (brain hemorrhage)

P6	Bronchial Asthma	P20	Hepatitis C	P34	Pneumonia
P7	Chronic cholestasis	P21	Hepatitis D	P35	Psoriasis
P8	Common Cold	P22	Hepatitis E	P36	Peptic ulcer diseae
P9	Cervical spondylosis	P23	Heart attack	P37	Tuberculosis
P10	Chicken pox	P24	Hypothyroidism	P38	Typhoid
P11	Drug Reaction	P25	Hyperthyroidism	P39	Urinary tract infection
P12	Dimorphic hemorrhoids(piles)	P26	Hypoglycemia	P40	Varicose veins
P13	Diabetes	P27	Hypertension	P41	(vertigo) Paroymsal Positional Vertigo
P14	Dengue	P28	Impetigo		

Source: https://github.com/itachi9604/healthcare-chatbot/blob/master/MasterData/symptom_Description.csv

The symptoms that serve as the causes of the diseases, also known as parameters/features, are as follows:

Table 2. Disease symptoms

No	Fitur	No	Fitur	No	Fitur
G1	itching	G46	fluid_overload (1)	G91	foul_smell_of urine
G2	skin_rash	G47	swelling_of_stomach	G92	continuous_feel_of_urine
G3	nodal_skin_eruptions	G48	swelled_lymph_nodes	G93	passage_of_gases
G4	continuous_sneezing	G49	malaise	G94	internal_itching
G5	shivering	G50	blurred_and_distorted_vision	G95	toxic_look_(typhos)
G6	chills	G51	phlegm	G96	depression
G7	joint_pain	G52	throat_irritation	G97	irritability
G8	stomach_pain	G53	redness_of_eyes	G98	muscle_pain
G9	acidity	G54	sinus_pressure	G99	altered_sensorium
G10	ulcers_on_tongue	G55	runny_nose	G100	red_spots_over_body

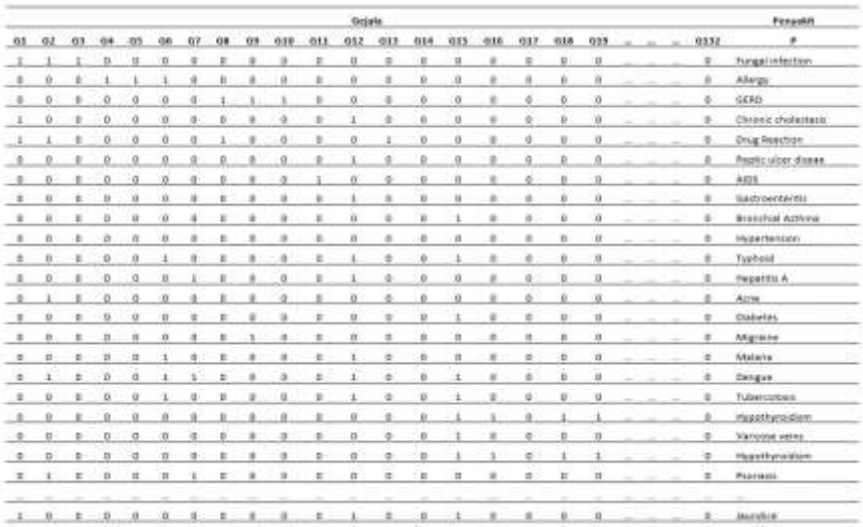
G11	muscle_wasting	G56	congestion	G101	belly_pain
G12	vomiting	G57	chest_pain	G102	abnormal_menstruation
G13	burning_micturition	G58	weakness_in_limbs	G103	dischromic _patches
G14	spotting_ urination	G59	fast_heart_rate	G104	watering_from_eyes
G15	fatigue	G60	pain_during_bowel_movements	G105	increased_appetite
G16	weight_gain	G61	pain_in_anal_region	G106	polyuria
G17	anxiety	G62	bloody_stool	G107	family_history
G18	cold_hands_and_feets	G63	irritation_in_anus	G108	mucoid_sputum
G19	mood_swings	G64	neck_pain	G109	rusty_sputum
G20	weight_loss	G65	dizziness	G110	lack_of_concentration
G21	restlessness	G66	cramps	G111	visual_disturbances
G22	lethargy	G67	bruising	G112	receiving_blood_transfusion
G23	patches_in_throat	G68	obesity	G113	receiving_unsterile_injections
G24	irregular_sugar_level	G69	swollen_legs	G114	coma
G25	cough	G70	swollen_blood_vessels	G115	stomach_bleeding
G26	high_fever	G71	puffy_face_and_eyes	G116	distention_of_abdomen
G27	sunken_eyes	G72	enlarged_thyroid	G117	history_of_alcohol_consumption
G28	breathlessness	G73	brittle_nails	G118	fluid_overload.1
G29	sweating	G74	swollen_extremeties	G119	blood_in_sputum
G30	dehydration	G75	excessive_hunger	G120	prominent_veins_on_calf
G31	indigestion	G76	extra_marital_contacts	G121	palpitations
G32	headache	G77	drying_and_tingling_lips	G122	painful_walking
G33	yellowish_skin	G78	slurred_speech	G123	pus_filled_pimples
G34	dark_urine	G79	knee_pain	G124	blackheads

G35	nausea	G80	hip_joint_pain	G125	scurring
G36	loss_of_appetite	G81	muscle_weakness	G126	skin_peeling
G37	pain_behind_the_eyes	G82	stiff_neck	G127	silver_like_dusting
G38	back_pain	G83	swelling_joints	G128	small_dents_in_nails
G39	constipation	G84	movement_stiffness	G129	inflammatory_nails
G40	abdominal_pain	G85	spinning_movements	G130	blister
G41	diarrhoea	G86	loss_of_balance	G131	red_sore_around_nose
G42	mild_fever	G87	unsteadiness	G132	yellow_crust_ooze
G43	yellow_urine	G88	weakness_of_one_body_side		
G44	yellowing_of_eyes	G89	loss_of_smell		
G45	acute_liver_failure	G90	bladder_discomfort		

Decision Table

The Decision Table is a technique used to identify business rules related to a dataset. This technique is commonly employed in data analysis to acquire useful information for decision-making purposes. Based on the disease data in Table.1 and symptom data in Table.2, the decision table is defined as follows:

Table 3. Symptoms and Disease



Result Analysis

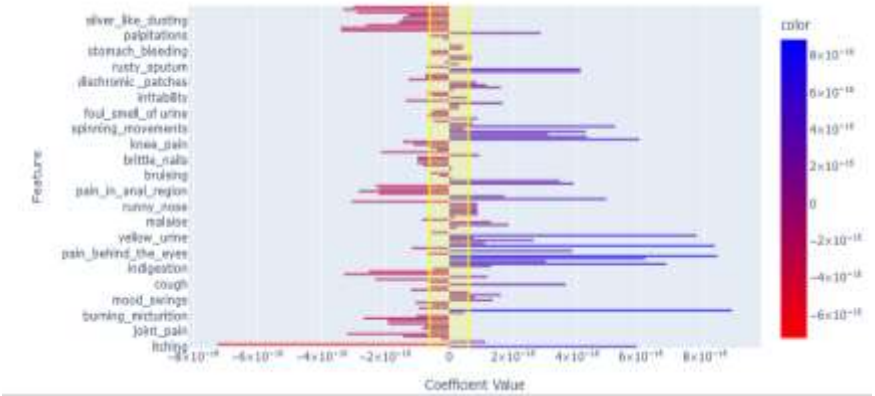
Initially, we will attempt to examine the datasets that will be used with Logistic Regression. This is due to the possibility of some irrelevant data if we employ all of the features. In this section, we will look at the coefficients and attempt to determine their threshold. The following pseudocode depicts the procedure flow:

Figure 2. Pseudecode.

Algorithm 1. Logistic Regression Check Analysis

1. Start
2. Input the required libraries.
3. Input the importance_threshold value.
4. Create a bar plot using Plotly Express:
a. Set X-axis as Coefficients (X)
b. Set Y-axis as Feature names (X_train.columns)
c. Set orientation as Horizontal
d. Set bar color based on coefficient values
e. Set color scale using red and blue
f. Label the x-axis as "Coefficient Value" and y-axis as "Feature"
g. Set title as "Important Features Based on Model Weights"
5. Add a vertical line for the importance threshold:
a. Use add_vline() from Plotly Express
b. Set x-coordinate as importance_threshold
c. Set line color as yellow
6. Add a vertical rectangular box for the area of importance threshold:
a. Use add_vrect() from Plotly Express
b. Set x-coordinates as importance_threshold and -importance_threshold
c. Set line width as 0 to hide the boundary lines
d. Set fill color as yellow with opacity level of 0.2
7. Display the plot.
8. Stop

Figure 3. Important Features based on Model Weights.



The above graph illustrates that the left side, indicated by the color red, represents features with coefficients less than or towards negative zero. Conversely, features with coefficients greater than 1 are marked towards the right side or with the color blue.

Table 4. Threshold and Accuracy

No	% Threshold	Accuracy (%)
1	20	99,86
2	25	99,66
3	30	97,29

The data above explains that increasing the threshold value will reduce the accuracy level. However, on the other hand, if there are too many features during the evaluation process, overfitting may occur, resulting in longer computations and excessive resource usage. This is because not all features play a significant role in relation to the target data..

Attribute Reduction

This stage uses RFE (Recursive Feature Elimination) to pick features. RFE is a method for reducing the number of characteristics in a dataset by continually deleting attributes that are deemed less essential. It is generally used on large datasets with many features to reduce overfitting and improve model interpretability. RFE aids in the identification of influential attributes in predicting the target variable. Furthermore, RFE aids in the construction of a simplified model with a subset of significant features. As seen below, this stage employs five model estimators (Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and Gradient Boosting) with the same pseudocode procedure :

Figure 4. Pseudecode Attribute Reduction.

Algorithm 2. Recursive Feature Elimination
1. Start 2. Remove the 'target' column from the dataset dataframe and assign it to the variable X.

a. Input: dataset dataframe
b. Output: X (dataset without 'target' column)
3. Retrieve the 'target' column from the dataset dataframe and assign it to the variable y.
a. Input: dataset dataframe
b. Output: y (target column)
4. Create an object for the chosen estimator model and assign it to a variable, for example, lr.
a. Input: None
b. Output: lr (estimator model object)
5. Create the object using the estimator object, such as Logistic Regression, and specify the desired number of features (e.g., 47). Assign it to a variable, for example, rfe_lr.
a. Input: estimator object, number of features
b. Output: rfe_lr (RFE object with specified estimator and number of features)
6. Perform Recursive Feature Elimination (RFE) on the dataset X and y using the RFE object rfe_lr.
a. Input: X, y, rfe_lr
b. Output: None
7. Display the feature selection results by printing "Selected Features: %s" % rfe_lr.support.
a. Input: rfe_lr
b. Output: None
8. Display the feature rankings by printing "Feature Ranking: %s" % rfe_lr.ranking_.
a. Input: rfe_lr
b. Output: None
9. Stop

Figure 5. Extract Feature

```
# tampilkan hasil seleksi fitur
print("Selected Features: %s" % rfe_lr.support_)
print("Feature Ranking: %s" % rfe_lr.ranking_)
```

```
Selected Features: [ True  True False  True False  True  True  True False False False  True
  True False  True False False False False  True False  True False False
  True  True False  True  True False  True  True  True  True  True  True
 False False  True  True  True  True False  True False False False False
  True False False False False False False False  True False False False
  True False False  True  True False False  True False False False False
 False False  True False False False False False  True  True  True False
 False  True False False False False False  True False False False False
  True  True  True False False False  True False False False  True False
 False False False False False  True False False False False False False
 False False False  True False  True False False False  True False False]
Feature Ranking: [ 1  1 12  1 18  1  1  1 19 52 17  1  1 30  1 81 78 83 54  1 41  1 32 79
  1  1  8  1  1 26  1  1  1  1  1  1 68  6  1  1  1  1 73  1 49 86 28 25
  1  3 39 71 80 77 82 84  1 46  9 37  1 23 56  1  1 60 29  1  7 70 85 76
 69 64  1  2 75 63 55 33  1  1  1 44 45  1 16 22 65 13 50  1 21 38 66 14
  1  1  1 53 57 31  1 34 67 62  1 36 51 20 72 61 43  1 15  5 59 42 74 47
 48 10 27  1  4  1 24 40 58  1 11 35]
```

Figure 6. Selected Feature

```
selected_features_lr = X.columns[rfe_lr.support_]
print("Selected Features: %s" % selected_features_lr)
```

```
Selected Features: Index(['itching', 'skin_rash', 'continuous_sneezing', 'chills', 'joint_pain',
       'stomach_pain', 'vomiting', 'burning_micturition', 'fatigue',
       'weight_loss', 'lethargy', 'cough', 'high_fever', 'breathlessness',
       'sweating', 'indigestion', 'headache', 'yellowish_skin', 'dark_urine',
       'nausea', 'loss_of_appetite', 'constipation', 'abdominal_pain',
       'diarrhoea', 'mild_fever', 'yellowing_of_eyes', 'malaise', 'chest_pain',
       'pain_in_anal_region', 'neck_pain', 'dizziness', 'obesity',
       'excessive_hunger', 'muscle_weakness', 'stiff_neck', 'swelling_joints',
       'loss_of_balance', 'continuous_feel_of_urine', 'irritability',
       'muscle_pain', 'altered_sensorium', 'dischromic _patches',
       'family_history', 'coma', 'blackheads', 'skin_peeling', 'blister'],
      dtype='object')
```

In this process, the selection of the number of features is done randomly using the 5 models mentioned above, and the selected feature results are saved for further accuracy testing in the next step.

Feature Selection

Feature selection is a technique used to select a group of features that are most relevant and informative in a dataset. The goal is to improve the accuracy of the model's predictions and facilitate understanding of the data. Feature selection results in a subset of relevant features.

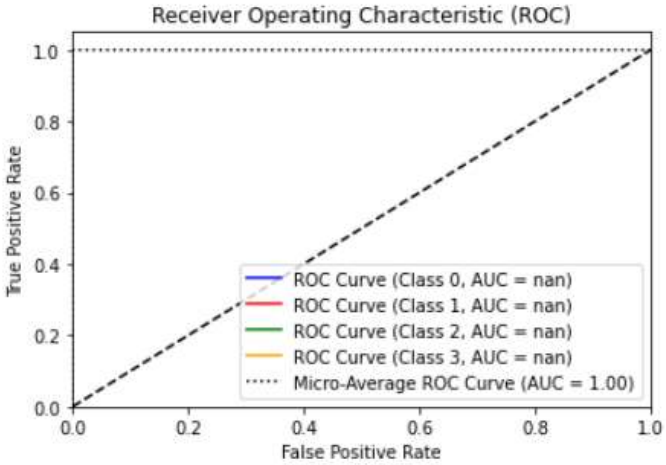
Output Result: Experimenting with different feature counts using Recursive Feature Elimination method with multiple estimators, achieving 100% accuracy.

Table 5. Accuracy Results based on Model and Features

No	Model / Estimator	Feature							
		46		47		48		49	
		Train	Test	Train	Test	Train	Test	Train	Test
1	Logistic Regression	0.9989	0.9989	100	100	100	100	100	100
2	Decision Tree	0,9899	0,9805	0,9905	0,9837	0,9910	0,9816	0.9867	0.9794
3	Random Forest	0.9945	0.9880	0,9932	0,9880	0,9967	0,9956	0,9929	0,9880
4	Support Vektor Machine	0,9986457	100	0,9986	100	0,9986	100	0,9972	100
5	Gradient Boosting	0.9794	0.9696	0,9788	0,9772	0.9840	0.9794	0.9823	0.9751

The Logistic Regression model with 47 features exhibited exceptional accuracy, as evidenced by its 100% Explained Variance Ratio. This indicates that the model captured all the variances present in the dataset, resulting in accurate predictions. Similarly, the SVM model achieved a perfect accuracy of 100% in both the Explained Variance Ratio and the ROC curve. [8]

Figure 7. ROC Curve



The 100% accuracy result implies that the model can tell the difference between both positive and negative categories in the test data. This shows that the model can classify data very effectively, with no prediction mistakes. In this situation, 100% accuracy means that the model predicts perfectly.

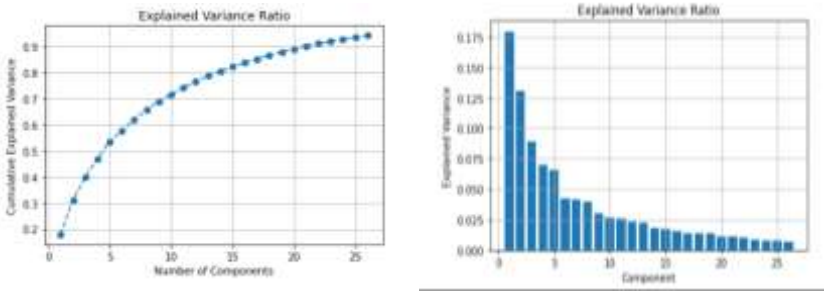


Figure 8. Plot Explained Variance and Histogram Explained Variance

The EVR plot demonstrates that each extra component contributes significantly to explaining the variation in the data. [9] whereas the X-axis Histogram increases consistently with the addition of components, it demonstrates that each additional component contributes significantly to explaining the variation in the data.

Figure 9. Classification Report

	precision	recall	f1-score	support
(vertigo) Paroymsal Positional Vertigo	1.00	1.00	1.00	23
AIDS	0.86	1.00	0.93	19
Acne	0.75	1.00	0.86	21
Alcoholic hepatitis	1.00	1.00	1.00	25
Allergy	1.00	1.00	1.00	29
Arthritis	1.00	1.00	1.00	29
Bronchial Asthma	1.00	1.00	1.00	25
Cervical spondylosis	1.00	1.00	1.00	24
Chicken pox	1.00	1.00	1.00	28
Chronic cholestasis	1.00	1.00	1.00	25
Common Cold	1.00	1.00	1.00	18
Dengue	1.00	1.00	1.00	21
Diabetes	1.00	1.00	1.00	21
Dimorphic hemmorhoids(piles)	1.00	1.00	1.00	23
Drug Reaction	1.00	1.00	1.00	29
Fungal infection	1.00	0.89	0.94	18
GERD	1.00	0.90	0.95	21
Gastroenteritis	1.00	0.93	0.96	28
Heart attack	0.93	1.00	0.97	28
Hepatitis B	1.00	1.00	1.00	17
Hepatitis C	1.00	1.00	1.00	20
Hepatitis D	1.00	1.00	1.00	22
Hepatitis E	1.00	1.00	1.00	12
Hypertension	1.00	1.00	1.00	27
Hyperthyroidism	1.00	1.00	1.00	20
Hypoglycemia	1.00	1.00	1.00	26
Hypothyroidism	1.00	1.00	1.00	19
Impetigo	1.00	0.75	0.86	16
Jaundice	1.00	1.00	1.00	28
Malaria	1.00	1.00	1.00	19
Migraine	1.00	1.00	1.00	20
Osteoarthristis	1.00	1.00	1.00	27
Paralysis (brain hemorrhage)	1.00	1.00	1.00	28
Peptic ulcer diseae	1.00	1.00	1.00	19
Pneumonia	1.00	1.00	1.00	19
Psoriasis	1.00	0.95	0.98	21
Tuberculosis	1.00	1.00	1.00	21
Typhoid	1.00	1.00	1.00	27
Urinary tract infection	1.00	1.00	1.00	26
Varicose veins	1.00	0.94	0.97	18
hepatitis A	1.00	1.00	1.00	17
accuracy			0.99	924
macro avg	0.99	0.98	0.99	924
weighted avg	0.99	0.99	0.99	924

Comparative Analysis

Feature selection can help improve the generalizability of machine learning models for disease risk prediction by reducing the number of features used in the model. The reason for this is that too many features can lead to overfitting, where the model becomes too complex and fits the training data too closely, resulting in poor performance on upcoming, unrecognized data. Feature selection can prevent overfitting and improve generalizability by selecting only the most informative features and removing noisy or irrelevant features. This can lead to more accurate predictions of disease risk and better outcomes for patients. [4]

In the context of feature selection, a study conducted by Huizhong et al. [10] proposed a hybrid feature selection method based on tree-reinforcing gradients and eliminating recursive features with cross-validation (RFECV) to reduce feature redundancy and retain the most discriminatory features. This method can balance the Simplicity model and learning performance to select the best subset of features. The proposed method has been compared to other basic models, such as Logistic Regression, Random Forest, Gradient Boosting, Extreme Gradient Boosting, Multilayer Perceptron, and 1D Convolutional Networks. The results show that the proposed method achieves the highest accuracy at 98.8%. In contrast, traditional feature selection methods such as filter methods, wrapper methods, and embedded methods may not be able to handle high-dimensional and complex datasets like the ACS dataset used

in this paper. These approaches may also exhibit overfitting or underfitting difficulties, as well as a lack of accessible learning aspects. As a result, the hybrid feature selection method described here, which is based on enhancer tree slope and recursive feature elimination by cross validation (RFECV), is a more effective and adaptable methodology for feature selection in the context of ACS risk prediction.

Conclusion

We successfully retrieved 47 features out of the initial 132 features after going through the feature selection process, which enhanced the accuracy. The evaluation stages and selection evaluations have been completed, including analyzing the feature coefficients, evaluating the five estimator models, and evaluating the outcomes using the confusion matrix. Based on the evaluation findings, it is possible to infer that the technique used efficiently selected suitable characteristics while addressing the issue of overfitting. It has been verified that the selected features effectively contribute to forecasting the desired target by taking into account the attribute coefficients and running several estimator models. The confusion matrix findings also show that the developed model performs effectively and delivers good prediction accuracy. This indicates that the model has a strong capacity to discriminate between positive and negative classes. Overall, the feature selection and accuracy evaluation stages optimized the model, minimized overfitting, and produced satisfactory results.

Bibliography

[1] W. C. Hsu and J. H. Li, "Visualising and mapping the intellectual structure of medical big data," *J. Inf. Sci.*, vol. 45, no. 2, pp. 239–258, 2019, doi: 10.1177/0165551518782824.

[2] M. Kwong, H. L. Gardner, N. Dieterle, and V. Rentko, "Optimization of Electronic Medical Records for Data Mining Using a Common Data Model," *Top. Companion Anim. Med.*, vol. 37, p. 100364, 2019, doi: 10.1016/j.tcam.2019.100364.

[3] R. N. Ndung'u, G. N. Kamau, and G. W. Mariga, "Using Feature Selection Methods to Discover Common Users' Preferences for Online Recommender Systems," *Int. J. Comput. Inf. Technol.*, vol. 10, no. 1, pp. 24–32, 2021, doi: 10.24203/ijcit.v10i1.71.

[4] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan, "A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction," *Front. Bioinforma.*, vol. 2, no. June, pp. 1–17, 2022, doi: 10.3389/fbinf.2022.927312.

[5] A. Site, J. Nurmi, and E. S. Lohan, "Systematic review on machine-learning algorithms used in wearable-based eHealth data analysis," *IEEE Access*, vol. 9, pp. 112221–112235, 2021, doi: 10.1109/ACCESS.2021.3103268.

[6] D. Panda, R. Ray, A. A. Abdullah, and S. R. Dash, "Predictive Systems: Role of Feature Selection in Prediction of Heart Disease," *J. Phys. Conf. Ser.*, vol. 1372, no. 1, pp. 1–9, 2019, doi: 10.1088/1742-6596/1372/1/012074.

[7] S. Balasubramaniam, K. Satheesh Kumar, V. Kavitha, A. Prasanth, and T. A. Sivakumar, "Feature Selection and Dwarf Mongoose Optimization Enabled Deep Learning for Heart Disease Detection," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–11, 2022, doi: 10.1155/2022/2819378.

[8] S. M. U. and B. C. Kovoor, "An aggregated deep convolutional recurrent model for event based surveillance video summarisation: A supervised approach," *IET Comput. Vis.*, vol. 15, pp. 297–311, 2021, doi: 10.1049/cvi2.12044.

[9] Y.-C. Cho, H. Choi, M.-G. Lee, S.-H. Kim, and J.-K. Im, "Identification and Apportionment of Potential Pollution Sources Using Multivariate Statistical Techniques and APCS-MLR Model to Assess Surface Water Quality in Imjin River Watershed, South Korea," *Water*, vol. 14, p. 793, 2022, doi: 10.3390/w14050793.

[10] H. Lin, Y. Xue, K. Chen, S. Zhong, and L. Chen, "Acute coronary syndrome risk prediction based on gradient boosted tree feature selection and recursive feature elimination: A dataset-specific modeling study," *PLoS One*, vol. 17, no. 11 November, pp. 1–24, 2022, doi: 10.1371/journal.pone.0278217.