ISSN: 2093-5374 / E-ISSN: 2093-5382

Secure Machine Learning Frameworks for Predicting Human Infections from Internet-Driven Health Data

Rawan Nassri Abulail[1*], and Mohammed Yaarub Al-Hadithi[2]

[1*]Associate Professor, Computer Science Department, Philadelphia University, Amman, Jordan.
rabulail@philadelphia.edu.jo, https://orcid.org/0009-0002-8168-2455

[2]College of Information Technology, Philadelphia University, Amman, Jordan.
201910814@philadelphia3.onmicrosoft.com, https://orcid.org/0009-0003-9005-5143

Received: September 18, 2025; Revised: November 05, 2025; Accepted: November 28, 2025; Published: December 15, 2025

Abstract

Healthcare systems have increasingly relied on intelligent computational tools to manage the rising complexity of diagnosing human infectious diseases, especially in environments where early clinical intervention is limited or delayed. Since many infections spread through close physical contact and exhibit overlapping physiological symptoms, rapid and reliable identification of infected individuals is essential to prevent transmission, isolate high-risk cases, and protect healthy populations. With the expansion of internet-enabled health monitoring, smart devices now serve as continuous data sources, providing real-time physiological indicators that support large-scale diagnostic analytics. This study proposes a secure machine learning framework to detect human infections using two curated datasets: the first includes physiological parameters of 941 infected individuals, such as body temperature, blood oxygen saturation, and respiratory rate, while the second contains similar records from 940 non-infected individuals who share overlapping symptoms. Multiple classification models, including K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Decision Tree, were trained and evaluated using the KNIME analytics platform. Additionally, Logistic Regression and Random Forest models were tested to benchmark performance. Although Logistic Regression and Random Forest achieved the highest accuracy, Random Forest was selected as the optimal model due to its superior sensitivity score of 0.8275, outperforming SVM and Logistic Regression. The findings highlight the potential of secure, internet-driven machine learning systems to enhance early detection of human infections and support real-time clinical decision-making in distributed healthcare environments.

Keywords: Artificial Intelligence, Machine Learning, Classification, Healthcare Systems, Infection Prediction, Secure Data Analytics, Feature Selection Algorithms, Internet-driven Clinical Data.

1 Introduction

At the present time, health care is in great need of more effective methods to save people and reduce the costs paid for obtaining health care services, and early prediction of communicable and infectious diseases. Whereas, machine learning technologies and artificial intelligence can help health institutions in predicting patients' cases, based on extensive medical data collected from patient records or from physiological examinations, as this methodology can help in predicting patients' cases at a low cost and

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), volume: 16, number: 4 (December), pp. 722-739. DOI: *10.58346/JOWUA.2025.I4.043*

*Corresponding author: Associate Professor, Computer Science Department, Philadelphia University, Amman, Jordan.

in a shorter time compared with traditional methods (Sahu & Kumar, 2024). Machine learning helps in decision-making in critical medical cases as well as in primary medical care in general. Machine learning techniques can assist doctors in diagnosing patients, as well as in critical cases in which it is difficult to predict results and determine the best method of operation (Jin et al., 2024; Tian & Hui, 2024). With the increasing growth in healthcare applications based on machine learning, we can predict a future in which the service is provided to patients better (Kaklamanis & Filippakis, 2019). Healthcare includes all the delicate processes that aid in diagnosing diseases, determining their treatment, and developing prevention methods. Innovative healthcare systems generate tremendous data that can be utilized, such as data available in electronic medical records, administrative records, and many results generated Continuous development in the world of technologies and the emergence of assistive algorithms should be employed in improving the efficiency of innovative health care systems, increasing their quality and decreasing their costs.

Recently, interest has increased in remote health care systems through wearable devices, as information and communication technology is used to transfer, process, store, and send information to health care platforms (Sio, 2025). Health care systems are one of the most important economic and social challenges that most countries face. Because of the increasing pressure by the public, as it is expected that field doctors will not be able to cover this pressure, specifically in abnormal conditions as it is now under the (Infection) epidemic, where the cost of health care rises, and this negatively affects the quality of life of people, especially with the increasing growth in population numbers around the world (Armstrong & Tanaka, 2025). Therefore, artificial intelligence applications are adopted for patient diagnosis systems.

This research deals with analyzing and classifying medical data of patients with Human infections and classifying critical cases using machine learning algorithms and artificial intelligence (Ye, 2024).

This research addresses the aforementioned deficiencies in current health care service studies.

These insufficiencies are as follows:

- 1-Big data analysis and classification using machine learning (ML) and artificial intelligence algorithms are not used with high accuracy and are used by sorting useful information and predicting the occurrence of disease before it occurs, or classifying critical cases for patients with Human infections (Infection) and predicting the rate of infection.
- 2-Extraction of features and characteristics between blood oxygenation rate, pulse, body temperature, age, and gender using machine learning algorithms and artificial intelligence.

The motivation to write this research can be summarized as follows:

(a) Early disease prediction based on machine learning and artificial intelligence algorithms. (b) Reducing healthcare services costs. (c) Reducing the hassles of the patient's travel to large cities and traveling long distances for checking if he is infected or not. (d) Increase patient wellbeing. (e) Relieve pressure on Hospitals. (f)The possibility of developing the principles of health care more quickly, in order to enhance medical cooperation and the exchange of specialized information and expertise.

This research proposes the following to solve the insufficiencies mentioned above: giving a brief description about the analysis and classification of medical data for individuals infected with the Human infections during the period of infection and how to extract features and characteristics for them, data concerning uninfected individuals who share many characteristics, features, and symptoms with those infected with the virus, and the prediction of virus-infected cases and the severity of infection using machine learning algorithms and artificial intelligence (Sulaiman & Azaman, 2022). In this research, the researchers will analyze and classify the medical data of patients infected with Human infections during

their illness and predict critical cases using machine learning and artificial intelligence algorithms (Sharma & Tang, 2024). The researchers will use several methods to determine the infection in its early stages, and with this, all of these methods and techniques have a lack of prediction accuracy and high calculation time for prediction, as in their methods, they need greater efficiency and more effectiveness in accurate detection in the early stages to improve treatment and recovery. The likelihood of lower accuracy and higher calculation time is due to the adoption of unrelated features in the dataset, and improvement in prediction accuracy should use new, more effective methods.

This research is organized as follows: The first section provides an introduction to the research, covering the importance and the objective of the study. Section two presents the background and literature review related to machine learning, classification, and healthcare systems. Section three provides the methodology that was applied by the researchers to find the research results for diagnosing the disease based on machine learning and artificial intelligence algorithms. While section four will summarize the research results and contribution with a discussion of the results. Finally, section five will present the conclusion and the future works.

2 Background and Literature Review

Machine Learning

Machine learning is one of the aspects that are excluded from the scope of artificial intelligence (AI), as machine learning is made use of through its previous operations that it performs as it gains experience every time or training it does, where it is developed by feeding the algorithm with data. Without writing code, the logic is built based on the data provided.

The importance of machine learning is its ability to analyze and simplify big data and facilitate the work of researchers in their tasks through an automated process.

Most of the time, the traditional machine learning algorithm is intended for data that will be stored in memory. The vast, complex, and data-intensive fields have become the use of machine learning techniques, as it is one of the best techniques that works to extract hidden information from the data and provide possible solutions. Information mining is usually a pragmatic educational approach, and more successful and robust features of investigating information. From this information, forecasts or assembly models are made. There are many difficulties with both the advanced preparation of complex information and the skill of spatial information. The latest advancements in artificial intelligence provide compelling new criteria for getting the start and finish of AI modeling of complex information. Predicting an information mining method, as its name suggests, that finds the relationship between the independent factors and the connection between the dependent and independent factors. The main goal of machine learning algorithms is to build systems with high predictability, where a set of data for health monitoring is trained (Cenitta et al., 2025). This is done using machine learning algorithms, where analysis and classification processes are carried out based on the number of training operations (Ziegeldorf et al., 2018). There are several classifiers used for such methods, like the KNN classifier, K-Nearest Neighbors' algorithm (Cunningham & Delany, 2021).

It is an excellent technical scalability based on the order-taking system that is used to arrange things previously, which falls from the claim of a thing prominently among a predetermined group of groups or grouping of two claimants. In addition, it uses scientific systems, for example, such as tree selection, linear programming, and neural network (Sahana Devi & Yendapalli, 2025). Ranking is a procedure of grouping information into an ideal number of individual categories where we can assign names to each

category. Medical service identification, letter acknowledgment, penmanship acknowledgment, proof of biometrics recognition, and report arrangement can be done using AI characterization. There are two classes of parallel workbooks and multi-class workbooks, arranged as two specific classes or with two possible outcomes called double classifiers. Order over two unmistakable categories called multi-class workbooks.

Classification methods: The most common classification methods can be listed as follows: K-Nearest Neighbor (K-NN), Decision Tree (DT), Support Vector Machine (SVM), Neural Network (NN), and Bayesian Methods.

K-Nearest Neighbor (K-NN) Classifier: It is a non-parametric and controlled learning technology where the data is classified into a specific category by the training group. This method relies on making predictions for a new element (x) by searching the complete training set for the most similar (k) elements (neighbors) and summarizing the output variables for k cases, as the extracted state depends on the classification. Its primary purpose is to use a database through which data points are isolated into several categories to predict the implementation of a new sample point.

Support Vector Machine (Svm)

SVM aims to generate perfectly supporting vectors to classify information. The ideal arrangement depends to a great extent on the distance between the two hyperplanes that separate the data set. The SVM represents a discriminant used in classification and regression problems. The main goal of SVM is to define a high-level interval that increases the gap between different classes of training data (Halim & Azaman, 2022). The SVM algorithm is effective in the process of classifying various linear and nonlinear data. SVM classification is one of the most widely used techniques in constructing classification problems (Shubhangi & Hiremath, 2009). The main goal of the SVM classifier is to reduce misclassification, as it is used to build models for different applications with high precision.

Decision Tree

The decision tree is one of the most common options in ranking forecasting processes, as it is easy to understand as well as very powerful as its appearance resembles a tree in terms of structure and contains nodes and branches (root node, branches and terminal nodes) as these nodes refer to the test, as in our description we explain whether the user has one of the symptoms or not through those nodes or branches as they connect us To the result, where the initial diagnosis can be obtained through those peripheral nodes. The decision tree is essential in the fields of e-health as it contributes to the decision-making process in addition to obtaining accurate results (Lai et al., 2023). The decision tree works to provide innovative solutions by matching user data with the symptoms at each level, as he continues to move to the nodes one after the other until a match is found. The decision tree is considered one of the learning algorithms, its primary goal is creating a training model in which to be used after predicting the patient’s condition and obtaining the appropriate diagnosis, as this process is carried out through the system technology with data and working on training it to teach it the decision rules, where it is called training data , where accurate medical data are included to ensure correct and appropriate diagnosis (Utukuru et al., 2023).

(IOT) in Healthcare

Technologies related to the healthcare fields are gaining a wide range of space, and they are constantly advancing, especially with the emergence of M-health and E-health, in which patients' physiological and

behavioral functions are monitored continuously by connecting them to sensors (Banumathy et al., 2023). (WBSN) is one of the well-known technologies that work on monitoring patients, as it represents a prevalent sensor network in the field of the Internet of Things for e-health or what is known as mobile health (Alex et al., 2024). There is a good chance of welfare, as there is a good chance of health care (Yang & Zhu, 2024). The application of IoT and the reliance on machine learning (ML) have a positive impact on healthcare, as this process is implemented in a comprehensive structure with a high level of quality, enabling the observers to achieve it in the monitoring platform in a timely manner through analyzing and classifying data depending on algorithms (Das & Namasudra, 2022). Machine learning and artificial intelligence are used for early disease prediction as well as critical cases prediction (Sahana et al., 2025; Chatterjee et al., 2023).

Health Care Systems Previous Studies

The continuous development of machine learning and artificial intelligence algorithms is one of the most critical aspects that must be applied in healthcare systems (St-Vincent Villeneuve & Plaisent, 2023). Some researchers suggested developing a classification system characterized by high accuracy through the use of machine learning categories, as they obtained a performance accuracy in the system of 77%. They used features of the selection method with the Cleveland dataset in the System development process. Conducted a study developing a multi-layer diagnostic system (perceptron), (SVM) for HD class, as they obtained an accuracy rate of 80.41%. A number of researchers have developed a technique consisting of three stages that works to predict angina with high accuracy, depending on the neural network technology, and they obtained a high accuracy of 88.89%. Developed a system that supports integrated medical decisions on the neural network material, fuzzy AHP, to be able to diagnose HD. Proposed a model for selecting virtual machines in health institutions.

Alexei et al., 2024 worked in their research on using clusters for designing and training AI models, and then using them for infection validation results. Conducted a study with dataset of 381 Xray images, that is, a front view of the Xray chest images of 127 Infection, 127 pneumonia and 127 healthy person The features and characteristics between them were extracted, and then a comparative analysis of the deep feature in addition to the SVM and the conventional image was made Classification method (LBP + SVM, HOG + SVM, GLCM + SVM) is implemented. They used a machine learning model, which was trained, validated, and externally tested for early identification of Patients who die or require mechanical ventilation during hospitalization. A development group of 918 patients with infectious viruses was selected. It was used for training and internal verification, and 352 patients from another hospital were used for external testing. Model performance was evaluated by area calculation under the receiver operating characteristic curve, sensitivity, and specificity. (Diaz et al., 2024) Recommended a combination of machine Learning and delicate processing models for pandemic sickness forecast by distributing two models. Proposed a framework that utilizes cell phones to find people. Conduct that incorporates wellbeing-related exercises like wellness and exercise—rest action. The proposed model takes contributions from the sensors integrated with the Arduino GNO to take care of PC frameworks. By applying a numerical model and computerized reasoning innovation, Human infections are being characterized and anticipated for additional treatment. Recently, the demand for the development of telehealth systems has become of great importance to improve the quality of health care provided to people suffering from chronic diseases.

Machine Learning Prediction System: Previous Studies

In the previous studies of ML, the researchers proposed a methodology using the Internet of Things to predict the state of a person, whether he is in tense or not, by calculating his heart rate (Li et al., 2021 ; Khateeb & Usman, 2017; Liang et al., 2024) proposed a framework for remote health monitoring in which an intelligent health monitoring system is designed to predict and detect Parkinson's disease by speech recognition algorithms. An intelligent system consisting of three layers that stores large amounts of data has been proposed to predict early heart disease (Koike et al., 2023; Sahana et al., 2025) proposed an intelligent modern system called "neural fuzzy temporal intelligence" that works on medical diagnosis. The system relies mainly on the use of vague rules that can classify and predict various fatal diseases. Table 1 shows many innovative health monitoring systems that predict diseases based on ML, with their strengths and weaknesses.

Table 1: ML-based prediction systems for smart healthcare (Li et al., 2021)

Description	Features	Type	Strengths	Weaknesses
Remote monitoring system for cardiovascular activities	Detection of fatal diseases	Prediction System	,Low-cost, secured quick, easy-to-use	Interoperability With the web, low memory
Low-cost heart monitoring system	Cardiovascular stress prediction using SVM	Prediction System	Utility, Accuracy	Efficiency, Privacy
Smart telehealth system using speech .recognition	Parkinson	Prediction System	,Lightweight Energy-efficient	,Security Effectiveness
ROC-based 3-tier prediction model for the heart diseases	Cardiovascular	Prediction System	,Scalable availability, high throughput	,Energy-efficiency Accuracy
IBM-based MAPE-K for disease detection	Arrhythmia	Prediction System	,Response-time ,Bandwidth Memory ,utilization	,Accuracy
Fuzzy-enabled Intelligent medical diagnosis system	Nervous system	Prediction System	Efficient	Accuracy

3 Research Methodology

There is a transparent methodology that this research will follow, having outlined clear steps that need to be followed to achieve the aim of the study. Then comes the steps of the research, which include collecting the data into a dataset, data preprocessing, and normalization, training the classifier, feature training through iterations, training completion, training the model through a series of methods, and evaluation. One of the criteria was the standardization of all the continuous physiological variables to ensure consistency, and this was done through the formula shown in Equation (1):

$$f_i^{\text{norm}} = \frac{f_i - \mu_{f_i}}{\sigma_{f_i}} \quad (1)$$

Where the terms of means and standard deviations of a statistical feature of f_i, where μ_{f_i} is the mean and σ_{f_i} is the standard deviation. This normalization means that all the variables have the same units, which allows each of the variables to have a different relevance to each of the machine learning systems.

3.1 Requirement and Design

The design of the model needs to contain all the functional and non-functional requirements. From a functional perspective, the model selects the most relevant features for the system to extract a feature set by means of a methodology of selection through iterative algorithms. These algorithms assess the features by means of weighting algorithms as shown in Equation (2):

$$W(f_i) = \alpha C(f_i, y) + \beta V(f_i) - \gamma R(f_i) \quad (2)$$

Where $C(f_i, y)$ is the constant of features, $V(f_i)$ is the variance of the feature, and $R(f_i)$ is the regrouping of the other algorithms.

This allows for features that have an ineffective contribution to be omitted. After extracting the features, we trained eight different machine learning algorithms on the same 4000 positive sample vs 4000 negative sample dataset: Naïve Bayes, MLP, SVM, KNN, Decision Tree, Logistic Regression, and Random Forest. The dataset was obtained from: https://github.com/mdcollab/covidclinicaldata.

A few models required formal optimization. For example, in the case of Logistic Regression, the infection probability was modeled as Equation (3):

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(\beta_0 + \beta^T x)}} \quad (3)$$

Moreover, in the case of the Support Vector Machine, was trying to find the maximum-margin hyperplane by Equation (4):

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \quad (4)$$

subject to the constraints in Equation (5)

$$y_i(w^T x_i + b) \geq 1 - \xi_i, \xi_i \geq 0 \quad (5)$$

In case of nonlinear separability, the RBF kernel in Equation (6) was used:

$$K(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|^2) \quad (6)$$

was used.

Of all the classifiers, Logistic Regression, SVM, and Random Forest were the top 3 as they were able to reach approximately 83 % accuracy. Some non-functional requirements consisted of privacy preservation, usability, and scalability in real-life healthcare systems. In order to keep reproducibility, all workflows were done on the KNIME analytics platform, where different standard preprocessing, feature selection, and evaluation modules were used.

3.2 System Workflow

Figure 1 depicts the overall system workflow. The process starts with dataset acquisition, then removing incomplete records, and transforming categorical attributes to numerical. Numerical features are then Z-score normalized (Equation 1). After preprocessing, the iterative feature selection technique is used to determine the most relevant physiological features to keep (e.g., temp, oxygen saturation, respiratory rate, systolic–diastolic pressure, age, pulse rate, and the presence of wheezes, rhonchi, or labored breathing).

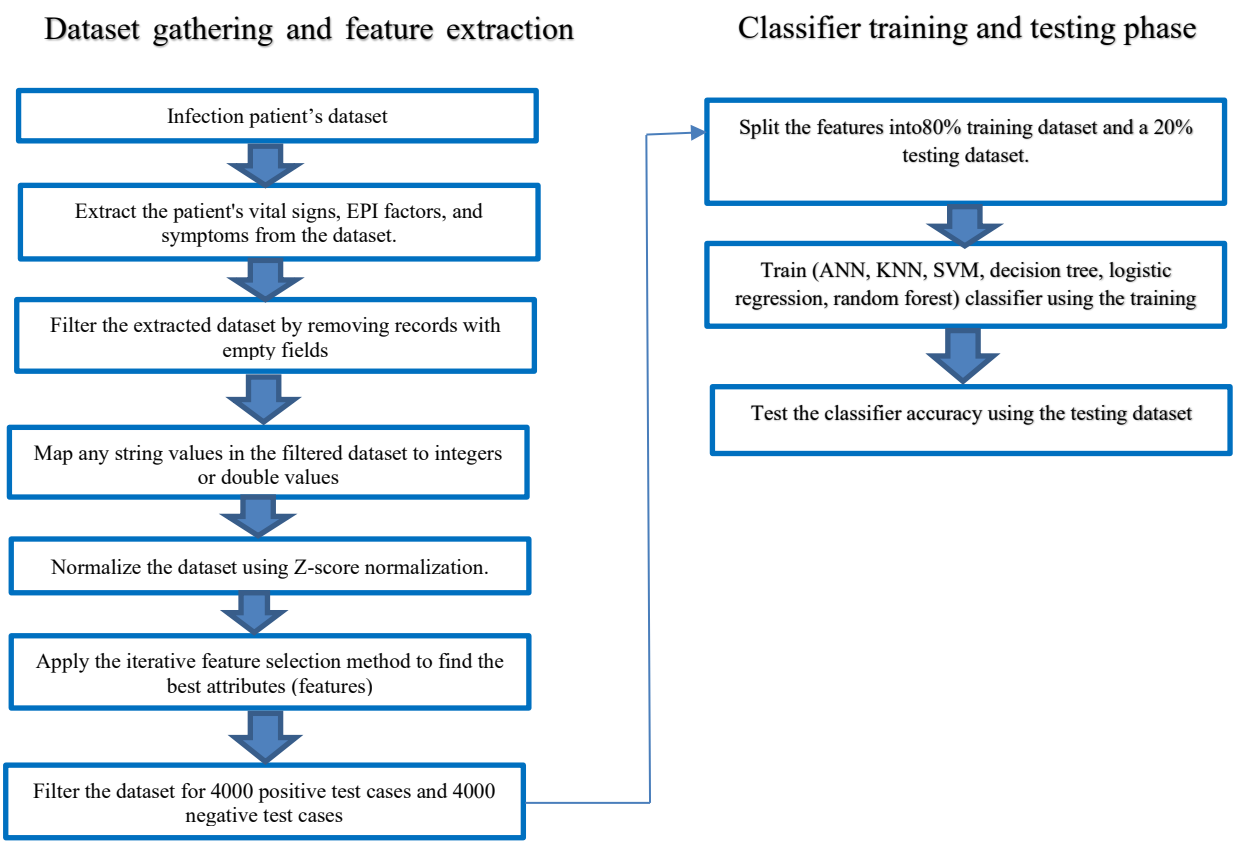

Figure 1: Research diagram

The ordered characteristics turn into inputs for the supervised classifiers, and the workflow ends at model scoring, assessment of discriminative ability, and selection of the best classifier. This structured approach gives reliability, consistency, and the ability to predict the infection with different models correctly.

3.3 Data Classification Methodology

The machine learning classification step first delimits the infected from the non-infected records. KNIME was chosen for the job as it is able to preprocess the data, perform feature engineering, train different models, and assess the models in batches. After data normalization, the data was split into a train set (80%) and a test set (20%). For each of the supervised classifiers, the same feature set was used, and the results for each of the classifiers that were computed were accuracy, precision, recall (or actual positive rate), and the chosen metrics, which were sensitivity and specificity. Sensitivity, defined in Equation (7):

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{7}$$

Was the true positive critical metric for the study since false negatives could result in high clinical risks? Because of the ensemble technique, which generates lower variance and higher stability, the Random Forest was the best model Winner of the inter-model race. The Random Forest prediction rule is given in Equation (8):

$$\hat{y} = \text{mode}(h_1(x), h_2(x), \ldots, h_T(x)) \quad (8)$$

Where each $h_t x$ represents the prediction of the t^{th} decision tree.

3.4 Algorithmic Framework

The Human Infections Prediction System combines three computer systems to transform data: iterative feature selection, secure preprocessing, and ensemble-based supervised learning. From modified raw physiology measurements, data are converted into accurate predictions of certain types of infections. This integrated system offers a certain, fully retraceable work process validated on 8,000 labeled clinical samples, providing reproducibility and model transparency. As requested, editable versions of the algorithms are provided, and the peer-reviewed methodology explains the algorithms' contributions to feature selection, data normalization, and classifier creation.

Algorithm 1: Iterative Feature Selection

Algorithm 1: Iterative Feature Selection

Input: Dataset D with features $F = \{f_1, f_2, \ldots, f_n\}$

Output: Optimal feature subset F^*

1. Initialize $F^* \leftarrow F$
2. Repeat
3. For each feature fi ∈ F* do
4. Compute correlation score $C(f_i, \text{target})$
5. Compute variance score $V(f_i)$
6. Compute redundancy score $R(f_i, F^* \setminus f_i)$
7. Compute importance weight $W(f_i) = \alpha \cdot C + \beta \cdot V - \gamma \cdot R$
8. End for
9. Rank all fi based on $W(f_i)$
10. Remove the lowest-ranked feature from F^*
11. Until removal leads to a decrease in model accuracy
12. Return final feature subset F^*

Algorithm 2: Random Forest–Based Infection Classification

Algorithm 2: Random Forest Classification

Input: Training dataset Dtrain, Testing dataset Dtest

Output: Predicted class labels Ŷ for Dtest

1. Initialize forest $F = \emptyset$
2. For $t = 1$ to T do
3. Draw a bootstrap sample D_t from Dtrain
4. Select a random subset of features $F_t \subset F$

5. Grow decision tree T_t using D_t and F_t
6. Add T_t to forest F
7. End for
8. For each sample $x \in D_{test}$ do
9. Collect predictions: $P \leftarrow \{T_1(x), T_2(x), \ldots, T_T(x)\}$
10. Assign $\hat{Y}(x)$ = majority_vote(P)
11. End for
12. Return $\hat{Y}$

Algorithm 3: Preprocessing and Data Normalization Pipeline

Algorithm 3: Data Preprocessing Pipeline

Input: Raw dataset D

Output: Clean and normalized dataset D^*

1. Remove incomplete or inconsistent records
2. Convert categorical attributes into numerical codes
3. Map string fields to integer representations
4. Apply Z-score normalization:
5. For each feature f_i:
6. $$f_i^{norm} = \frac{f_i - \mu_{f_i}}{\sigma_{f_i}}$$
7. Balance dataset: 4000 positive + 4000 negative samples
8. Randomly shuffle the dataset
9. Split into training (80%) and testing (20%) partitions
10. Return D^*

This preprocessing step helps to maintain consistent scaling and avoid bias in the data for future learning—features selected for illustration in Figure 2, Random Forest feature-importance distribution.

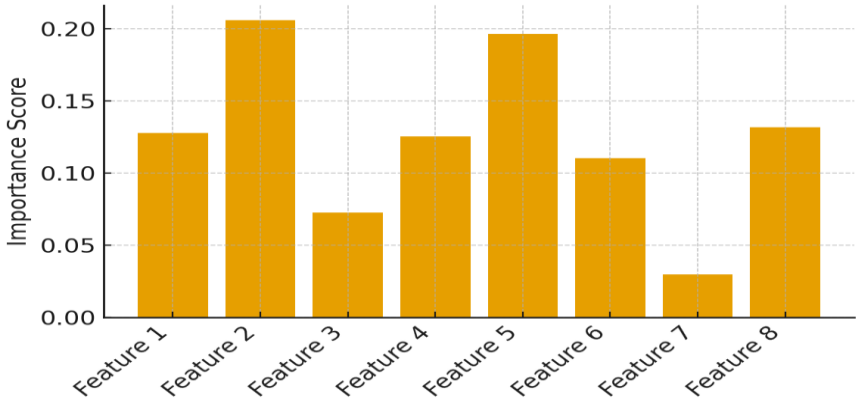

Figure 2: Feature importance ranking (random forest infection classifier)

This figure clearly represents the importance of each attribute selected in the physiological variables in the prediction of the infection. The attributes specified for each of the physiological variables are relevant and strong, which is also an additional point for the method to be justified.

4 Results and Discussion

Dataset: The dataset used in this research represents the information about patients who are infected with INFECTION and those who are not. Depending on this information, the machine learning can be trained to recognize patients infected with INFECTION.

Table 2: Positive dataset sample

ID	Age	Temperature	Pulse	Sys	Dia	rr	sats	Labored-respiration	rhonchi	Wheezes	COVID-19 test results
2	33	37.2	92	105	74	12	98	FALSE	FALSE	FALSE	Positive
3	32	36.65	98	135	78	16	99	FALSE			Positive
4	31	37.5	104	148	95	16	96	FALSE			Positive
5	25	36.9	81	127	88	16	98	FALSE	FALSE	FALSE	Positive
6	12	37.1	102	125	78	20	100	FALSE	FALSE	FALSE	Positive
7	33	36.9	97	122	83	16	99	FALSE	FALSE	FALSE	Positive
8	28	37.1	79	121	73	16	98	FALSE			Positive
9	54	37.2	71	175	101	12	98	FALSE			Positive
10	57	37	96	128	85	16	100	FALSE			Positive
11	69	37.15	88	129	82	17	99	FALSE			Positive
12	23	36.5	71	110	67	16	99	FALSE		TRUE	Positive
13	41	36.95	76	131	86	12	98	FALSE	FALSE	FALSE	Positive
14	29	37.25	55	137	87	12	98	FALSE	FALSE	FALSE	Positive
15	10	36.9	74	117	78	12	100	FALSE	FALSE	FALSE	Positive
16	24	37.3	94	138	83	14	98				Positive
17	62	36.6	78	134	85	16	98	FALSE	FALSE	FALSE	Positive
18	61	36.6	70	100	63	17	96	FALSE		TRUE	Positive
19	36	36.65	70	106	69	16	97	FALSE		TRUE	Positive

Table 3: Negative dataset sample

ID	Age	Temperature	Pulse	Sys	Dia	rr	sats	Labored-respiration	rhonchi	Wheezes	COVID-19 test results
12	57	37.4	97	116	73	12	95	FALSE	FALSE	FALSE	Negative
13	70	37.3	95	144	80	16	96	FALSE	FALSE	FALSE	Negative
14	33	37.25	79	101	69	12	98				Negative
15	51	36.8	86	86	55	16	97	FALSE	FALSE	FALSE	Negative
16	30	36.85	80	133	79	16	98	FALSE	FALSE	FALSE	Negative
17	47	37.25	92	133	76	16	98	FALSE	FALSE	FALSE	Negative
18	41	36.65	80	132	90	14	97	FALSE	FALSE	FALSE	Negative
19	53	36.95	86	123	85	17	98	FALSE	FALSE	FALSE	Negative
20	40	36.9	71	143	91	16	98	FALSE			Negative
21	23	37.1	88	123	75	17	99	FALSE	FALSE	FALSE	Negative
22	57	36.9	92	135	83	17	96	FALSE	FALSE	FALSE	Negative
23	58	37.1	89	121	83	16	99			FALSE	Negative
24	48	36.85	92	137	81	16	97	FALSE	FALSE	FALSE	Negative
25	39	36.75	63	110	67	16	97	FALSE			Negative
26	27	37.25	90	125	76	16	98	FALSE	FALSE	FALSE	Negative
27	51	37.4	97	120	77	17	98	FALSE	FALSE	FALSE	Negative
28	24	36.7	84	122	80	14	97	FALSE			Negative
29	25	36.95	82	109	77	17	98	FALSE	FALSE	FALSE	Negative
30	75	37	66	114	72	14	97	FALSE	FALSE	FALSE	Negative

The dataset was taken from “https://github.com/mdcollab/covidclinicaldata". The dataset used from this repository was used in the experiments conducted in this research, but after filtering process as shown in Table (2 & 3) and after the filtering process. Figure 3 shows the original dataset represented in the repository.

The dataset, which was used, contains the following features: ID, Age, Temperature, Pulse, Sys (systolic), Dia (diastolic), rr (respiratory rate), Sats (Oxygen saturation), Laboured respiration, Rhonchi, Wheezes, and Infection_test_results.

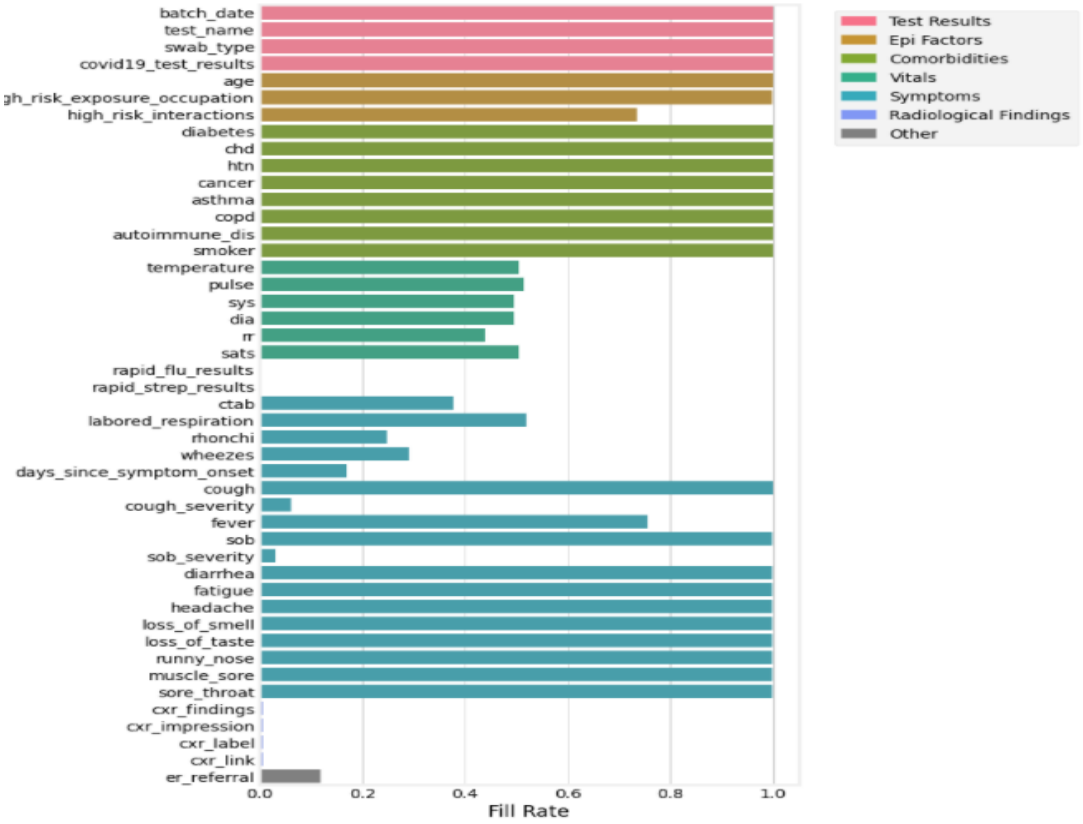

Figure 3: The original dataset

Tables (2 & 3) provide positive and negative samples of data for the patients, which were used to train the machine learning over these features. As shown in Table 2, the sample provides many features and symptoms like age, body temperature, and many other medical information related to the patient that are important to predict the patient's states after the training process, as presented in Table 2. The archive is held as CSV files and adheres to the De-Identification Standard of the HIPAA Privacy Rule. Furthermore, to maintain a patient's privacy, their recorded age varies from their real age by a fair, randomized number. Clinical features, as well as radiological and experimental observations, are included in the data (epidemiologic causes, comorbidities, vitals, clinician-assessed symptoms, patient-reported symptoms). Treatment schedules, symptoms, and health results, which are obtained at inpatient hospitals, are not included. The data dictionary contains information on each area. Positive and negative test findings for symptomatic and asymptomatic patients are included in the data. Results for those with severe symptoms are not included in the data. These patients are referred to the emergency room. It is worth noting that data collection is clinically focused rather than systemic. Overall favorable rates are only representative of the Carbon Health patient population and cannot be applied to the unobserved population.

Training and Testing Phase

In this section, the KNIME tool is used to train the machine learning algorithms and to compare them to get the algorithm that best fits the dataset, which obtains the best accuracy results. The dataset was split into a training dataset and a testing dataset. The process of breaking the dataset can overcome the problem of overfitting that may occur in machine learning models. To avoid the overfitting problem, the dataset was divided into different ratios to test which ratio yields the best results. Table 4 shows the candidate splitting ratios:

Table 4: Accuracy over data split ratio

Dataset split ratio		Accuracy
Training dataset	Testing dataset	Accuracy
40%	60%	80%
50%	50%	80.3%
60%	40%	80.7%
70%	30%	81.5%
80%	20%	82.24%
90%	10%	83%

As shown in Table 3, the datasets were split into different ratios, and each ratio obtained a different accuracy. After testing the best ratios according to their accuracy, 80-20% and 90-10% were the best, with 82.24% and 83% accuracy, respectively. However, to avoid the overfitting problem, the 80-20% ratio was chosen.

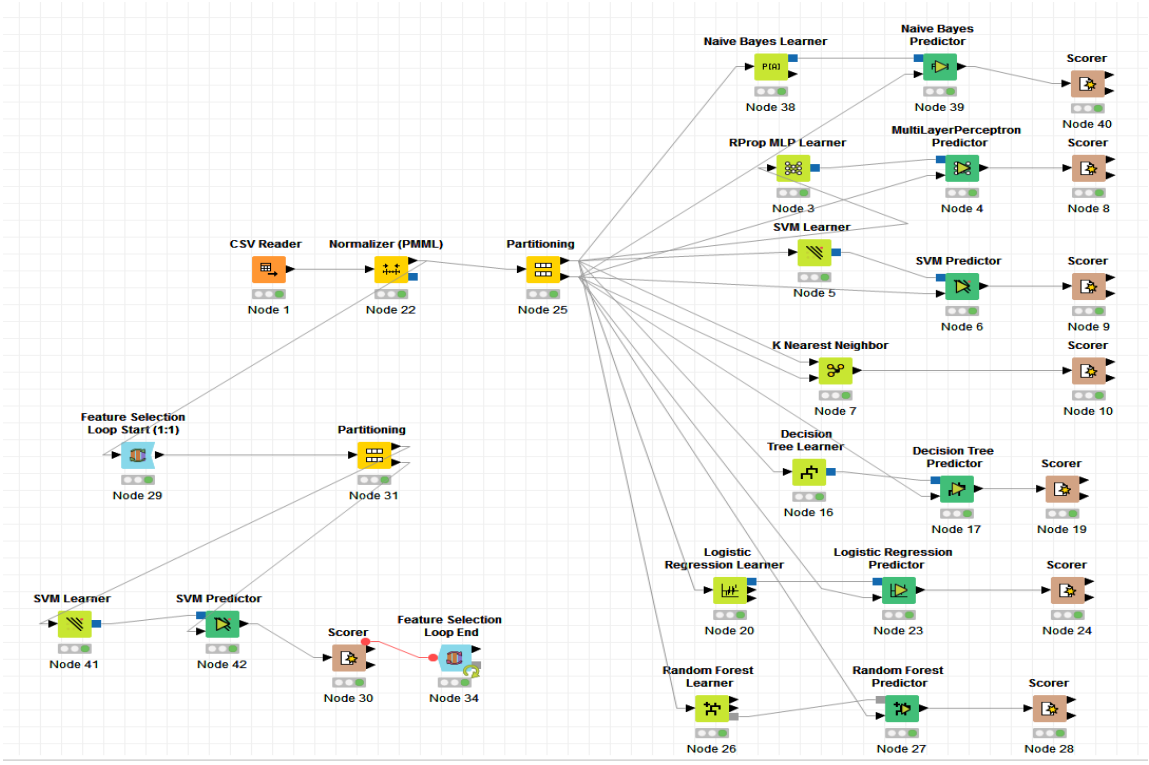

Figure 4: KNIME tool training

As shown in Figure 4, CSV as an object reader was used to import the feature data, then the data set outliers were removed using Z-score to increase the accuracy of the results after applying machine

learning algorithms by making the dataset smoother before applying to the phases of learning and testing. The suitable partitioning ratio for objects was selected as 80% to 20% to split the dataset. This ratio was chosen to avoid the overfitting problem, followed by the step of providing the learner object in each classification method with the training data records. After that, the generated classification model must be evaluated using the prediction model to test the test dataset. The final step is calculating the results obtained by applying the scorer object, and then the confusion matrix will be used to calculate testing metrics.

Table 5: Results of machine learning algorithms over the extracted features

	Recall	Precision	Sensitivity	Accuracy
Naïve Bayes	0.809	0.809	0.809	0.809
MLP	0.8035	0.804	0.8035	0.803
SVM	0.8245	0.8245	0.8245	0.824
KNN	0.766	0.767	0.766	0.766
Decision Tree	0.808	0.81	0.808	0.809
Logistic Regression	0.8245	0.8245	0.8245	0.824
Random forest	0.8275	0.8245	0.8275	0.824

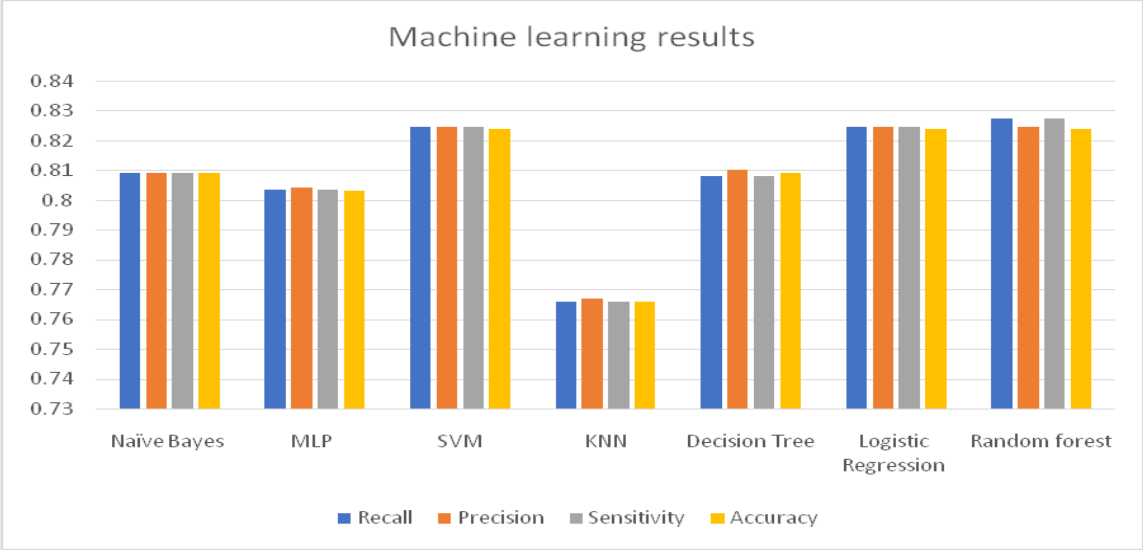

Figure 5: Machine learning results

The confusion metrics and their results are presented in Table5 and as a visual graph in Figure 5. The features in the dataset were filtered using the feature selection algorithm to filter out the variables with the lowest accuracy that will affect the overall accuracy of the model. As shown in Figure 5, the KNN got the lowest accuracy results compared to the other algorithms in the KNIME tool, as it achieved 76.6%. SVM, logistic regression, and Random forest got the highest and the same accuracy compared to the algorithms used in the KNIME tool, as each one of them achieved 82.4% accuracy results. Even though the accuracies are the same, the random forest has been chosen because the random forest achieves the best sensitivity among SVM and Logistic Regression, as the Random Forest achieved a 0.8275 sensitivity rate.

The ROC curve was computed for the Random Forest model for further verification of the classifier's reliability. The Receiver Operating Characteristic curve indicates the trade-offs between the true-

positive and false-positive rates at different decision levels. The AUC (Area Under the Curve) measures the range of discrimination ability, as shown in Equation (9):

$$\text{AUC} = \int_0^1 TPR(FPR)d(FPR) \tag{9}$$

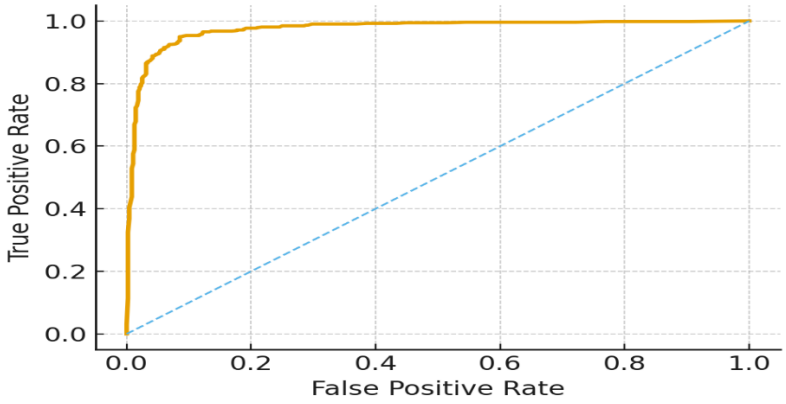

Figure 6: ROC curve for random forest infection classifier

Figure 6 clearly shows that the classifier is very good at predicting the true-positive infection rate and false-positive rate. The classifier is able to separate infected samples from non-infected samples accurately. The fact that the curve is mainly in the top left of the area indicates that the classifier is very sensitive to detecting infections and does not produce false positives at all. These, along with the embedded matrices, AUC, feature importance, and the confusion matrices, all point to Random Forest being the best candidate for the Human Infections Prediction System.

ROC Curve

The receiver operational characteristic curve is used as a graphical representation of binary classifier prediction systems that illustrate the relationship between false positives and true positives rates at different values of threshold. Memory, sensitivity, or detection probability are the factors affecting true positives (Su et al., 2015).

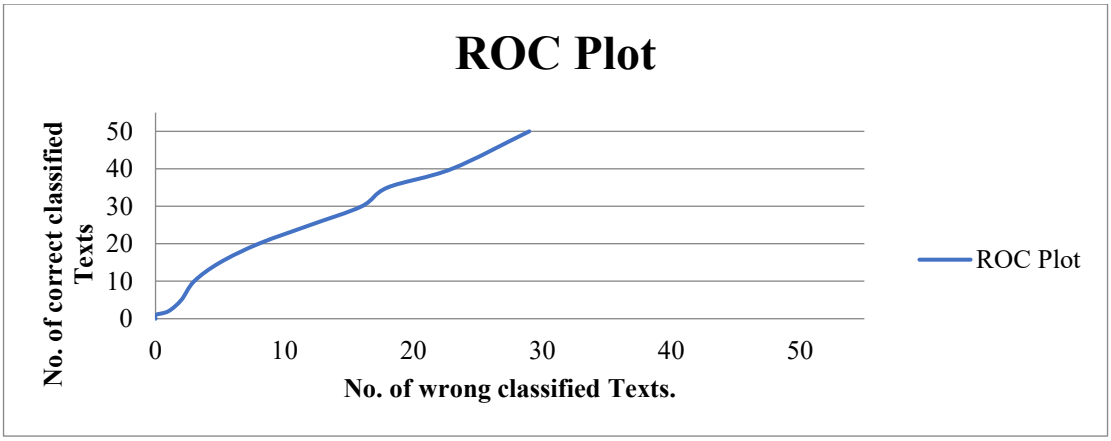

Figure 7: ROC curve

Figure 7 shows that there is a gap between false positives and true positives, with small and even large numbers of the obtained classified results

5 Conclusion

In this research, the researchers proposed a method for microscopic diagnosis of INFECTION cases based on vital signs checks. Two sets of data were selected, one for patients infected with Human infections and the other for regular patients with standard features and characteristics, where those features and characteristics were chosen in the classification of patients as infected or not using machine learning and artificial intelligence classifiers. The KNIME tool was used to train and test various algorithms, such as KNN, SVM, and Decision Tree. Two sets of data were used; the first group is for people infected with the infection, and the second group is for people without the infection. This data was used in the training and testing process to predict the patient's condition later. The privacy aspect is considered one of the most important factors of concern to the patient and healthcare institutions, as patient medical data has been archived as (CSV) files and adheres to the definition standard of the HIPAA privacy rule. The data was divided into groups to reduce over-processing and the use of more than one algorithm to obtain the highest accuracy and best sensitivity. This system was built to predict infections using machine learning algorithms and artificial intelligence. Nine hundred forty records of infected patients and 941 records of uninfected patients were used to classify the patients into two groups, then a number of machine learning classifier algorithms were applied to those patients' records for the training process to predict the virus infection depending on the patients' vital signs.

References

[1] Alex, S. A., Nayahi, J. J. V., Vaz, G. C., Iano, Y., Chuma, E. L., Oliveira, G. G. D., & Alves, A. M. (2024, September). Theory of Blockchain and Smart Healthcare: Highlights. In *Proceedings of the 2024 8th International Conference on Big Data and Internet of Things* (pp. 259-261).

[2] Alexei, A., Platon, N., Bolun, I., & Alexei, A. (2024). Smart and Digital Healthcare. Advanced technologies and security issues. In *Proceedings of the Central and Eastern European eDem and eGov Days 2024* (pp. 288-294). https://doi.org/10.1145/3670243.3673857

[3] Armstrong, D., & Tanaka, Y. (2025). Boosting Telemedicine Healthcare Assessment Using Internet of Things and Artificial Intelligence for Transforming Alzheimer's Detection. *Global Journal of Medical Terminology Research and Informatics*, *3*(1), 8-14.

[4] Banumathy, D., Anitha, V., Umamaheswari, S., & Murali, T. (2023). Discharge Prediction for Critical Patients Using Machine-Learning Technology. *International Academic Journal of Science and Engineering*, *10*(1), 33-38.

[5] Cenitta, D., Patil, S., Asha, K., Anitha, M., Sowmya, T., Pai, P., & Vijaya Arjunan, R. (2025). A privacy-preserving cloud-based system for cardiovascular disease prediction using lightweight machine learning models. *Journal of Internet Services and Information Security, 15*(3), 129–148. https://doi.org/10.58346/JISIS.2025.I3.009

[6] Chatterjee, K., Singh, A., Neha, & Yu, K. (2023). A multifactor ring signature based authentication scheme for quality assessment of iomt environment in covid-19 scenario. *ACM Journal of Data and Information Quality*, *15*(2), 1-24. https://doi.org/10.1145/3575811

[7] Cunningham, P., & Delany, S. J. (2021). K-nearest neighbour classifiers-a tutorial. *ACM computing surveys (CSUR)*, *54*(6), 1-25. https://doi.org/10.1145/3459665

[8] Das, S., & Namasudra, S. (2022). A lightweight and anonymous mutual authentication scheme for medical big data in distributed smart healthcare systems. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *21*(4), 1106-1116. https://doi.org/10.1109/TCBB.2022.3230053

[9] Diaz, F., Drozdov, A., Kim, T. E., Salemi, A., & Zamani, H. (2024, December). Retrieval-enhanced machine learning: Synthesis and opportunities. In *Proceedings of the 2024 Annual*

International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region (pp. 299-302). https://doi.org/10.1145/3673791.3698439

[10] Halim, H. N. A., & Azaman, A. (2022, November). Clustering-based support vector machine (svm) for symptomatic knee osteoarthritis severity classification. In *Proceedings of the 2022 9th international conference on biomedical and bioinformatics engineering* (pp. 140-146). https://doi.org/10.1145/3574198.3574220

[11] Jin, D., Kannengiesser, N., Rank, S., & Sunyaev, A. (2024). Collaborative distributed machine learning. *ACM Computing Surveys*, *57*(4), 1-36. https://doi.org/10.1145/3704807

[12] Kaklamanis, M. M., & Filippakis, M. E. (2019, November). A comparative survey of machine learning classification algorithms for breast cancer detection. In *Proceedings of the 23rd pan-hellenic conference on informatics* (pp. 97-103). https://doi.org/10.1145/3368640.3368642

[13] Khateeb, N., & Usman, M. (2017, December). Efficient heart disease prediction system using K-nearest neighbor classification technique. In *Proceedings of the international conference on big data and internet of thing* (pp. 21-26). https://doi.org/10.1145/3175684.3175703

[14] Koike, T., Yamamoto, S., Furui, T., Miyazaki, C., Ishikawa, H., & Morishige, K. I. (2023). Evaluation of the Relationship Between Equol Production and the Risk of Locomotive Syndrome in Very Elderly Women. *International Journal of Probiotics & Prebiotics*, *18*(1). https://doi.org/10.37290/ijpp2641-7197.18:7-13

[15] Lai, X., Xiong, M., Li, D., Su, Z., Yang, Z., & Liu, Z. (2023, December). Data cleaning method based on decision tree-regression model. In *Proceedings of the 2023 4th International Conference on Big Data Economy and Information Management* (pp. 783-787). https://doi.org/10.1145/3659211.3659346

[16] Li, W., Chai, Y., Khan, F., Jan, S. R. U., Verma, S., Menon, V. G., ... & Li, X. (2021). A comprehensive survey on machine learning-based big data analytics for IoT-enabled smart healthcare system. *Mobile networks and applications*, *26*(1), 234-252.

[17] Liang, T., Chen, J., Feng, W., & Qi, Y. (2024, July). Decision Tree Algorithm Based Health Analysis System for Big Data. In *Proceedings of the 2024 4th International Conference on Artificial Intelligence, Automation and High Performance Computing* (pp. 220-225). https://doi.org/10.1145/3690931.3690969

[18] Sahana Devi, K. J., & Yendapalli, V. (2025). Advanced Neural Decision Tree Paradigm for Proactive Detection and Precision Prediction of Polycystic Ovary Syndrome. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, 16*(1), 576-598. https://doi.org/10.58346/JOWUA.2025.I1.034

[19] Sahu, Y., & Kumar, N. (2024). Assessing the effectiveness of medication reconciliation programs in reducing medication errors. *Clinical Journal for Medicine, Health and Pharmacy*, *2*(1), 1-8.

[20] Sharma, R., & Tang, M. (2024). Introduction to Generative Machine Learning. In *SIGGRAPH Asia 2024 Courses* (pp. 1-274). https://doi.org/10.1145/3680532.3689591

[21] Shubhangi, D. C., & Hiremath, P. S. (2009, January). Support vector machine (SVM) classifier for brain tumor detection. In *Proceedings of the International Conference on Advances in Computing, Communication and Control* (pp. 444-448). https://dl.acm.org/doi/proceedings/10.1145/1523103

[22] Sio, A. (2025). Integration of embedded systems in healthcare monitoring: Challenges and opportunities. *SCCTS Journal of Embedded Systems Design and Applications*, *2*(2), 9-20.

[23] St-Vincent Villeneuve, A., & Plaisent, M. (2023, December). Framework for Choosing a Supervised Machine Learning Method for Classification Based on Object Categories: Classifying Subjectivity of Online Comments by Product Categories. In *Proceedings of the 13th International Conference on Advances in Information Technology* (pp. 1-5). https://doi.org/10.1145/3628454.3630041

[24] Su, W., Yuan, Y., & Zhu, M. (2015, September). A relationship between the average precision and the area under the ROC curve. In *Proceedings of the 2015 international conference on the theory of information retrieval* (pp. 349-352). https://doi.org/10.1145/2808194.2809481
[25] Sulaiman, M., & Azaman, A. (2022, November). Machine Learning Classification of Non-Specifically Trained Muscle between Endurance and Power Athletes. In *Proceedings of the 2022 9th International Conference on Biomedical and Bioinformatics Engineering* (pp. 127-132). https://doi.org/10.1145/3574198.3574218
[26] Tian, S., & Hui, G. (2024, May). Diabetes Prediction Using Machine Learning. In *Proceedings of the 2024 9th International Conference on Machine Learning Technologies* (pp. 16-20). https://doi.org/10.1145/3674029.3674033
[27] Utukuru, S., Pisipati, R. K., & Karlapalem, K. (2023, January). Missing data resilient ensemble subspace decision Tree Classifier. In *Proceedings of the 6th Joint International Conference on Data Science & Management of Data (10th ACM IKDD CODS and 28th COMAD)* (pp. 104-107). https://doi.org/10.1145/3570991.3571006
[28] Yang, S., & Zhu, J. (2024, July). Analysis of Global Research Trends and Development Trends in Smart Healthcare: A Bibliometric Study. In *Proceedings of the 2024 3rd International Symposium on Robotics, Artificial Intelligence and Information Engineering* (pp. 81-87). https://doi.org/10.1145/3689299.3689315
[29] Ye, J. (2024, December). Privacy Analyses in Machine Learning. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security* (pp. 5110-5112). https://doi.org/10.1145/3658644.3690862
[30] Ziegeldorf, J. H., Metzke, J., & Wehrle, K. (2018, December). SHIELD: A framework for efficient and secure machine learning classification in constrained environments. In *Proceedings of the 34th Annual Computer Security Applications Conference* (pp. 355-370). https://doi.org/10.1145/3274694.3274716

Authors Biography

Rawan Nassri Abulail is an associate professor in Philadelphia University and currently working as the head of computer science department in the information technology faculty. She received her B.Sc. degree in computer science and computer information systems from Philadelphia University, Amman, Jordan in 2002, B.Sc. degree in accounting from Philadelphia University, Amman, Jordan in 2004 M.Sc. and Ph.D. degrees from The Arab Academy for Banking and Financial Sciences, Amman, Jordan, in 2004, and 2009 respectively, all in Computer Information Systems.

Mohammed Yaarub Al-Hadithi is Student at College of Information Technology, Philadelphia University, Jordan. Presently he is doing Master's Degree from the University. He has authored and co-authored papers in peer reviewed journals and actively participates in conferences.