Cardiovascular Disease Prediction Using Hybrid Feature Selection with Random Forest and SVM

Keerthika N[1*] and Nithyanadam S[2]

[1*]Department of Computer Science and Engineering, Ponnaiyah Ramajayam Institute of Science and Technology (PRIST) Deemed to be University , Thanjavur, 613403, Tamil Nadu, INDIA.

[2]Department of Computer Science and Engineering, Ponnaiyah Ramajayam Institute of Science and Technology (PRIST) Deemed to be University , Thanjavur, 613403, Tamil Nadu, INDIA.

*Corresponding author(s). E-mail(s): n.keerthi.edu@gmail.com;
Contributing authors: snsirvp@gmail.com;

Abstract

Our lives are better because of the Internet of Things, seamless communication between people and things, and its fusion with the cloud. With improved artificial intelligence, predictive analytics in the medical field can assist in transforming a reactive healthcare approach into a proactive one, and with the widespread use of machine learning techniques in the healthcare sector, machine learning and deep learning have the revolutionary capacity to accurately and quickly analyze vast volumes of data, draw perceptive conclusions, and successfully solve complicated issues. To provide preventative treatment alongside early intervention for at-risk people, reliable and timely disease prediction is essential. The suggested system gathers data from IoT devices, and patient history-related electronic clinical data stored in the cloud is subjected to predictive analytics. The proposed work has two phases, the first of which is hybrid feature selection with the help of filters and wrappers. In this work, we will concentrate on filter-based methods for feature selection such as Chi-square, Fast Correlation Based Filter (FCBF), Gini Index (GI), RelieF, and wrapper-based methods for feature selection such as Backward Feature Elimination (BFE), Exhaustive Feature Selection (EFS), Forward Feature Selection (FFS), and Recursive Feature Elimination (RFE). The second phase is hybrid classification using machine learning techniques like Random Forest, Decision Tree, Naive Bayes, K-nearest neighbours, and

Support Vector Machine. Heart disease prediction is predicted by the probabilistic ensemble voting method. The performance metrics (F1 Score, Precision, and Recall) for all the techniques revealed that Naive Bayes is better without parameter tuning. At the same time, the Random Forest algorithm proved to be the best technique for hyperparameter tuning.

Keywords: feature selection, classification, IoT-cloud, Heart disease prediction, health care system

1 Introduction

Traditional methods of feature selection in heart disease prediction often involved manual selection based on expert knowledge and statistical techniques like

The analysis of each feature individually to determine its correlation with the target variable (heart disease). Standard techniques include statistical tests like t-tests, ANOVA (Analysis of Variance) [1], or correlation coefficients. Forward selection or backward elimination methods iteratively add or remove features based on specific criteria (like p-values or information criteria) until the best subset of features is identified. The Principal Component Analysis (PCA) [2] technique reduces dimensionality by creating a new collection of features from the original orthogonal variables called principal components, which aim to retain most of the variability in the data. Wrapper Methods Algorithms are used to evaluate different subsets of features by training models and selecting the subset that results in the best model performance. Examples include Recursive Feature Elimination (RFE) or Genetic Algorithms. Current approaches in combined hybrid feature selection for heart disease prediction involve more sophisticated and adaptive methods [3]like Feature selection techniques are integrated into several machine learning systems. For instance, decision trees and ensemble methods like Random Forests perform implicit feature selection by considering feature importance scores during model training. Ensemble techniques combine multiple feature selection methods to maximize their strengths, aiming to improve predictive performance by selecting robust and relevant features. Hybridization of Multiple Techniques Combining various feature selec- tion methods [4][5][6][7][8], including filter, wrapper, and embedded methods, to exploit the strengths of each approach and enhance predictive accuracy.

Current research and advancements in machine learning and artificial intelligence continuously evolve. New hybrid feature selection approaches for heart disease prediction may emerge, aiming for better accuracy, interpretability, and computational efficiency. This paper focuses on Combined hybrid feature selection for heart disease prediction in the cloud-based IoT health care system. Healthcare has undergone a transformative evolution by integrating cutting-edge technologies, notably Cloud-based Internet of Things (IoT) systems, in predictive diagnostics and personalized treatment. Among the pivotal applications lies heart disease prediction, where these innovative systems offer unparalleled potential in revolutionizing early detection, monitoring, and proactive intervention strategies.

The integration of IoT-enabled monitoring of health devices is the cornerstone of this paradigm shift, cloud infrastructure, advanced data analytics, and machine learning algorithms. IoT devices, ranging from wearable sensors to implantable monitors, gather physiological data, including heart rate, blood pressure, electrocardiogram (ECG) signals, and activity levels, continuously collecting information from individuals in real time. These data

streams are channelled securely to the cloud infrastructure, the backbone of this sophisticated healthcare ecosystem. Cloud computing furnishes the computational power, storage capabilities, and analytical tools required to process and obtain valuable insights from the enormous amounts of health data these IoT devices produce. At the heart of this system lies the predictive prowess of machine learning algorithms designed explicitly for classification tasks in heart disease prediction. These algorithms, including logistic regression, support vector machines, neural networks, and ensemble methods, are trained on integrated and preprocessed datasets derived from diverse IoT sources. Feature engineering and selection techniques help distil pertinent health indicators instrumental in identifying patterns and predicting the likelihood of heart disease onset or progression.

However, the significance of this system extends far beyond predictive modelling. This framework's real-time predictions and decision support systems empower healthcare providers with timely insights and recommendations, facilitating early interventions and personalized healthcare strategies. Patients benefit from continuous monitoring, receiving alerts and actionable guidance based on their health profiles, promoting proactive and preventive care measures. The assurance of data security, privacy, and regulatory compliance constitutes a pivotal pillar of this technological innovation. Robust encryption protocols, stringent access controls, and adherence to healthcare are essential to protect patient privacy and confidentiality; laws l.kLawse Healthcare Portability and Accountability Act (HIPAA) in the US are essential.

This Cloud-based IoT healthcare system [9][10] transcends the traditional reactive approach to healthcare by fostering a continuous feedback loop. Regular updates, model retraining, and optimization based on incoming data and feedback ensure that the predictive models remain accurate and adaptive to evolving health conditions and patient needs. Integrating Cloud-based IoT systems for heart disease prediction marks a watershed moment in healthcare. This amalgamation of cutting-edge technologies enhances predictive capabilities. It lays the groundwork for a paradigm shift towards personalized, proactive, and data-driven healthcare initiatives, which ultimately result in better patient outcomes and a more wholesome society. In modern healthcare, the convergence of Cloud-based Internet of Things (IoT) systems with the power of Machine Learning (ML) has brought about a transformative wave. This fusion of technologies promises to redefine healthcare delivery by enabling predictive diagnostics, personalized treatments, and proactive wellness management.

At its core, a Cloud-based IoT healthcare system harnesses the interconnectedness of devices and the omnipresence of cloud infrastructure. IoT devices, from wearable sensors to implantable monitors, continually capture diverse health data streams. These data encompass many physiological metrics such as heart rate variability, blood pressure fluctuations, ECG patterns, activity levels, and more, creating a comprehensive digital footprint of an individual's health status. Transiting this voluminous and real-time health data to the cloud establishes a robust foundation for healthcare analytics and intervention. Cloud computing provides the necessary computational muscle, storage capacity, and sophisticated analytical tools pivotal in processing, analyzing, and deriving actionable insights from these massive datasets.

Embedded within this ecosystem, Machine Learning algorithms are the linchpin in driving predictive modelling and decision-making. These algorithms, from traditional regression models to advanced neural networks and ensemble methods, operate on amalgamated and

refined datasets from IoT sources. Through feature engineering, these algorithms discern intricate patterns and correlations within the data, enabling the prediction of diseases, prognoses, and treatment outcomes with increasing accuracy. The significance of this ML-driven approach extends far beyond predictive analytics. Real-time decision support systems and predictive models empower healthcare providers with invaluable insights, facilitating timely interventions and personalized patient care pathways. Moreover, patients benefit from continuous monitoring, receiving personalized recommendations and alerts based on their health data, fostering a culture of proactive healthcare management.

However, using Machine Learning in a Cloud-based IoT healthcare system is not merely confined to predictive modelling. It operates within a continuous improvement loop, where models are continuously trained, validated, and optimized using incoming data. This iterative process ensures that predictive models evolve and adapt to changing health dynamics and patient needs, ensuring relevancy and accuracy. Furthermore, the deployment of robust security measures is imperative within these systems. Encryption protocols, stringent access controls, and adherence to healthcare regulations ensure the safeguarding of sensitive patient data, preserving confidentiality and privacy.

Integrating Machine Learning into Cloud-based IoT healthcare systems signifies a monumental stride towards a more efficient, proactive, and personalized healthcare landscape. This fusion of technologies augments predictive capabilities and establishes a framework for a data-driven, patient-centric healthcare paradigm. It sets the stage for transformative healthcare delivery, driving improved patient outcomes and fostering a healthier global community. The following contributes can be summarizing are

- The novelty of this paper is heart disease prediction, and the approach involves a combination of hybrid feature selection and machine learning classification.
- The novelty of sixteen feature selection (FS) methods and a domain knowledge database (DKDB) are utilized. The DKDB-based FS method prioritizes features based on cardiologists' expertise.
- The proposed approach employs six hybrid ML classifiers and two voting classifiers.
- The paper is evaluated independently using publicly available data sets, including Cleveland, Statlog, and Z-Alizadeh Sani.
- The proposed methods are tested on various scenarios, including complex vs. non-complex, linear vs. nonlinear, balanced vs. imbalanced, and sparse vs. non-sparse data sets.
- Performance The results of the accuracy of the proposed model are reported as 94.58%, 93.85%, and 94.97% on the Z-Alizadeh Sani, Statlog, and Cleveland data sets, respectively.

The remainder of this paper is as follows:. Section II provides an overview of related work. In Section III, the proposed system and formulation of IoT based health care system are described. In Section IV, the experiment setup details. Section V contains the results of experimental evaluations and corresponding discussions. Finally, concluding remarks are presented in Section VI.

2 Related work

[11] introduced innovative ensemble feature selection methods and assessed them along- side computational and domain knowledge-based approaches. Utilizing a mix of six single classifiers and four ensemble voting classifiers, the methods underwent hyper-parameter optimization through grid-search techniques. The evaluation covered diverse publicly available datasets, such as Cleveland, Statlog, and ZAlizadeh Sani, rigorously testing the proposed methods across various dataset characteristics. Results showcased their effectiveness across complex vs. noncomplex, linear vs. nonlinear, balanced vs. imbalanced, and sparse vs. nonsparse datasets. Performance analysis highlighted impressive accuracy rates: 91.78% on Z-Alizadeh Sani, 85.55% on Statlog, and 85.47% on the Cleveland dataset, demon- strating the model's efficacy across multiple datasets with differing attributes. [12] The RF-FSFC model aims to enhance heart disease classification accuracy compared to estab- lished models by leveraging sensitivity and correlation techniques for input variable selection. Sensitivity-based selection emphasizes crucial features in assessing CHD risk, while sim- ilarity analysis identifies feature relationships. Extensive experiments using performance metrics like accuracy, confusion matrix, PPV, and NPV validate the model's predictive effi- ciency and reliability in heart disease prediction. [13] The objective is to develop a Hybrid Adadelta Stochastic Gradient Classifier-based Healthcare Hash Big Data Storage (HADSGC- HHBS) method. This method efficiently stores and manages clinical information gathered from diverse locations within a distributed setting. The primary goal is to achieve optimal data storage while minimizing space usage and ensuring faster processing times. [14]Three distinct feature selection approaches - chi-square, ANOVA, and mutual information - were employed to identify feature subsets denoted as SF1, SF2, and SF3, respectively. Imple- menting a meta-heuristic feature selection approach aimed to boost overall classification accuracy by minimizing feature dimensions. This reduction is geared towards refining accu- racy and speed in classification. This supervised machine-learning task utilised Heart disease datasets from the UCI Machine Learning repository [15]. While heart disease has been extensively researched, solving it requires sophisticated algorithms rather than simplistic machine-learning models. The project explores algorithms such as linear regression (LR) and decision tree (DT) to address this. These analyses involve the application of various feature selection methods to the datasets.

[6]A comparative study was proposed to gauge how feature selection methods and population sizes impact results. Notably, WOA tended to select fewer features, though thoroughly assessing population size proved challenging. While KNN topped in classification success, using reduced features from the CS algorithm resulted in higher average success rates. [16]The investigation aimed to assess the efficiency of different machine learning algorithms in forecasting heart disease. The study employed several methodologies to construct predictive models, such as random forest [15], decision tree classifier, multilayer perceptron, and XGBoost [17].

We implemented k-mode clustering as a preprocessing step to enhance model convergence to scale and refine the dataset. The dataset utilized in this research is publicly accessible on Kaggle. All computational tasks, including preprocessing and visualization, were executed using Python on Google Colab. [18] The Sequential Feature Selection (SFS) technique initiates with an empty set and progressively includes elements that offer the most substantial

contribution to the intended objective in the initial phase. Subsequently, starting from the second phase, the remaining features are meticulously added to the current subset. The algorithm selects multiple features from a pool of available features effectively. It evaluates their impact on model iteration, iteratively reducing or enhancing the number of features to attain optimal performance and desirable outcomes.

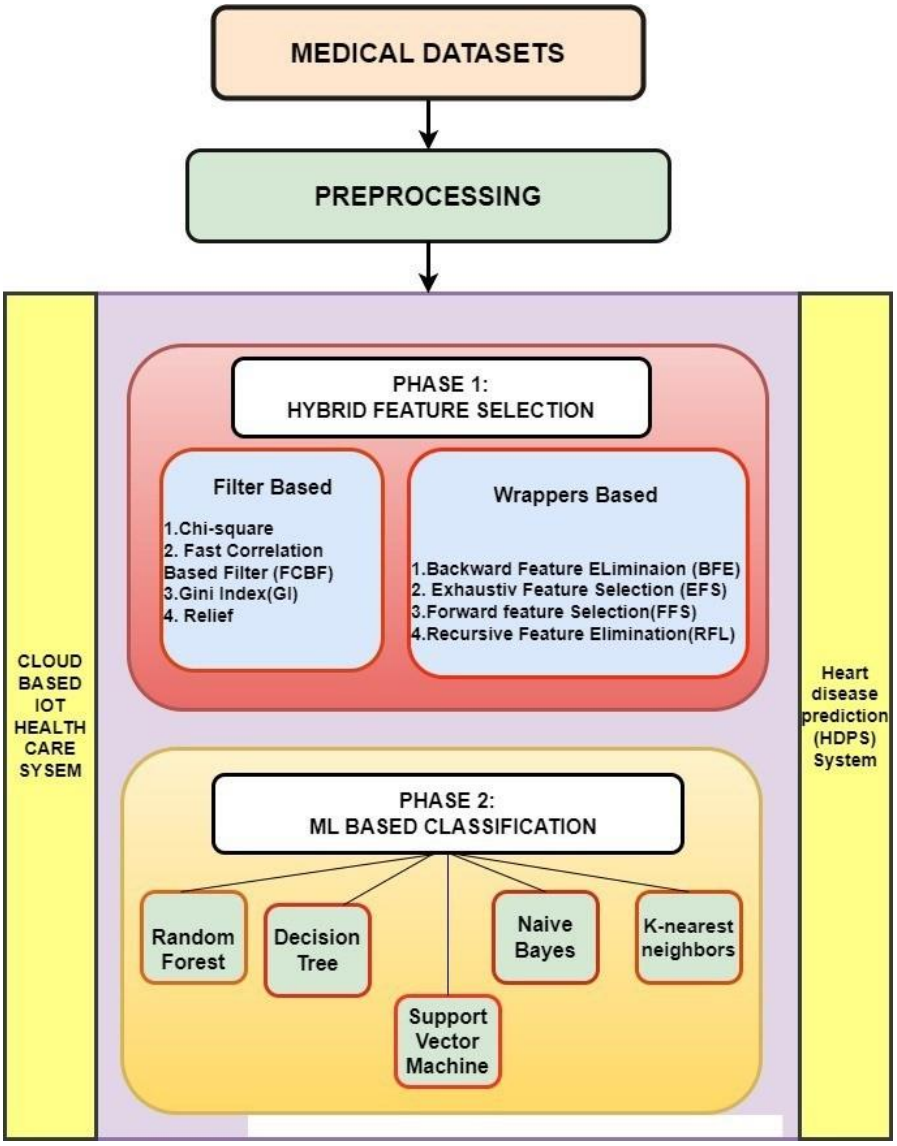

Fig. 1: Proposed work Framework diagram.

3 PROPOSED SYSTEM AND FORMULATION

The process of choosing only particular features from the feature set to improve classification accuracy is known as feature selection. The likelihood that every component in the dataset will be helpful for building. A model is accurate when building a machine learning model. Repetitive variables reduce the model's classification accuracy and may limit a classifier's maximum performance. Additionally, the model's overhead increases as more and more parameters are added to a specification. From the n variables produced for determining maximum and minimum risk of cardiovascular disease, we choose factors based on significance in the link to good classification.

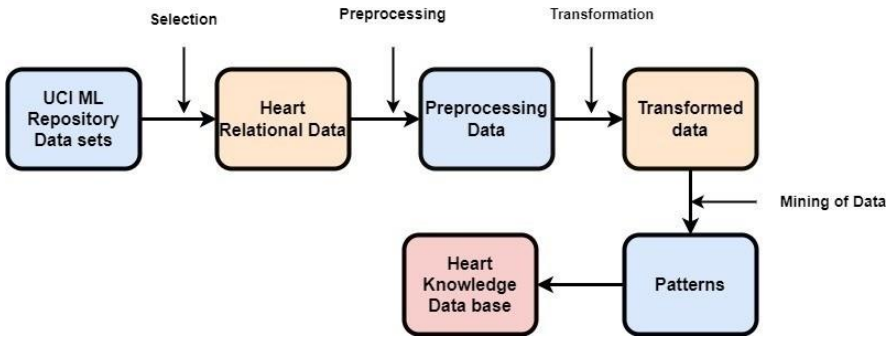

Fig. 2: Heart Knowledge database

3.1 Datasets

The paper following three datasets from taken from the Kaggle archives.

- **Z-Alizadeh Sani:** A patient may fall into one of two categories: Normal or CAD. If a patient's diameter narrowing is more than or equal to 50%, they are classified as having CAD; otherwise, they are classified as usual. Note 1: This extension now includes LAD, LCX, and RCA functions. When any of the three arteries above is stenotic, at least one develops CAD (the final column in the dataset). Only one of the LAD, LCX, RCA, or Cath (Result of The procedure) must be present in the dataset to use; the remaining ones must be removed to classify them. Note 2: The LAD, LCX, and RCA arteries can all be diagnosed with stenosis using this dataset in addition to CAD identification.
- **Cleveland:** Although there are 76 attributes in this database, only 14 are used in the published studies. ML researchers have only utilized one database, specifically the Cleveland database. The patient's heart disease status is indicated in the "goal" field. It is an integer between 0 (no presence) and 4. The focus of Cleveland database experiments has been to differentiate between presence (values 1, 2, 3, 4) and absence (value 0). Recently, the social security numbers and patients' names were removed from the database and replaced with dummy entries. A single file containing the Cleveland database has been "processed". There are also the four raw files in this subfolder.

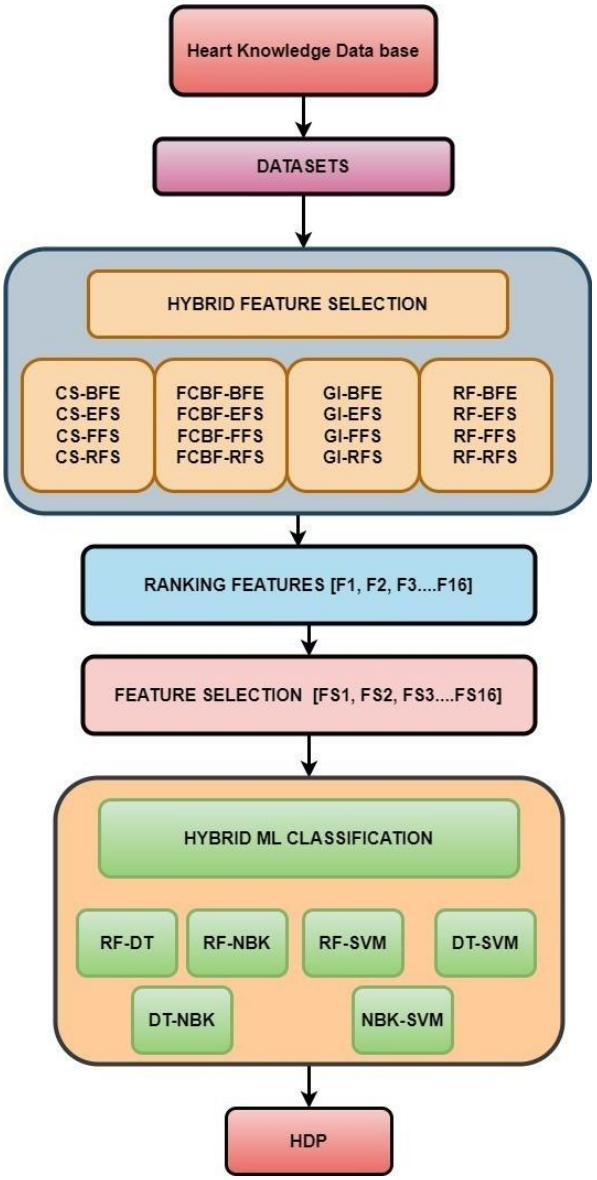

Fig. 3: Schematic representation of the probabilistic ensemble proposed work.

- **Statlog:** This dataset is a modified rendition of an existing repository housing heart disease databases, sharing similarities with the dataset currently under discussion. Its relevance lies within health and medicine, specifically tailored to a specialized categorization task.
  - **Multivariate Nature:** Encompasses multiple traits or variables within its structure.
  - Subject Area: Primarily focused on heart health and associated medical domains.

- **Related Tasks:** Tailored for classification purposes, particularly aiding in predicting or diagnosing heart disease. Feature Types: Encompasses absolute and categorical features, representing fundamental attributes and categorical data, likely comprising clinical indicators and various health factors.
- **Dataset Size:** Comprises 270 instances or observations.

This dataset stands as a crucial resource within the field of health and medicine, offering a platform conducive to developing and refining classification models. Its emphasis on heart-related data and its varied feature types make it instrumental for researchers and medical professionals in enhancing diagnostic capabilities and formulating effective treatment strategies for heart-related ailments.

The Heart Disease dataset collection encompasses four distinct datasets, among which the Cleveland [19] dataset was chosen for analysis due to its relatively lower incidence of missing values than the others. Within this dataset, originally consisting of 76 features, we opted to work with 13 features devoid of missing values. The Cleveland dataset contains 303 samples, while the Statlog dataset comprises 270. Notably, only the Cleveland dataset manifests six missing values, omitted from this study rather than undergoing data correction procedures.

In both the Cleveland and Statlog datasets, the classification of each sample is determined based on a critical health indicator: the narrowing of blood vessels. Samples are categorized as (i) healthy if their vessels demonstrate less than a 50% narrowing, while those exceeding this threshold are labelled as (ii) CAD, signifying coronary artery disease. The decision to utilize the Cleveland dataset over others within the Heart Disease collection primarily stemmed from its relatively lower count of missing values across its 76 features. In our analysis, we chose to work with 13 features within the Cleveland dataset that were free of any missing data, ensuring robustness and integrity in our examination.

Contrarily, the Z-Alizadeh Sani [20] dataset presents a diverse set of health indicators encompassing 303 samples and 55 features, meticulously organized into distinct categories such as 'Demographics', 'Symptom and Examination', 'ECG', and 'Laboratory and Eco'. In this dataset, each sample is classified as either (i) healthy or (ii) unhealthy, a broader categorization that allows for a more comprehensive understanding of health conditions beyond the specific vessel-narrowing criteria used in the Cleveland dataset.

For a deeper understanding of the characteristics and comprehensive descriptions of both the Cleveland and Z-Alizadeh Sani datasets, we refer interested readers to our earlier detailed studies [21], [22], where an in-depth analysis and breakdown of these datasets have been provided.

3.2 Feature selection method

Feature selection methods are techniques for choosing the most relevant and informative features from a dataset. These methods aim to improve model performance, reduce overfitting, enhance computational efficiency, and enhance interpretability by selecting the most critical attributes for modelling. There are five common types of feature selection methods.

1. **Filter Methods:** Filter methods assess the intrinsic characteristics of features independently of any specific machine learning algorithm. These methods involve statistical tests, correlation coefficients, or information gain metrics to rank or score features based on relevance. They are computationally efficient and can handle large datasets well.

2. **Wrapper Methods:** Wrapper methods select features by evaluating subsets of features using a specific machine learning algorithm. Algorithms like Forward Selection, Backward Elimination, and Recursive Feature Elimination (RFE) are used to search for the best feature subset that optimises model performance. Considers feature interactions and their impact on model performance. It is computationally expensive, especially with a large number of features.
3. **Embedded Methods:** Embedded methods integrate feature selection while training a machine learning model. Algorithms like Lasso Regression, Decision Trees, Random Forests, and Gradient Boosting Machines inherently perform feature selection during model training by assigning importance scores to features. Simultaneously, it learns feature importance and model parameters.
4. **Hybrid Methods:** Hybrid methods combine aspects of different feature selection approaches to leverage their strengths. Techniques that merge filter and wrapper methods or use a combination of embedded and wrapper methods to obtain a better subset of features. Tries to mitigate the limitations of individual methods by combining their strengths. It may increase computational complexity and require fine-tuning.
5. **Dimensionality Reduction Techniques:** Techniques like Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) reduce the dimensionality of data by transforming features into a lower-dimensional space. Techniques: These meth- ods project high-dimensional data onto a lower-dimensional subspace while retaining the most essential information. It helps reduce computational complexity and address multicollinearity. Reduced interpretability due to transformed features. Choosing the appropriate feature selection method depends on the dataset size, the relationship between features, computational resources, and the specific machine-learning problem. Experimentation and testing with different methods are often necessary to determine the most effective approach for a given task.

The proposed model is based on the filter and wrapper feature selection hybrid. The hybrid selection techniques are crucial in enhancing model performance by identifying and utilizing the dataset's most relevant subset of features. Filtering and wrapping methods are two distinct approaches employed for this purpose, each with its characteristics and implications for model training and performance.

Filter methods excel in computational efficiency as they do not involve model training, making them significantly faster. Conversely, wrapper methods are computationally expensive since they repeatedly train models to evaluate different feature subsets. Filtering methods rely on statistical measures to assess the relevance of features within a subset, aiming to retain the most informative ones. In contrast, wrapper methods employ cross-validation, utilizing the performance of models trained on different feature subsets to determine feature relevance. Filtering methods may only sometimes identify the best feature subset due to their reliance on statistical measures, potentially missing intricate relationships between features. Through their exhaustive search and evaluation process, Wrapper methods aim to identify the optimal subset of features, providing a more comprehensive selection. Utilizing a subset of features selected via wrapper methods makes the model more robust, considering complex interactions between features. However, relying solely on the subset of features from filter methods might render the model more prone to underfitting or oversimplification, potentially limiting its capacity to capture nuanced patterns in the data.

While filter methods offer faster processing and can efficiently select relevant features, their inability to capture intricate relationships might affect the model's predictive capabilities. Wrapper methods, though computationally expensive, provide a more exhaustive search, enhancing the chances of optimizing model performance by selecting the most pertinent features. Choosing filter and wrapper methods for feature selection in Machine Learning involves a trade-off between computational efficiency and optimality in identifying the most relevant feature subset. While filter methods offer speed, wrapper methods provide more comprehensive and potentially superior subsets, which can significantly impact the model's predictive performance and robustness.

3.3 ML based Classification

Machine learning-based categorisation is a method for classifying data into different classes or categories based on trends and data found in the dataset. It uses various machine learn- ing techniques trained on labelled training data to create estimates or assign class labels to previously unseen or fresh data. The subsequent phase involves employing diverse ML techniques for classification tasks. Various algorithms such as Random Forest, Decision Tree, Naive Bayes, K-nearest neighbours (KNN), and Support Vector Machine (SVM) categorize or classify data points into distinct classes or categories.

In the realm of Machine Learning (ML), the extraction of features from temporal data holds paramount importance. Time-based feature extraction involves deriving meaningful and informative features from timestamps or time-related data. These extracted features serve as crucial inputs for ML models, enabling them to comprehend temporal patterns, trends, and dependencies within datasets. In the context of time-series data or datasets containing temporal information, feature extraction involves various methodologies to effectively capture and represent temporal characteristics. These methodologies pave the way for unveiling insightful patterns and aiding predictive modelling in numerous domains. Time-based feature extraction is a cornerstone in enabling ML models to comprehend and exploit temporal relationships within datasets. Whether predicting stock prices, forecasting demand, or analyzing user behaviour, incorporating time-based features empowers models to make more accurate predictions and generate insights crucial for decision-making.

Despite the strides made in time-based feature extraction, challenges persist, including handling irregular time intervals, coping with missing data, and effectively capturing complex temporal dependencies. Future endeavours in ML may focus on developing more robust methodologies that can adapt to diverse temporal patterns and better accommodate real-world complexities. Time-based feature extraction in Machine Learning is a pivotal step in comprehending temporal dynamics within datasets. By leveraging various techniques to extract meaningful temporal features, ML models can unlock valuable insights, empowering enterprises and scholars to make knowledgeable choices and predictions based on the temporal intricacies of their data.

3.4 IoT healthcare System

In recent years, the fusion of Internet of Things (IoT) technology with healthcare systems has ushered in a new era of innovation, fundamentally reshaping the landscape of patient care, management, and healthcare delivery. The convergence of interconnected devices, data

analytics, and advanced connectivity within the IoT framework has propelled the healthcare sector towards a future prioritizing personalized, efficient, and proactive approaches to well-ness and treatment. The implementation of IoT in healthcare begins with an intricate web of intelligent, interconnected devices and sensors designed to monitor and collect real-time health data. These technological marvels continuously gather comprehensive health met- rics, from wearable devices tracking vital signs and activity levels to implantable sensors measuring glucose levels or medication adherence.

The seamless transmission of this data through robust connectivity protocols allows instantaneous transfer to centralized healthcare systems and cloud platforms. Here, the information undergoes meticulous analysis, leveraging cutting-edge machine learning and analytics techniques to extract valuable insights and identify meaningful patterns. These understandings enable healthcare professionals to make wise choices. They were prompt, leading to more accurate diagnoses, timely interventions, and personalized treatment plans. IoT's most significant effect on healthcare is its ability to engage and empower patients in their health management. User-friendly mobile applications are gateways for patients to access their health data, receive personalized recommendations, set medication reminders, and communicate directly with healthcare professionals. This newfound connectivity fos- ters a sense of ownership over one's health, promoting proactive wellness behaviours and facilitating more meaningful patient-provider interactions.

IoT-enabled remote monitoring and telemedicine have transcended geographical barri-ers, enabling healthcare services to reach individuals in remote areas or those with limited mobility. This transformative shift not only reduces the burden on traditional healthcare facil-ities but also improves access to care and enhances the efficiency of healthcare delivery. Real-time monitoring and early detection of anomalies allow for proactive interventions, min-imizing hospital readmissions and preventing health complications. Amidst the revolutionary advancements, data security, interoperability, and regulatory compliance challenges loom large. Protecting sensitive patient information, ensuring seamless communication between diverse systems, and adhering to stringent healthcare regulations like HIPAA (Health et al. Act) demand continuous innovation and vigilance from healthcare organizations and tech-nology providers. The future of IoT in healthcare holds immense promise. Advancements in wearable technology, AI-driven predictive analytics, and the integration of IoT devices with Electronic Health Records (EHRs) are poised to revolutionize healthcare delivery further. As healthcare continues to evolve, the amalgamation of IoT technology with patient-centric care will remain at the forefront, empowering individuals, optimizing healthcare resources, and ultimately, transforming how we perceive and experience healthcare.

The fusion of IoT and healthcare represents a technological convergence and a paradigm shift towards a patient-centric, data-driven healthcare ecosystem. As this transformation con-tinues to unfold, the collaborative efforts of healthcare professionals, technology innovators, and regulatory bodies will be pivotal in harnessing the full potential of IoT to deliver more accessible, efficient, and personalized healthcare solutions for all.

3.5 Proposed Work

The primary aim of FS is to assign significance scores to features in a dataset, guiding the selection of features based on these scores. High scores are assigned to essential features,

Table 1: Hybrid Feature Selection and probabilistic score calculation

S.No	Hybrid Feature Selection	Ensemble Score Feature	Probabilistic Feature Selection Score
1	(CS-BFE)	3.12	0.78
2	(CS-EFS)	3.23	0.8075
3	(CS-FFS)	3.01	0.7525
4	(CS-RFS)	2.12	0.53
5	(FCBF-BFE)	1.65	0.4125
6	(FCBF-EFS)	1.98	0.495
7	(FCBF-FFS)	1.78	0.445
8	(FCBF-RFS)	1.99	0.4975
9	(GI-BFE)	3.23	0.8075
10	(GI-EFS)	3.26	0.815
11	(GI-FFS)	3.16	0.79
12	(GI-RFS)	3	0.75
13	(RF-BFE)	2.89	0.7225
14	(RF-EFS)	3.23	0.8075
15	(RF-FFS)	2.87	0.7175
16	(RF-RFS)	3.67	0.9175

while uninformative and redundant ones receive low scores. FS is advantageous for improving performance results and reducing computational time [23]. However, each FS method has unique effects, leading to variations in classification performance [24]. Notably, a fea- ture considered essential by one FS method may receive a low score from another, creating challenges in establishing a consistent feature score with a high confidence level [25].

Sixten FS methods in hybrid FS in Fig.3 on the Z-Alizadeh Sani dataset significantly deem the 'Tinversion' feature. However, it is regarded as irrelevant by the DKDB-based FS.

In the existing exhaustive ensemble FS approach method, there is a risk of overlooking low-scoring features during the classification process, potentially missing opportunities for improved performance. So, the proposed method, the probabilistic ensemble FS approach, has been proposed to counter this concern. In this probabilistic ensemble FS method, a probabilistic score is computed for each feature using various FS methods. These probabilistic scores play a pivotal role in determining feature selection rates. Subsequently, different sets of features are selected and evaluated in the classification process. Feature selection plays a pivotal role in machine learning, influencing the performance and efficiency of models. Traditional approaches often rely on a single feature selection method, which may lead to inconsistencies and variations in results across different datasets. In response to these challenges, the Probabilistic Ensemble Feature Selection approach has emerged as a promising solution, aiming to enhance the robustness and reliability of feature selection by considering the probabilistic nature of feature scores. FS identifies and selects a subset of relevant features from Z-Alizadeh Sani, Cleveland and Statlog datasets. The goal is to improve model performance, reduce computational complexity, and enhance interpretability.

Probabilistic scores are computed for each feature using a hybrid FS of sixteen different FS methods. These scores represent the likelihood or probability of a significant feature based on the outcomes of sixteen methods. Table 1 presents an illustrative instance of the probabilistic score computation for sixteen distinct features using the ensemble scores. The ensemble

score for each feature is calculated as the average of importance scores assigned by various hybrid FS methods, denoted in the second column in Table 1; the probabilistic scores for the features are determined by dividing the ensemble feature score by the total ensemble score. This calculation results in probabilistic scores that function as selection probabilities for each feature.

To clarify, the selection probability of a feature corresponds to the ratio of its ensem- ble feature score to the total ensemble score. A schematic representation of our probabilistic ensemble FS method is depicted in Fig. 4. The feature selection process involves generating a random number, and based on the probabilistic scores of the features, the feature corresponding to that random number is chosen. This approach allows for the selection of varying numbers of features, which are then tested in the subsequent classification process. Probabilistic scores play a crucial role in establishing the selection rates for features. Features with higher probabilistic scores are more likely to be selected for further analysis or model training. The approach allows for the consideration of multiple feature selection methods simultaneously. This flexibility mitigates the risk of relying too heavily on the outcomes of any single method, promoting adaptability to diverse datasets. This work endeavours to diagnose heart disease and experiments utilizing three publicly available CAD datasets, incorporating sixteen computational FS methods, a domain knowledge database (DKDB) FS method, six hybrid classifiers, and an ensemble classifier with four variations. To ensure the robustness of our analysis, samples with missing features were excluded from the datasets rather than imputed with synthetic data.

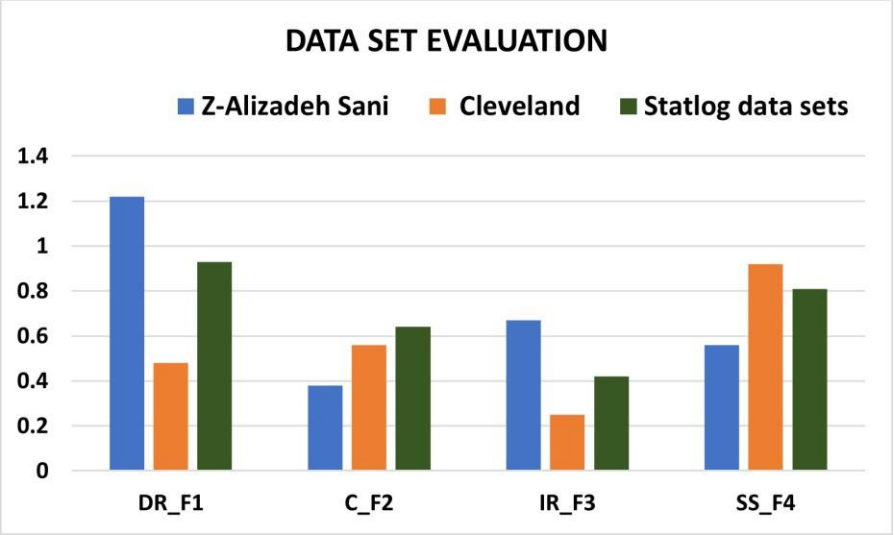

Fig. 4: Data set Evaluation

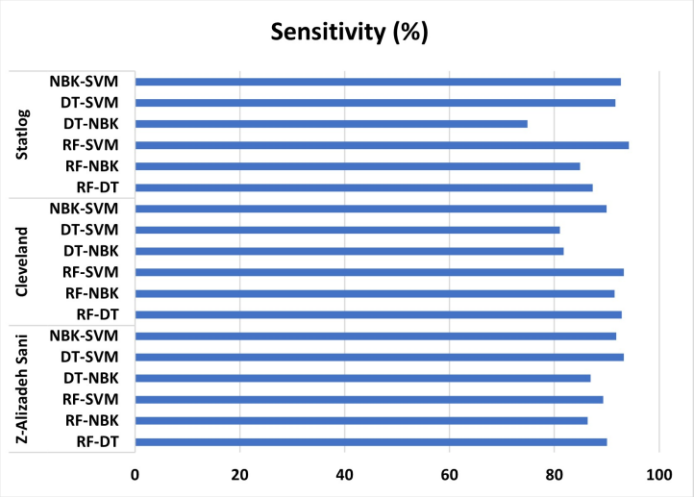

Fig. 5: Sensitivity Evaluation

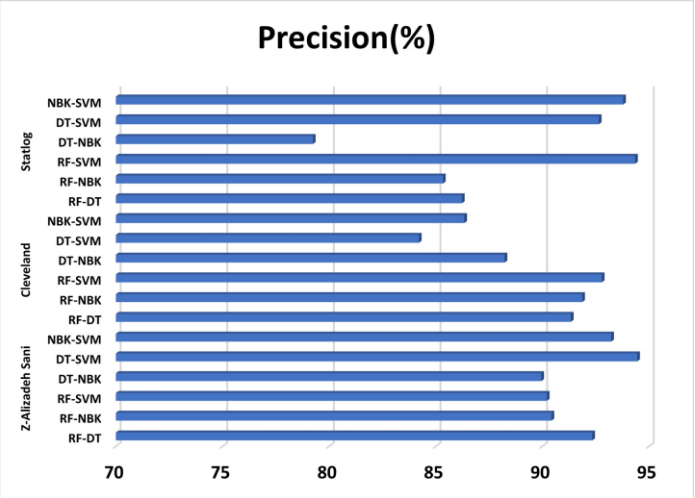

Fig. 6: Precision Evaluation

4 Experimental setup

Importance scores of features were computed using eight diverse FS methods on the Cleveland, Statlog, and Z-Alizadeh Sani datasets. The exhaustive ensemble FS approach: Ensemble scores of features were computed 2^{16} times, representing each combination of the sixteen FS methods 2^{16} different lists, each showcasing feature scores, are generated. Varying numbers of features (t) are tested for each list, ranging from 1-2^{16}. Computational efficiency considerations set M to 25 for the Z-Alizadeh Sani dataset and 22 for the Cleveland and Statlog

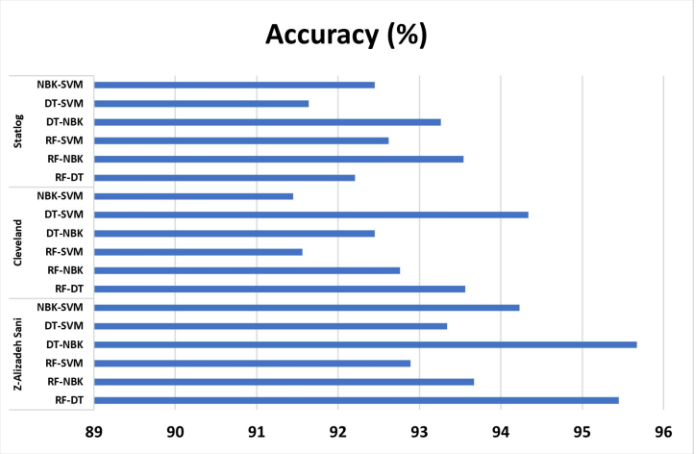

Fig. 7: Accuracy Evaluation

datasets. For each combination of FS methods, 21 (25 - 16 + 1) and 8 (22 - 16 + 1) combinations of top t features were generated for the Z-Alizadeh Sani and Cleveland/Statlog datasets, respectively.

The probabilistic ensemble FS approach is explicitly applied to the Z-Alizadeh Sani dataset. Due to the limited number of features (9–25) in the Cleveland and Statlog datasets, the probabilistic ensemble FS approach needed to be applied. On the Z-Alizadeh Sani dataset, feature sets ranging from 1 to 2^{16} are selected and separately tested in each classification.

The experiments highlighted the impact of computational time on the Z-Alizadeh Sani dataset when the number of features exceeded 25. Thus, this dataset is set to 25, generating 21 combinations of top-t features for each combination of FS methods. For the Cleveland and Statlog datasets, M was set to 12, generating eight combinations of top-t features for each combination of FS methods. Our meticulous experimental design, involving the implementation of exhaustive and probabilistic ensemble FS approaches, allows for a comprehensive evaluation of feature selection methods in diagnosing heart disease. The adaptation of these approaches to dataset-specific characteristics demonstrates our commitment to robust and applicable methodologies in the realm of heart disease prediction. This research employs a diverse set of hybrid classifiers comprising RF-DT, RF-NBK, RF-SVM, DT-NBK, DT-SVM, and NBK-SVM, as well as ensemble classifiers with varying configurations (S14, S24, S34, S44, H14, H24, H34, and H44). Soft and hard voting strategies are employed, with the number of classifiers specified by the following numerical indicators. Stratified 16-fold cross-validation is utilized for evaluation. Feature selection is conducted using Weka, and classification methods are implemented through Python with the Scikit-Learn library. The proposed exhaustive ensemble feature selection (FS) method is applied to the Z-Alizadeh Sani dataset, generating many models for comprehensive analysis. The potential of hybrid and ensemble classifiers with exhaustive feature selection enhances classification performance. Including diverse hybrid classifiers and ensemble configurations aims to identify optimal models for three datasets from DKDB.

The experiments employ six hybrid classifiers: RF-DT, RF-NBK, RF-SVM, DT-NBK, DT-SVM, and NBK-SVM. Each classifier integrates different machine learning algorithms, contributing to a diverse set of models for evaluation. Eight ensemble classifier variations (S^4, S^4_1 S^4_2 S^4_3 H^4_4 H^4_1 H^4_2 H^4_3) are utilized, incorporating soft and hard voting methodologies. The numerical indicators following S and H indicate the number of classifiers used in each ensemble, contributing to a broad spectrum of ensemble configurations. Stratified 16-fold cross-validation ensures robust evaluation and generalization of the classification models.

All feature selection processes are executed using Weka, and the classification methods are implemented in Python with the Scikit-Learn library. This combination allows for the seamless integration of diverse machine-learning techniques. The proposed exhaustive ensemble feature selection method is applied to the Z-Alizadeh Sani dataset, resulting in many models. Specifically, $(2^16 * 25 * 6)$ models are generated, providing a comprehensive basis for evaluating the impact of feature selection on classification performance. For the Cleveland and Statlog datasets, a total of $(2^16 * (25 + 10) * 6)$ models are generated using the exhaustive ensemble FS method. The probabilistic ensemble FS approach generates $(2^16 * 25 * 6)$ models for the Z-Alizadeh Sani dataset.

To advance classification performance by exploring hybrid classifiers and ensemble configurations coupled with exhaustive feature selection. The extensive number of models generated facilitates a thorough analysis of the proposed methodologies, providing valuable insights into the effectiveness of feature selection in enhancing classification accuracy across diverse datasets.

Table 2: Performance results of Hybrid classifier on three different Heart data sets

Dataset types	Hybrid ML classification	Number of Features	Sensitivity (%)	Precision(%)	F-Measure	Area Under Curv	Accuracy (%)
Z-Alizadeh Sani	RF-DT	25	90.05	92.34	0.89	0.098	95.45
	RF-NBK	25	86.32	90.43	0.87	0.89	93.67
	RF-SVM	23	89.34	90.21	0.92	0.92	92.89
	DT-NBK	21	86.91	89.95	0.78	0.93	95.67
	DT-SVM	20	93.23	94.45	0.89	0.91	93.34
	NBK-SVM	23	91.85	93.23	0.88	0.9	94.23
Cleveland	RF-DT	22	92.87	91.34	0.87	0.923	93.56
	RF-NBK	24	91.46	91.87	0.85	0.892	92.76
	RF-SVM	24	93.23	92.79	0.88	0.92	91.56
	DT-NBK	25	81.78	88.23	0.79	0.94	92.45
	DT-SVM	21	81.03	84.21	0.86	0.92	94.34
	NBK-SVM	20	89.99	86.34	0.87	0.93	91.45
Statlog	RF-DT	10	87.32	86.23	0.87	0.95	92.21
	RF-NBK	11	84.87	85.34	0.81	0.93	93.54
	RF-SVM	9	94.21	94.35	0.89	0.93	92.62
	DT-NBK	11	74.91	79.24	0.81	0.92	93.26
	DT-SVM	10	91.69	92.65	0.89	0.93	91.64
	NBK-SVM	9	92.67	93.79	0.88	0.94	9

5 Performance Results

This delves into the performance evaluation of various hybrid machine learning classifiers across three distinct datasets: Z-Alizadeh Sani, Cleveland, and Statlog. Each classifier is assessed based on essential metrics, including sensitivity, precision, F-measure, area under the curve (AUC), and overall accuracy shown in Table 2. The aim is to provide a thorough understanding of how different hybrid classifiers perform in diverse dataset scenarios.

Hybrid ML classifiers combine multiple base classifiers to leverage the strengths of individual algorithms. The study focuses on six hybrid classifiers: RF-DT, RF-NBK, RF-SVM, DT-NBK, DT-SVM, and NBK-SVM.

Sensitivity (%): Indicates the classifier's ability to identify positive instances correctly, shown in Fig. 5. **Precision (%):** Measures the accuracy of optimistic predictions , shoen in Fig.6. **F-Measure:** Harmonic mean of precision and sensitivity, providing a balanced metric. **Area Under Curve (AUC):** Represents the classifier's discrimination ability. **Accuracy (%):** Overall correctness of predictions shown in Fig. 7. Z-Alizadeh Sani Dataset: The RF-SVM hybrid classifier stands out with 92.89% accuracy, combining the strengths of Random Forest and Support Vector Machine. Notable performances are observed with DT-SVM (93.34% accuracy) and NBK-SVM (94.23% accuracy). RF-DT: Achieves an impressive accuracy of 95.45%, demonstrating a harmonious balance between sensitivity (90.05%) and precision (92.34%). RF-NBK: Maintains competitive accuracy at 93.67%, showcasing a balanced performance across various metrics. RF-SVM: Demonstrates high accuracy (92.89%) and AUC (0.92), indicating effective discrimination. DT-NBK: Yields an accuracy of 95.67%, emphasizing a balanced performance across multiple metrics. DT-SVM: Stands out with an accuracy of 93.34%, showcasing balanced precision and sensitivity. NBK-SVM: Combines vital precision (93.23%) and sensitivity (91.85%) for an accuracy of 94.23%. Cleveland Dataset: RF-DT exhibits strong performance with an accuracy of 93.56%, showcasing the synergy between Random Forests and Decision Trees. Both RF-NBK and RF-SVM also demonstrate competitive accuracy, emphasizing the versatility of these hybrid classifiers.RF-DT: Leads with an accuracy of 93.56%, excelling in precision, sensitivity, and AUC. RF-NBK: Maintains a high accuracy of 92.76%, demonstrating a well-balanced performance. RF-SVM: Shows resilience with a 91.56% accuracy, emphasizing consistent performance. DT-NBK: Despite lower sensitivity, it achieves a notable accuracy of 92.45DT-SVM: Stands out with an accuracy of 94.34%, showcasing balanced performance metrics. NBK-SVM: Achieves a competitive accuracy of 91.45%, emphasizing overall effectiveness.

Statlog Dataset: RF-SVM achieves the highest accuracy of 92.62%, showcasing the effectiveness of combining Random Forest and Support Vector Machines for this dataset. Notably, all hybrid classifiers outperform their single algorithm counterparts, emphasizing the advantages of leveraging multiple algorithms. RF-DT: Maintains strong performance with an accuracy of 92.21%, demonstrating balanced precision and sensitivity. RF-NBK: Achieves a commendable accuracy of 93.54%, with balanced precision and sensitivity. RF-SVM: Demonstrates effective discrimination with an AUC of 0.93 and high accuracy (92.62%). DT-NBK: Despite lower sensitivity, it attains an accuracy of 93.26%. DT-SVM: Performs consistently well, achieving an accuracy of 91.64%. NBK-SVM: Yields a competitive accuracy of 92.79%, showcasing balanced precision and sensitivity.

The hybrid classifiers consistently outperform individual classifiers across all datasets. Combining Random Forest with other algorithms proves effective in various scenarios,

Table 3: Top three Sensitivity(%)

Data set types	Hybrid Classification Types	Number of Features	Sensitivity (%)
Z-Alizadeh Sani	DT-SVM	20	93.23
Cleveland	RF-SVM	24	93.23
Statlog	RF-SVM	9	94.21

Table 4: Top three Precision(%)

Data set types	Hybrid Classification Types	Number of Features	Precision(%)
Z-Alizadeh Sani	DT-SVM	20	94.45
Statlog	RF-SVM	9	94.35
Statlog	NBK-SVM	9	93.79

Table 5: Top three Accuracy (%)

Data set types	Hybrid Classification Types	Number of Features	Accuracy (%)
Z-Alizadeh Sani	RF-DT	26	95.45
Cleveland	DT-NBK	21	95.67
Cleveland	DT-SVM	21	94.34

with SVM consistently contributing to solid performance. The comprehensive evaluation highlights the efficacy of hybrid machine learning classifiers in handling diverse datasets. The study emphasizes the importance of selecting hybrid classifiers based on the specific characteristics of the dataset. These findings contribute valuable insights for practitioners and researchers seeking optimal classification models for different applications. Across all datasets, hybrid classifiers consistently outperform individual classifiers, highlighting the efficacy of combining multiple algorithms. The choice of the best-performing classifier varies based on the dataset, showcasing the importance of adapting models to specific data characteristics. RF-SVM emerges as a robust performer across multiple datasets, emphasizing its versatility and discrimination ability. The hybrid classifiers consistently demonstrate balanced performance metrics, showcasing their suitability for diverse applications. In conclusion, the comparative analysis underscores the adaptability and effectiveness of hybrid machine learning classifiers across diverse datasets. The findings provide valuable insights for practitioners selecting classifiers tailored to specific dataset types and characteristics. Analyzing the hybrid classification models across different datasets from Table 3, Table 4, and Table 5, we find noteworthy performance metrics for the Z-Alizadeh Sani dataset. Specifically, the DT-SVM model exhibits outstanding performance with a sensitivity (%) of 93.23% and a precision (%) of 94.45%. These values indicate the model's ability to identify positive instances and minimize false positives correctly. However, from an accuracy perspective, the hybrid classification model RF-DT stands out as the best performer on the Z-Alizadeh Sani dataset,

achieving an impressive accuracy of 95.45%. This suggests that the RF-DT model excels in overall predictive accuracy, considering both positives and negatives. In the Z-Alizadeh Sani dataset, the DT-SVM model is a standout choice if the focus is on achieving a balance between sensitivity and precision. On the other hand, if the primary concern is overall predictive accuracy, the RF-DT hybrid classification model proves to be the top performer with an accuracy of 95.45%.

6 Conclusion

Integrating the Internet of Things (IoT), advanced communication technologies, and artificial intelligence and machine learning presents immense transformative potential within the healthcare sector. The focus on predictive analytics, particularly for early disease prediction, signifies a pivotal shift from reactive to proactive healthcare methodologies. The proposed system, which harnesses data from IoT devices and electronic clinical records stored in the cloud, implements a two-phase approach involving hybrid feature selection and classification through machine learning techniques. The first phase underscores the significance of feature selection, employing a combination of filter-based methods (such as Chi-square, FCBF, GI, and RelieF) and wrapper-based methods (including BFE, EFS, FFS, and RFE). This meticulous selection aims to optimize disease prediction accuracy and efficiency by identifying pertinent features from the extensive dataset. The second phase entails implementing various machine learning techniques for hybrid classification, including Random Forest, Decision Tree, Naive Bayes, K-nearest neighbours, and Support Vector Machine. Utilizing the probabilistic ensemble voting method for heart disease prediction demonstrates the system's adaptability in addressing intricate medical scenarios. Despite the impressive performance of the DT-SVM model in Sensitivity and Precision, the RF-DT hybrid classification model emerges as the frontrunner from an accuracy standpoint on the Z-Alizadeh Sani dataset, attaining a remarkable accuracy of 95.45%. This highlights the RF-DT model's superior ability to balance true positives and negatives, making it an optimal choice for applications where predictive accuracy is paramount. The Z-Alizadeh Sani dataset's choice between the DT-SVM and RF-DT models depends on specific priorities. Opting for the DT-SVM model would be advantageous when seeking a harmonious balance between sensitivity and precision. Conversely, if the primary objective is to maximize overall predictive accuracy, the RF-DT hybrid classification model stands out as the top-performing choice, with an impressive accuracy of 95.45%. The proposed system epitomizes a holistic and advanced approach to healthcare, leveraging the synergies of IoT, cloud computing, artificial intelligence, and machine learning. By embracing proactive measures through predictive analytics, this system has the potential to revolutionize healthcare practices, offering timely and precise predictions for enhanced patient outcomes and an overall enhancement in public health. For future work, it would be beneficial to explore the scalability and interoperability of the proposed system across diverse healthcare settings. Additionally, investigating the integration of real-time data streams and continuous model learning could enhance the system's predictive capabilities. Moreover, conducting longitudinal studies to assess the long-term efficacy and impact of the system on healthcare outcomes would provide valuable insights for its refinement and broader adoption.

Declarations

Ethical Approval and Consent to participate

Not Applicable

Human and Animal Ethics

Not Applicable

Consent for publication

Not Applicable

Availability of supporting data

1. CNN / Daily Mail summarization dataset
   dataset link :https://github.com/abisee/cnn-dailymail

2. DUC 2002 single document summarization dataset
   dataset link :https://www-nlpir.nist.gov/projects/duc/guidelines/2002.html

Competing interests

The authors declare that they have no competing interests.

Funding

This paper does not get funds from any funding agency, institute, or company

Author's Contribution

All authors equally contributed to the studies given in the manuscript. All authors read and approved the final manuscript."

Acknowledgment

I want to express my sincere gratitude to the National Institute of Technology Tiruchirapalli for providing access to the DGX and Param Porul supercomputing facility. This facility has been instrumental in carrying out my research work, and its availability has been an invaluable resource to me.

References

[1] Zaini, N.A.M., Awang, M.K.: Hybrid feature selection algorithm and ensemble stacking for heart disease prediction. International Journal of Advanced Computer Science and Applications **14**(2) (2023)

[2] Deenathayalan, D., Narayanan, B.: Predicting heart disease using ftgm-pca based informative entropy based-random forest. CURRENT APPLIED SCIENCE AND TECHNOLOGY, 10 (2023)

[3] Zhu, Y., Li, W., Li, T.: A hybrid artificial immune optimization for high-dimensional feature selection. Knowledge-Based Systems **260**, 110111 (2023)

[4] Li, X., Zhang, J., Safara, F.: Improving the accuracy of diabetes diagnosis applications through a hybrid feature selection algorithm. Neural processing letters **55**(1), 153–169 (2023)

[5] Wei, G., Zhao, J., Feng, Y., He, A., Yu, J.: A novel hybrid feature selection method based on dynamic feature importance. Applied Soft Computing **93**, 106337 (2020)

[6] Zhang, S., Khattak, A., Matara, C.M., Hussain, A., Farooq, A.: Hybrid feature selection-based machine learning classification system for the prediction of injury severity in single and multiple-vehicle accidents. PLoS one **17**(2), 0262941 (2022)

[7] Sadeghian, Z., Akbari, E., Nematzadeh, H.: A hybrid feature selection method based on information theory and binary butterfly optimization algorithm. Engineering Applications of Artificial Intelligence **97**, 104079 (2021)

[8] Nasarian, E., Abdar, M., Fahami, M.A., Alizadehsani, R., Hussain, S., Basiri, M.E., Zomorodi-Moghadam, M., Zhou, X., Pławiak, P., Acharya, U.R., *et al.*: Association between work-related features and coronary artery disease: A heterogeneous hybrid feature selection integrated with balancing approach. Pattern Recognition Letters **133**, 33–40 (2020)

[9] Ramkumar, G., Seetha, J., Priyadarshini, R., Gopila, M., Saranya, G.: Iot-based patient monitoring system for predicting heart disease using deep learning. Measurement, 113235 (2023)

[10] Akhbarifar, S., Javadi, H.H.S., Rahmani, A.M., Hosseinzadeh, M.: A secure remote health monitoring model for early disease diagnosis in cloud-based iot environment. Personal and Ubiquitous Computing **27**(3), 697–713 (2023)

[11] Kolukisa, B., Bakir-Gungor, B.: Ensemble feature selection and classification methods for machine learning-based coronary artery disease diagnosis. Computer Standards & Interfaces **84**, 103706 (2023)

[12] Saranya, G., Pravin, A.: A novel feature selection approach with integrated feature sensitivity and feature correlation for improved prediction of heart disease. Journal of Ambient Intelligence and Humanized Computing **14**(9), 12005–12019 (2023)

[13] Senthil, R., Narayanan, B., Velmurugan, K.: Develop the hybrid adadelta stochastic gradient classifier with optimized feature selection algorithm to predict the heart disease at earlier stage. Measurement: Sensors **25**, 100602 (2023)

[14] Biswas, N., Ali, M.M., Rahaman, M.A., Islam, M., Mia, M.R., Azam, S., Ahmed, K., Bui, F.M., Al-Zahrani, F.A., Moni, M.A., et al.: Machine learning-based model to predict heart disease in early stage employing different feature selection techniques. BioMed Research International **2023** (2023)

[15] Mishra, S., Thakkar, H.K., Singh, P., Sharma, G.: A decisive metaheuristic attribute selector enabled combined unsupervised-supervised model for chronic disease risk assessment. Computational Intelligence and Neuroscience **2022** (2022)

[16] Bhatt, C.M., Patel, P., Ghetia, T., Mazzeo, P.L.: Effective heart disease prediction using machine learning techniques. Algorithms **16**(2), 88 (2023)

[17] Absar, N., Das, E.K., Shoma, S.N., Khandaker, M.U., Miraz, M.H., Faruque, M., Tamam, N., Sulieman, A., Pathan, R.K.: The efficacy of machine-learning-supported smart system for heart disease prediction. In: Healthcare, vol. 10, p. 1137 (2022). MDPI

[18] Ahmad, G.N., Ullah, S., Algethami, A., Fatima, H., Akhter, S.M.H.: Comparative study of optimum medical diagnosis of human heart disease using machine learning technique with and without sequential feature selection. ieee access **10**, 23808–23828 (2022)

[19] Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J.-J., Sandhu, S., Guppy, K.H., Lee, S., Froelicher, V.: International application of a new probability algorithm for the diagnosis of coronary artery disease. The American journal of cardiology **64**(5), 304–310 (1989)

[20] Alizadehsani, R., Habibi, J., Hosseini, M.J., Mashayekhi, H., Boghrati, R., Ghandeharioun, A., Bahadorian, B., Sani, Z.A.: A data mining approach for diagnosis of coronary artery disease. Computer methods and programs in biomedicine **111**(1), 52–61 (2013)

[21] Karaboga, D., et al.: An idea based on honey bee swarm for numerical optimization. Technical report, Technical report-tr06, Erciyes university, engineering faculty, computer . . . (2005)

[22] Kolukisa, B.: Development of data mining methodologies and machine learning models to understand cardiovascular disease mechanisms. Master's thesis, Abdullah Gül Üniversitesi, Fen Bilimleri Enstitüsü (2020)

[23] Chandrashekar, G., Sahin, F.: A survey on feature selection methods. Computers & Electrical Engineering **40**(1), 16–28 (2014)

[24] Hira, Z.M., Gillies, D.F.: A review of feature selection and feature extraction methods applied on microarray data. Advances in bioinformatics **2015** (2015)

[25] Tsai, M.-J., Wang, C.-S., Liu, J., Yin, J.-S.: Using decision fusion of feature selec- tion in digital forensics for camera source model identification. Computer Standards & Interfaces **34**(3), 292–304 (2012)