Computers and Electrical Engineering 105 (2023) 108528

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

Developing lung cancer post-diagnosis system using pervasive data analytic framework

Mohamed Shakeel Pethuraj *, Burhanuddin bin Mohd Aboobaider, Lizawati Binti Salahuddin

Faculty of Information & Communication Technology, Universiti Teknikal Malaysia Melaka, Durian Tunggal, Melaka, Malaysia

ARTICLE INFO

Keywords:
Auto encoder learning
Butterfly optimization
Jaya optimization
Lung cancer
Pervasive data analysis
Wearable sensors
Feature correlation
Data segregation

ABSTRACT

The data from lung cancer patients using wearable sensors and clinical assessments after observation is available to predict the disease's recurrence. In recurrence prediction, pervasive data analysis is required to prevent flaws in clinical correlations and data observations. This article proposes a Pervasive Data Analytical Framework (PDAF) for recurrence prediction. The proposed framework incorporates three processes: data segregation using Butterfly Optimisation, feature correlation using Jaya Optimisation, and autoencoder prediction. First, the data from the wearable sensor is segregated using observation count for its availability and discreteness. It prevents missing errors under different observation sequences for which the correlation rate is determined using the next optimization. In the Jaya optimization process, the features correlate with the clinical assessments to improve precision. The autoencoder predicts the occurrence of previous missing and non-correlated inputs for maximizing the detection rate. Using the proposed framework, the maximum gains of 9.22% in accuracy, 9.29% in detection, and 7.96% in recommendations.

1. Introduction

Lung cancer occurs in tissues present in air passages. It is caused mainly due to smoking habits. The two major lung cancers are small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). Data analysis is a process that analyses the data necessary to perform certain tasks in an application [1]. Cigarette smokers are 15 to 30 times more susceptible to developing lung cancer and also have a greater risk of dying from the disease. The chances of lung cancer are increased even with infrequent or light smoking. The amount of cigarettes smoked daily and the length of time a person has smoked determines the ill effects and leads to potential danger. There is a chance that lung cancer survivors who continue to smoke will develop a second lung cancer. If a person's parent, sibling, or child has lung cancer, their risk of developing the disease may be higher. For example, individuals may share a common environment with a radon source or be exposed to other carcinogens at work or home. The data analysis process forms the backbone for the decision-making process for providing a proper data set to reduce aggregation errors. The data analysis process for lung cancer is a crucial task to be performed at healthcare centres [2]. Data analysis requires accurate data captured by various functions and operations. Both SCLC and NSCLC need appropriate information for the diagnosis process. The data analysis process improves the quality of life and accuracy rate in providing patient services [3].

* Corresponding author.
E-mail address: shakeelji@ieee.org (M.S. Pethuraj).

https://doi.org/10.1016/j.compeleceng.2022.108528
Received 24 August 2022; Received in revised form 30 November 2022; Accepted 1 December 2022
Available online 10 December 2022

An X-ray of the lungs could show an abnormal tumor, often called a nodule. Small lung lesions may go undetected on an X-ray. However, a CT scan can pick them up quickly. Cytology of sputum is a diagnostic procedure. It is possible to detect lung cancer cells by examining sputum samples from people. The Lung Cancer Symptom Scale (LCSS) is verified using the data analysis process of lung cancer. LCSS provides augmented data for diagnosis and identification [4]. The data analysis process identifies critical features and patterns of lung cancer, providing an optimal data source for LCSS. The Deep Convolutional Neural Network (DCNN) algorithm is employed for lung cancer data validation. DCNN detects the exact data type for generating precise data for different processes [5]. The only way to confirm the presence of cancer is through a biopsy. Examining tissue and cell samples under a microscope is a common laboratory procedure. All the cells in a normal organism appear as same sized and neatly arranged. Cancer cells are less uniform in size and shape and appear disorganized.

Pervasive computing, also known as ubiquitous computing, reduces computer interaction. It improves the communication and interaction process among users [6]. Pervasive computing techniques are widely used in various fields to reduce the computation cost and the latency rate in providing user services. Wireless networks and sensors are used in pervasive computing techniques to gather information related to specific tasks [7]. Pervasive computing techniques are also used in the data analysis process of lung cancer. It uses the segmentation method to classify the type of data presented in the database [8]. Computed Tomography (CT) scans are primarily used in lung cancer detection, providing information about lung cancer [9]. The pervasive computing technique gets information from CT scans and identifies the important aspects of the database. It reduces the time and energy involved in the computation process, thereby improving its efficiency [10]. Physical activity and a well-balanced diet are essential to prevent cancer. They go hand-in-hand with warding off chronic diseases, including cancer. Normal individual's Body Mass Index (BMI) goals typically fall between 18.5 and 25.0. A person can find Body Mass Index (BMI) by visiting our Health Library or visiting a doctor. Eating a healthy diet including varied nutrients that the body needs and cutting down on high-fat, processed foods is another way to reduce the risk of developing cancer.

An Optimisation algorithm is a procedure that determines the optimal solution to solve problems. It is deployed in various fields to improve the performance and reliability of the system, and mostly for lung cancer prediction, requiring precise data accumulation for detection [11]. Consequently, data analysis may conclude that applying the Pervasive Data Analytical Framework can yield useful outcomes for lung cancer detection. The designed fuzzy soft expert system is helpful for physicians in determining whether a patient has lung cancer. Lung cancer prediction is difficult to perform in every healthcare centre. It provides varying features for delivering services and diagnosis recommendations [12]. Analyzed data is applied to the lung cancer prediction process and generates accurate augmented data. Hence, the process leverages accuracy and precision [13].

The lung cancer detection process requires patient-centric clinical and observation data for accurate correlation. Image processing and Machine learning methods are assimilated into lung cancer detection to enhance the system's likelihood and efficacy [14]. According to research observations, health controls differ between patients with lung malignancies. Lung patients at stage still had considerably more plasma than the control group.

Nonetheless, a separate study that similarly used PDAF as a screening tool could not distinguish and predict between two groups containing thousand high-risk smokers. Hence, the development of lung cancer could not be anticipated. The Whale Optimisation Algorithm (WOA) is also used for lung cancer detection using analyzed datasets. WOA reduces the time required for identification and classification processes and improves accuracy [15]. The article's remaining sections are organized as follows: Section 2 lists the relevant works. The proposed framework is elaborated on in Section 3. Section 4 discusses the results, Section 5 examines the comparative analysis, and Section 6 discusses the research paper's conclusion.

2. Related works

The research article [16] introduced a novel lung nodule detection method named the Two-Stage Convolutional Neural Network (TSCNN) approach. The u-net segmentation method used in the research segments the data necessary for the detection and prediction. The classification and identification process plays a significant role in providing accurate information for the detection process. The classification process suppresses latency in the detection. The proposed TSCNN approach achieves better accuracy in detecting nodules, augmenting the system performance.

The authors of [17] designed a novel, network-based multi-omics clustering method for cancer data analysis. Aggregation multi-omics and Wasserstein Distance Clustering (aWCluster) are used here to analyze the datasets and generate analytical data for processing. The important features and patterns are identified here that reduce the identification latency. This method enhances the effectiveness and detection system reliability.

The authors of [18] introduced a Dependent Nearest Neighbour (DNN)-based recurrence method for big data in cancer diagnosis. Cancer recurrence prediction is crucial and complex to perform in every healthcare centre. TSVR algorithm is also used here to ascertain the patterns using the mining method. The proposed DNN-based approach improves the overall detection accuracy.

The authors of [19] developed a novel logistic regression prediction model for Long Non-Coding RNA (lncRNA). A Random Forest (RF) algorithm is available to find out the exact disease pattern and produce analytical data sources for diagnosis and prediction. RF reduces both time and energy consumption while improving detection accuracy. Butterfly Optimisation, which employs X-rays to create cross-sectional images of the chest, is the gold standard for identifying lung cancer. Scans using radio waves and powerful magnets (MRI) provide detailed images of soft tissue.

The research paper [20] introduced a new dynamic Bayesian network for lung cancer screening. National Lung Screening Trial (NLST) data is used for decision-making, and Machine Learning (ML) is employed techniques to perform optimization and classification processes. NLST enhances the decision-making accuracy of the lung cancer screening process. It increases the diagnosis

process's significance level.

The authors of [21] designed a Deep Neural Network (DNN) model for lung cancer detection. The gene expression approach is used here to address the exact problem of the patients. Dimension and imbalanced data are identified by the gene expression approach that produces the definitive analysis data. This method achieves fair accuracy and service outcomes for the patients.

The research article [22] introduced a biomarker identification method for Non-Small Cell Lung Cancer (NSCLC). Micro RNA (miRNA) and gene expression data are assimilated for dense data generation and are used for detection. The microarray data analyses the datasets necessary for the biomarker detection process. Biomarkers are mainly used for the clinical data management process containing exact details about lung cancers. This method provides accurate biomarkers that are presented in NSCLC.

In [23], authors have proposed a meta-data analysis method for Prophylactic Cranial Irradiation (PCI) NSCLC detection. The proposed analysis method is used mainly to assess the impact of PCI in NSCLC that provide appropriate information related to lung cancer. Capable variables and parameters that produce optimal data for lung cancer detection are identified here. The analysis process improves the efficiency and effectiveness of the data management system.

In [24], authors have introduced a retrospective analysis approach to Small Cell Lung Cancer (SCLC). The system is used mainly for the long-term survival prediction process. Concurrent Chemoradiotherapy (CCRT) is used for SCLC detection, generating dense analytical data. This method identifies lung cancer patients' overall survival (OS) rate during diagnosis.

Based on the above survey, there is a lack of proper tools in existing models. Hence, this paper has suggested the Pervasive Data Analytical Framework model using a Butterfly Optimisation algorithm. The following section discusses the proposed Pervasive Data Analytical Framework model briefly.

3. Proposed framework

The early stage of lung cancer is completely diagnosed through surgical tumor resection. However, even after the complete resection of the tumor, 30–55% of cancer-treated patients are exposed to recurrence after 2-5 years of the surgery. Therefore, this proposed framework's main challenge is identifying lung cancer-treated patients with an elevated risk of recurrence post-surgery. This recurrence prediction helps patients with continuous observation and personalized adjuvant treatment. PDAF is performed by validating predictive models and diagnostic recommendations for providing adjuvant therapy in the initial stages. Individualized risk assessment, rather than lung cancer prediction post-first diagnosis based on Auto Encoder learning, would be helpful for adaptable diagnosis for these patients. Our continuous observations through wearable sensors gave varying recurrence predictions for patients with similar feature exhibitions. In some patients, irregular comments and miscellaneous food practices result in recurrence and post-surgical procedures. In contrast, many patients ensure continuous compliance and pervasive data analysis after five years without recurrence. The PDAF for the recurrence prediction of each patient relies on clinical correlation powered by Jaya Optimisation. The accuracy of this technique is estimated using the bonding between PDAF and the clinical assessments through independent auto-encoder validations. Fig. 1 presents the proposed framework.

The proposed technique for lung cancer recurrence prediction is designed to identify the flaws in clinical correlation and data observations. The balancing factor of recurrence prediction and pervasive data analysis is required for stabilizing a patient's health based on continuous observations. The auto encoder's input and hidden layers can benefit from adding noise, increasing the learning accuracy. Designers conduct a thorough investigation of the impact that various forms of noise and degrees of corruption have on network performance.

It ensures availability and discreteness in data analysis through Butterfly Optimisation for recurrence prediction. This technique uses three distinct processes: Butterfly Optimisation, Jaya Optimisation, and Auto Encoder for recurrences of lung cancer prediction post-first diagnosis. Usually, lung cancer recurrences in patients after the suppression of original cancer following several years. The techniques of chemotherapy and targeted therapies may not be able to prevent a recurrence. Hence the method prediction model can lengthen a patient's life expectancy. The above processes are used for sequential input and output computation for heterogeneous data analysis. First, the observed continuous data from wearable sensors are segregated for their availability, and the discreteness of the data relies on the observation count performed by Butterfly Optimisation. This optimization helps to address missing errors under distributed observation sequences, for which the feature correlation is determined. Second, the determining features from the available

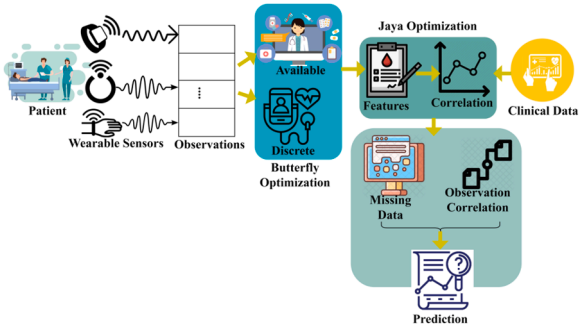

Fig. 1. Proposed PDA Framework.

instances are correlated with the clinical assessments through Jaya Optimisation for improving precision. Third, the autoencoder predicts the recurrence of the previously missing data and non-correlated inputs for augmenting detection and diagnosis recommendation.

From the pervasive data analysis-assisted processing of the input and output validation, the correlation of the features is determined under distinguishable continuous observations. Therefore, the proposed technique is responsible for distinguishing clinical assessments pipelined for constant estimation of input and output at different time intervals with missing errors and flaws. The recurrence prediction is modeled for the previous missing errors and non-correlated input features. The location and subtype of recurrent lung cancer are major factors in determining the best course of treatment. Lung cancer returns to the patient with severity, and treatment options are limited. Hence, this technique aims to maximize the detection of such recurrence and give diagnosis recommendations for the complete resection of the tumor. The clinical assessment relies on input and output estimation for harmonized data analysis resulting in missing data in observation correlation does not predict the recurrences respectively, then

$$\sum_{o^b=Pt}(I_p,O_p)_{W_s}=\sum_{W_s=Pt}\left\{\sum_{o^b=t}(I_p)_{W_s o^b}+\left(1-\left[\frac{(I_p)_{o^b}}{\sum(I_p+O_p)_{o^b}}\right]\right)\right\} \tag{1}$$

In Eq. (1), the variableI_p and O_p used to denote the probability of continuous input and output validation for data analysis are pipelined $\exists$ through wearable sensors W_s from the post-operative patientPt. The maximum probability of recurrence prediction of lung cancer$R^p=1$ achieves high I_p and O_p computation based on clinical correlation and data observations for the adjuvant treatment, if o^b represents the observation count of these patients. Instead, Pt and t are not idle due to $\exists$ as $R^p\in[0,1]$ for the clinical assessments. Therefore, the condition $R^p\in[0,1]$ is accounted for stabilizing a patient's health at different time sequences$t$, preventing flaws. This flaw is addressed for recurrence prediction and pervasive data analysis. The continuous observation data using wearable sensors and patient clinical assessments are pipelined for input and output estimation for data analysis, maximizing precision. Depending on the constant observation, the missing data instances are classified as represented in Fig. 2.

The observations are classified$as\forall R^P\in[0,1]$ where I_p and$R$ are valid under different classes. In the classification process, $R^P\in[0,1]$ is validated forW_Sob (discrete) andC_{ob} (continuous) data assessments. This is required to prevent missing R(or)I_p detection provided$0<t<I_p\in R$. However, this is identified from the consecutive sequence other than the current I_p(Refer to Fig. 2).

3.1. Correlation for observations

From the wearable sensor-assisted data observations, pervasive optimization should be carried out for lung cancer recurrence prediction to prevent clinical correlation and data observation flaws. The continuous input and output validation performed for heterogeneous data analysis were observed using wearable sensors. It is used to identify lung cancer recurrence in a post-operative patient, as in Eq. (1). The probability of input and output validations are pipelined at t sequences without flaws. Therefore, the probability of observation data $\rho(OB^d)$ is given by,

$$\rho(OB^d)=\frac{\sum_{o^b\in t}W_s}{\sum_{o^b\in Pt}C_a}*F_L\left(-\frac{I_p}{Pt}.V.C_c\right) \tag{2}$$

From Eq. (2), the variables W_s and C_a are used to denote the wearable sensors and clinical assessments analysis using the clinical correlation C_c at different time sequences t and the actual data observations, respectively. From the continuous data observations, the pervasive optimization for lung cancer prediction of $1-\left[\frac{(I_p)_{o^b}}{\sum(I_p+O_p)_{o^b}}\right]$ is validated using$V$. The condition for maximizing $\rho(OB^d)$ is $R^p=1$ as the clinical assessment predicts the chances of recurrence for stabilizing a patient's health, and therefore, the recurrence prediction is validated based onC_c,$W_s\in t$. The observation data of C_c in t helps to validate the input and output for clinical assessments and identify flawsF_L. This clinical correlation is computed using Eq. (3) and is valid for t observation sequences.

$$C_c,\ W_s\in t=[(1-V)\ \frac{I}{C_a}.\ R^p.\frac{I}{\exists}-\left(C_a-OB^d\right),\ o^b\in Pt \tag{3}$$

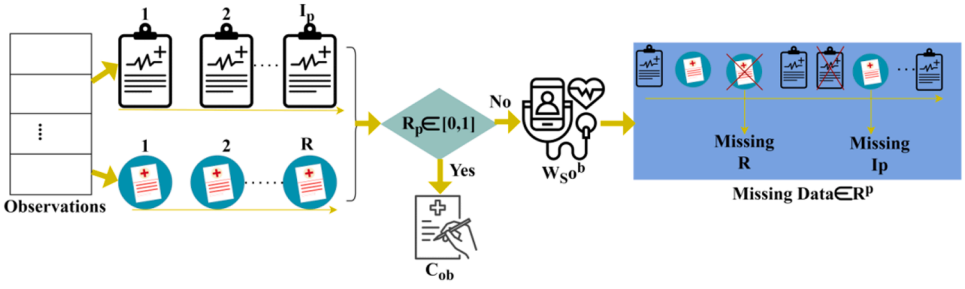

Fig. 2. Missing Data Classification Instance.

In Eq. (3), the input and output computations are pipelined to prevent flaws in the available C_a. And it is analyzed for a post-operative patient at different sequences. If this condition C_c,$W_s \in t$ exceeds, then recurrence is required. The flaws in clinical correlation and data observation augment$V$ to identify the chances of recurrence prediction. The wearable sensors contain inputs and outputs of OB^d at different time intervals$t$. The outputs for the multiple observations are classified as available and discrete through Butterfly Optimisation.

3.2. Butterfly optimisation for data segregation

First of all, Butterfly Optimisation is based on the behavior of the butterflies as they involve in food-finding and replications. The butterflies, with their senses, should reach their target (food source). The butterflies use their senses of touch, hearing, smell, taste, and sight while searching for food and mating partners. This sensing capability is especially useful as they move from place to place. These two processes confront two issues; generally, the formulation of fragrances is compared to observations of the availabilityA_v, and variation of stimulus intensity is compared to discreteness in observationsD_r. The actual stimulus intensity is exploited for incrementing the observations; the magnitude factor is responsible for its continuity. The fragrance is a common factor that attracts other neighbors from which new intensities are estimated. This technique is used to segregate the data observations based on the count. Data segregation uses Butterfly Optimisation, and the observation count is represented in Eq. (4).

$$OB^{Count} = \sum_{W_s \in Pt}^{o^b} Ds_g (A_v)^{vc} \tag{4}$$

In Eq. (4), the data segregation is represented asDs_g and the exponent for the power of the availability and vc represents the varying observation count. In many cases, the variables Ds_g and vc are bound between “zero” and “one” using traditional Butterfly Optimisation. The butterfly migrating locations are marked using the fragrance and estimated corresponding fitness. This estimation relies on the existing (current) and preserved (previous) fitness values. Their availability and discreteness segregate the continuous data observation and clinical assessment of post-operative patients through observation count. Data observation increments the count using the traditional Butterfly Optimisation algorithm, a new metaheuristics algorithm that takes cues from butterflies’ social and foraging behaviors. This paper provides three variants of the Butterfly Optimisation algorithm that improve upon the original by providing a fair balance between their exploratory and exploitative capabilities. It also avoids the algorithm’s tendency to get stuck in local optimums.

The segregation for continuity and discreteness is performed for the accumulated patient information. In this sequence, the optimization estimates the availability of the discrete observation. The clinical correlation and comments will help diagnose lung cancer recurrence prediction for other patients by Eq. (4) at an early stage. Segregating the availability and discreteness of different observation sequences are the two primary phases of this technique. It is possible to monitor motor-related symptoms, such as balance, gait, and spasticity, with the help of flexible wearable sensors. Employing more digital biomarker systems can aid in continuously monitoring the patient’s activity and improve their care. Also, sensors with improvised designs, along with their placement on the hips, knees, and legs, will help gather patient information. The patient takes the step of the continuous observation C_{ob} for the best precision Z_p, and the related equation is shown in Eq. (5).

$$(C_{ob})_i^{o^b+1} = (C_{ob})_i^{o^b+1} + \left(OB^d \times Z_p - (C_{ob})_i^{o^b+1}\right) \times OB_i^{Count} \tag{5}$$

In Eq. (5), the output is given by $(A_v)_i^{o^b+1}$ for ith patients in its continuous observations, the current best precision from all the outputs in the current wearable sensor data is represented as$Z_p \in [0, 1]$, ?denoted asOB^d and the availability of ith patients is represented as A_{v_i}. Eq. (6) defines the mathematical representation of the observation count. It is predicted that 30 to 50 percent of individuals who undergo treatment for stage 1 lung cancer will experience recurrence. This is possible even after a successful operation, as cancer can manifest locally or in distant areas of the body. Usually, cancer spreads from its original location and rarely reappears in the same spot. The wearable sensor data observations for jth and kth patients in observation are given by $(C_{ob})_j^{o^b+1}$ and $(C_{ob})_k^{o^b+1}$, respectively.

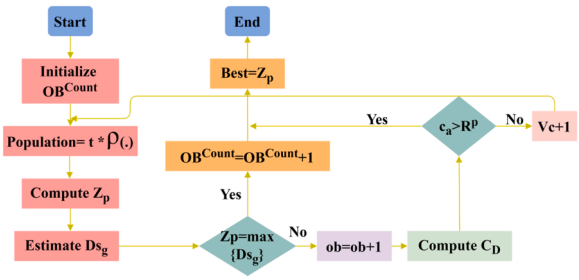

Fig. 3. Butterfly Optimisation Flow.

$$(C_{ob})_i^{o^b+1} = (C_{ob})_i^{o^b+1} + \left[\left[\left(OB^d\right)(C_{ob})_j^{o^b+1}\right] - (C_{ob})_m^{o^b+1}\right] \times OB_i^{Count} \tag{6}$$

For both wearable sensors and clinical assessments of chances of recurrence prediction, availability and discreteness in data observations are segregated. In Butterfly Optimisation, the segregation of availability and discreteness through observation count prevents missing errors. The Butterfly Optimisation flow is illustrated in Fig. 3.

The optimization process initializes$OB^{count}$for$t^*\rho(OB^d)$ Population. Using theZ_p, theDs_g is estimated for which max$\{Ds_g\} \in t$ is identified. If this condition is satisfied, then Z_p (current) is the best solution, and the new count is initialized. On the contrary, if the condition fails, thenC_a is estimated to verify if$C_a > R^P$ is true/ false. If this condition fails, variations are observed (Refer to Fig. 3). The available features and correlations are determined based on the distinguishable observation sequences using the next optimization to improve precision.

3.3. Jaya optimisation

Jaya Optimisation algorithm is used in this proposed technique for improving the precision while detecting the recurrence detection of lung cancer in a post-operative patient and stabilizing the patient's health. This proposed optimization modifies the worst result to the best-fit solution through precise population initialization and re-initialization. This algorithm generates and updates towards one single best-fit development for leveraging performance. $f(p)$ is the target function to be achieved for a best-fit solution, and some precisions $\{p^1, p^2, \ldots, p^n\}$ are initialized. This initialization is defined using the autoencoder process, which is random; the tumor dimension is represented by the variable$n$. These starting precisions are computed using the$f(p)$, and the best result p_b, and the worst result p_w is estimated. The maximum recurrence is computed through repeated observation validation for the target function. At any observationo^b, the features from the available instances are correlated with the clinical assessments as follows:

$$q_{i,j} = p^{i,j} + r_i^1\left(p^{i,b} - \left|p^{i,j}\right|\right) - r_i^2\left(p^{i,w} - \left|p^{i,j}\right|\right) \tag{7}$$

where $p^{i,b}$ is the variable i for the best-fit precision whereas $p^{i,w}$ is the variable i for the worst precision. The variable $q_{i,j}$ is the updated value of$p^{i,j}$; r_i^1 and r_i^2 are the two random integers for the ith variable during the ith observation of the data [0, 1]. The condition $r_i^1\left(p^{i,b} - \left|p^{i,j}\right|\right)$ represents the recurrence chances of the prediction given as the best-fit solution. The condition $r_i^2\left(p^{i,w} - \left|p^{i,j}\right|\right)$ represents the recurrence chances of the prediction to mitigate the least possible precision. The variablesr^1 and r^2 are randomly assigned for the optimal clinical assessment using feature correlation. The absolute recurrence prediction of these patients' result $|p^{i,j}|$ is considered to enhance the feature correlation of the algorithm further. The precision detection (best and worst) process is portrayed in Fig. 4.

The C_{ob} and $C_{ob} \in (ob + 1)$ are used as inputs for clinical correlation with the existing data. From the correlation$\forall\rho(OB^d)^*t \in R^P$, the $q_{i,j}$ is classified as $p^{i,b}$(best) and $p^{i,w}$(worst) based on$C_a > R^P$ (true/false) condition. This is an optimal classification for$Ds_g \oplus max\{Z_p\} \in t$ and$Ds_g \oplus \min\{Z_p\} \in R^D$ for precision detection. The first case is the high precision, whereas the second is the low-level precision for which C_c is further performed. The absolute precision is validated using the$f(p)$ function. If the new precision is optimal over the actual, it is substituted with the previously known value; the current value is retained otherwise. The term "precision medicine" refers to a medical approach in which a patient's medical history and the results of certain laboratory tests are used to develop individualized treatment plans. It can aid in precise diagnosis and better treatment in specific situations. Therefore, $q_{i,j}$ is accepted if it offers the best precision. The final results of the observations are retained by pursuing the clinical assessments. These results are conducted in the next observation correlation through Auto Encoder.

3.4. Autoencoder learning

In this autoencoder learning, the data observed from wearable sensors are analyzed and used to predict the occurrence of the previous missing errors and non-correlated inputs. The clinical assessments and data observations from the available sequences are correlated for maximizing detection and diagnosis recommendation. In particular, the different observation sequences are processed to prevent missing errors and non-correlated features, as in Eq. (1). The probability of predicting the occurrence Δ is estimated as follows:

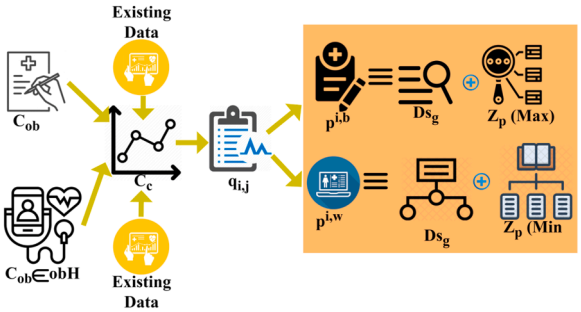

Fig. 4. Precision Detection Process.

$$\rho(\Delta) = \frac{\rho(I_p \cap O_p)}{\rho(O_p)} \tag{8}$$

The processes are pipelined for sequential input and output validation for heterogeneous data analysis using Butterfly Optimisation, Jaya Optimisation, and autoencoder learning for improving precision. Fig. 5 presents the autoencoder process for output detection.

The observation inputs rely on multiple q possibilities, $\in (i^*j)$; the validation for$p^{i,b}$ and$p^{i,w}$ are induced for $Z_p \geq q_{i,j}$ and $Z_p < q_{i,j}$ conditions. The conditions are used for identifying $\rho(\Delta)$, from which recurrence detection andvc are identified as outputs. The detection process relies on bothC_{ob} and$c_{ob} \in (Ob + 1)$ in identifying$\rho(\Delta)$. On the contrary, vc is estimated from the failing$'q'$ possibilities and$C_a < R^P$ conditions (Refer to Fig. 5). This technique predicts lung cancer recurrence chances after surgery. The data observations and clinical correlation are validated with the available features using optimization. The Butterfly Optimisation validates the classifier's ability through the availability and discreteness in data observations. It relies on non-redundant correlation for distinguishable clinical words. Jaya Optimisation is used for maximizing correlation precision by exploiting the available features from the training set. This autoencoder learning generates and validates the worst-fit solutions for the above-generated validation set. The worst-fit solution is computed as the weight-distributed summation of multiple observations. The computed risk rates are crucial in identifying and classifying the patients based on severity level.

4. Results and discussion

The proposed framework's performance is validated using the dataset from [25]. This dataset provides 309 male and female test results relating to lung cancer detection. Using the given data under 13 fields, the prediction is performed. The values are filled with the dataset's Yes (2) and No (1) options. Then, C_c is performed for outputs other than individual values for ease of recurrence prediction. Firstly, the correlation factor of varied observation (1 to 17) fields is presented along with its continuous andvc instances.

The correlation increases with the observed fields from the dataset for which OBandvc impacts are validated. The continuous OB leads to increased C_c Which is due to$C_a > R^P$ achievement; on the other hand, the fall in $(i, j) \in p^{i,w}$ result in C_c failures due to the breaks in ob and $(ob + 1)$ resulting invc; the adverse case is reported as the missing data. From this sequence, the autoencoder and the classification from Butterfly Optimisation require a newOB^{Count}. Therefore, the consecutive sequence stabilizes varied fields that improve precision (Refer to Fig. 6). After this process, the recurrence factor for different age groups under required observation fields (1 to 17) is presented in Fig. 7.

The considerable factors vary for different age groups based on the observation period. Depending on the observation period, classification for $p^{i,b}$ and$p^{i,w} \forall Z_p$ (max and min) is performed, through which the cumulative$\rho(A)$ of the considered 6 patients (each) is estimated. The average observation required is also presented in Fig. 7 above. In connection with the proposed framework, the independent process and its contributions are furnished in Table 1.

The classification optimization and autoencoder processes are analyzed for the observations, trials, errors,C_c, and precision in Table 1. The individual process stabilizes itself in different trials to improve consecutive processes. First, the $Ds_g$and vc are classified, followed by$p^{i,b}$ and$p^{i,w}$ estimation. In the AC process, the $\rho(\Delta)$ and precision-based competitions are performed. Similarly, the best and worst solutions for the different methods are computed in Table 2, and the AE stabilizes the $p^{i,b}$, whereas it is lost inDs_g. For people of different age groups, lung cancer in its initial stages typically does not manifest outwardly in the body. In addition, it may be years before a person with lung cancer experiences any symptoms. The squamous cell carcinoma is a generic form of lung cancer example where the size of cells are 30 mm only after having progressed for around eight years.

The first classification process alone achieves highDs_g as the$A \in$ does not perform classification. Therefore, the optimization/$A \in$ does not maximize classification. Contrarily, the optimization stabilizes$p^{i,b}$ other than$P^{i,w}$ Compared to the other processes.

4.1. Comparative study

This subsection presents the comparative study for the metrics precision, detection ratio, recommendations, error rate, and analysis time. The observation count and intervals (min) are the variables for analyzing the performance. Alongside the proposed framework, aWCluster [17], D-TSVR [18], and MRDCM [23] methods from the related works section are augmented in this study.

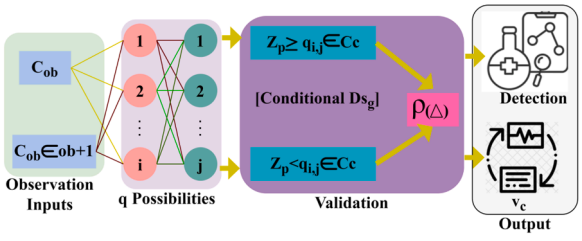

Fig. 5. Autoencoder Process.

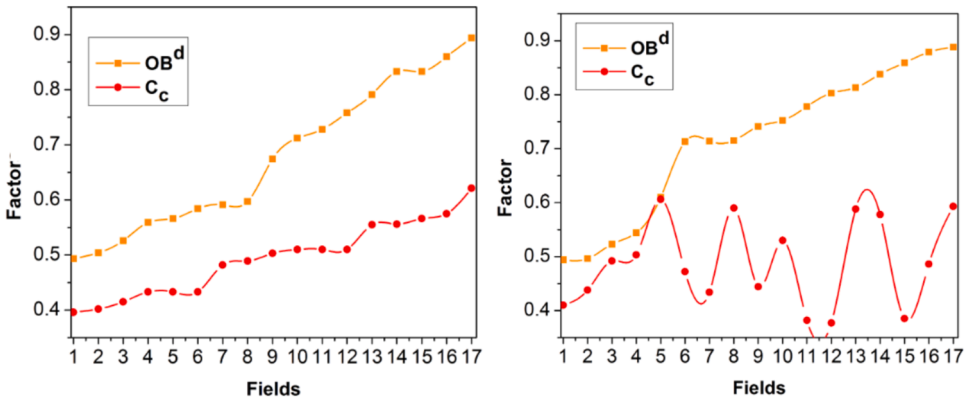

Fig. 6. C_c for OBand vc.

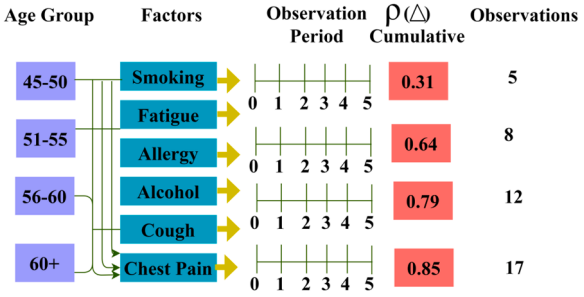

Fig. 7. ρ(Δ) for Different Age Groups.

Table 1
Independent Process Output.

Process	Observations	Trials	Error	C_c	Precision
Classification	5	4	0.31	0.37	0.638
	10	6	0.28	0.43	0.721
	15	10	0.21	0.52	0.796
Optimization	5	6	0.19	0.55	0.814
	10	8	0.15	0.62	0.872
	15	13	0.12	0.74	0.895
AE	5	6	0.1	0.77	0.921
	10	10	0.058	0.84	0.901
	15	17	0.031	0.96	0.964

Table 2
Best and Worst Solutions for Different Processes.

Process	Observation	$p^{i,b}$	$p^{i,w}$	Ds_g
Classification	C_{Ob}	0.63	0.21	0.91
	$C_{Ob} \in (Ob+1)$	0.49	0.36	0.74
Optimization	C_{Ob}	0.95	0.06	0.66
	$C_{Ob} \in (Ob+1)$	0.89	0.09	0.43
AE	C_{Ob}	0.82	0.13	0.19
	$C_{Ob} \in (Ob+1)$	0.75	0.2	0.03

4.1.1. Precision

Lung cancer recurrence is high due to irregular observation and miscellaneous food practices after surgery, as represented in Fig. 8. This proposed framework satisfies less precision for identifying the recurrence and missing data. In this continuous data observation using wearable sensors and clinical assessments, in many instances, is represented as $\rho(OB^d)$ and$R^p = 1$ such that $C_c, W_s \in t$ is validated until feature correlation takes place. The three distinct processes of the proposed framework predict the recurrence chances and provide diagnostic recommendations for adjuvant treatment. From the condition, the Butterfly Optimisation using data segregation

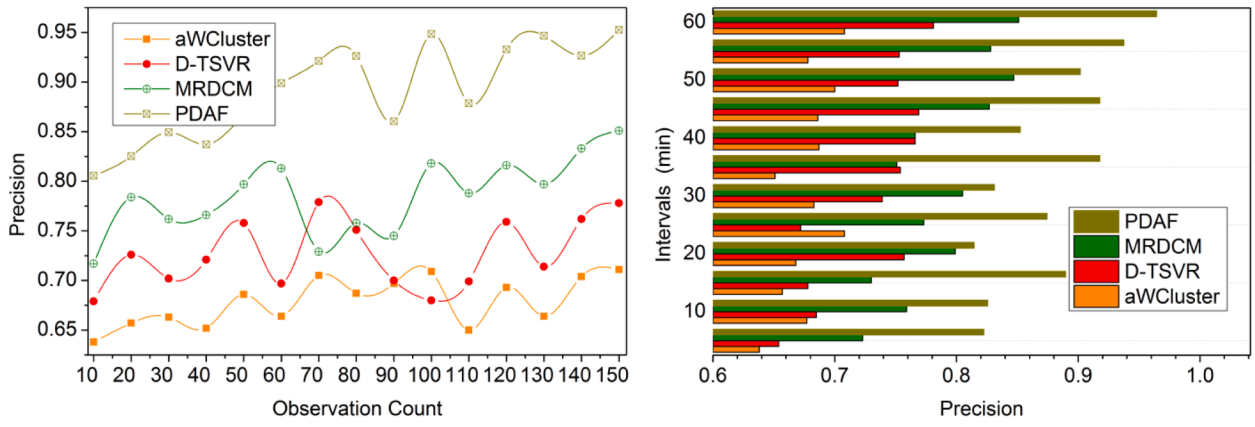

Fig. 8. Precision Comparison.

relies on availability and discreteness for harmonized data analysis through autoencoder learning. The missing data identification in pervasive optimization and the features of the available sequences are correlated based on clinical assessments. The different observation instances prevent missing errors in the clinical correlation, and it retains augmenting detection along with diagnostic recommendations. Hence, the recurrence prediction is validated for maximizing the diagnosis recommendation and data analysis for high precision.

4.1.2. Detection

In Fig. 9, predicting recurrence chances of lung cancer for a post-operative patient does not require clinical correlation and data observations at different time intervals. The input and output validation for harmonized data analysis prevents flaws and missing data through the correlation process. Jaya Optimisation is used to identify missing data, and observation correlation relies on clinical assessments to stabilize a patient's health and improve precision. The missing data is identified and analyzed by using Jaya Optimisation based on $r_i^1(p^{i.b} - |p^{i.j}|)$ and $r_i^2(p^{i.w} - |p^{i.j}|)$ such that $q_{i,j}$ achieves successive recurrence prediction of lung cancer, preventing missing errors, and therefore, further data segregation is performed for diagnostic recommendation is not presented. The three distinct processes satisfy high precision for continuous observations based on the flaws. Missing data is identified through the next optimization, and the correlation rate is reduced, preventing high detection due to continuous observation sequences.

4.1.3. Recommendations

The proposed framework achieves high diagnosis recommendations for missing data and observation correlation prediction through autoencoder at different time intervals aided in identifying the flaws and errors (Refer to Fig. 10). The available features are correlated for distinguishable observation sequences using $\frac{\rho(I_p \cap O_p)}{\rho(O_p)}$, the input and output validation through the aforesaid processes identifying the irregular observations and miscellaneous food practices for a post-operative patient. The main aim of these processes is to continue monitoring tumor size changes to provide adjuvant treatment. The occurrence prediction is due to pervasive data analysis using an autoencoder, analyzing the missing data detection to reduce errors. The addressed flaws and errors based on the clinical assessments rely on the condition $[(OB^d)(C_{ob})_j^{\phi^b+1}]$ for segregating data. An autoencoder prediction requires detection and diagnostic

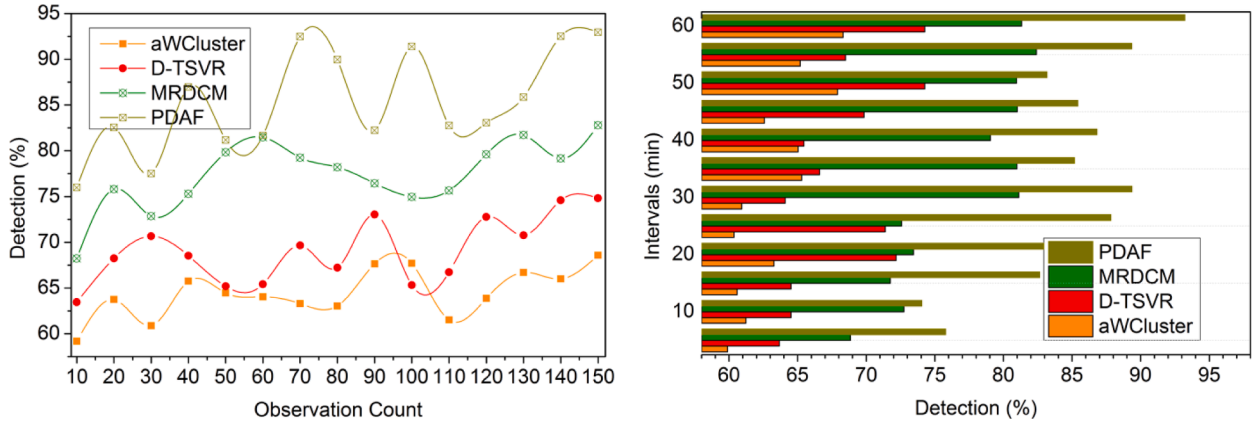

Fig. 9. Detection Comparison.

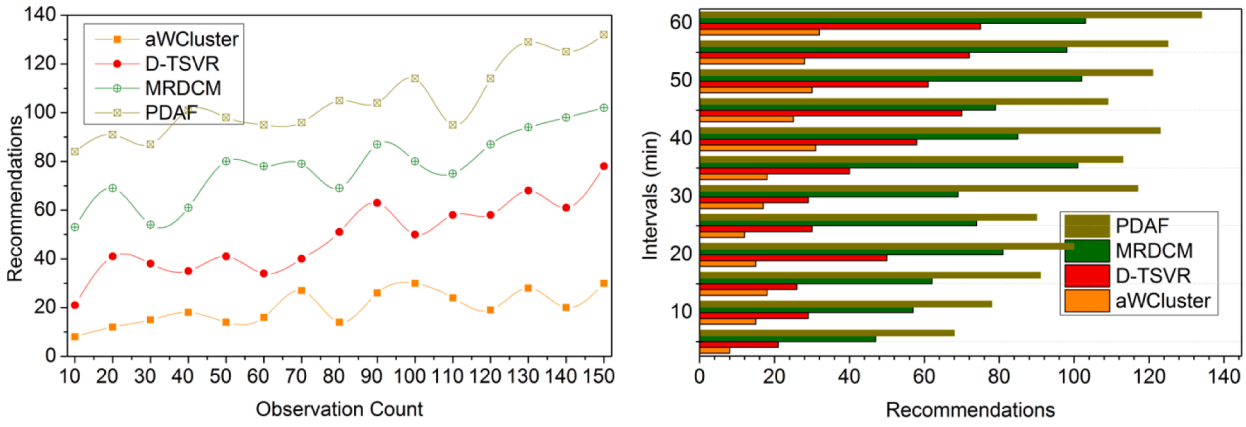

Fig. 10. Recommendations Comparison.

recommendations for distinguishable observation sequences. Therefore, the detection of missing data increases the stability of patient health as it verifies the flaws that depend on other factors in lung cancer recurrence prediction. As a result, the diagnostic recommendation is high, and observation also increases.

4.1.4. Error rate

In Fig. 11, the available and discrete wearable sensor data are segregated through Butterfly Optimisation for recurrence prediction of lung cancer using continuous data observation for a post-operative patient, as it does not provide diagnostic recommendations at different time intervals. The availability and discreteness of data from the observation count validation under distributed observation sequences, the flaws considered for recurrence prediction, and pervasive data analysis for the conditions $(C_{ob})_j^{o^b+1}$ and $(C_{ob})_k^{o^b+1}$ continuously using Auto Encoder learning. The observation correlation and missing data are addressed through a data segregation process that relies on data analysis of $(A_v)_i^{o^b+1}$ for the *ith* patient in its continuous observations, it involves further processing while preventing missing data. The irregular observation and miscellaneous food practices are analyzed for a post-operative patient that requires lung cancer recurrence prediction. Through Butterfly Optimisation, data segregation is performed for detection and diagnostic recommendation at different time intervals. The distinguishable observation sequences for which the correlation rate is determined and the proposed framework satisfy less error rate.

4.1.5. Analysis time

The flaws and error rate identification in recurrence prediction of lung cancer incorporates three distinct processes, as illustrated in Fig. 12. This proposed framework requires less analysis time where validation of the input and output depends on wearable sensors and clinical assessments in any intervals with its feature correlation instances. In this, error and flaw detection from the observation sequences $F_L\left(-\frac{I_p}{P_t}.V.C_c\right)$ is performed and analyzed for stabilizing a patient's health. The recurrence prediction and pervasive data analysis mitigate data segregation depending upon the continuous data observations and clinical correlation for its availability and discreteness using observation count. The pervasive data analysis is preceded by using Eqs. (4)–(7) estimations. The data segregation and available feature correlation are performed in this proposed framework using two optimizations to validate recurrence prediction further. It is possible to distinguish between seasonal changes and true insights with the help of analysts trained to spot outliers and irregularities. Validated clinical assessments are integrated into a pipeline for real-time estimation of inputs and outputs. This regular observation prevents flaws and missing data under different observation instances [as in Eq. (8)] through the autoencoder. Therefore, the missing data is high compared to the other methods considered. This consecutive heterogeneous data analysis estimates the analysis time for data observations. The above discussion is summarised in Tables 3 and 4 for observation count and interval, respectively. The findings from this summarisation are also presented.

5. Conclusion

This article discusses the process and performance of a pervasive data analytical framework for lung cancer recurrence prediction. The proposed framework operates on clinical and observation data for predicting lung cancers. This framework assimilates Butterfly and Jaya Optimisations and autoencoders to improve detection accuracy and recurrence prediction precision. Butterfly Optimisation classifies the discrete and continuous sequences to identify missing data. The classified data is optimized for their best and worst solutions using the Jaya optimization process. The optimization identifies the recurrent best solution through precise conditional validation in this process. The autoencoder is responsible for determining better precision for prediction. The classified best-fit results from the previous two optimizations are used in this process. This is recurrent for identifying better prediction-confining errors; the errors in classification and worst-fit solutions are confined to the previous operating processes. Depending on the clinical correlation,

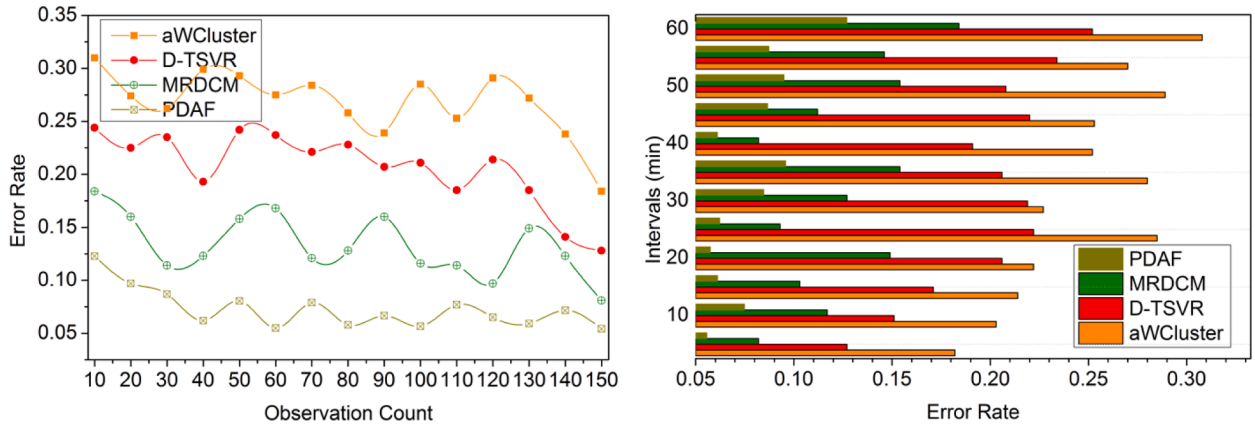

Fig. 11. Error Rate Comparison.

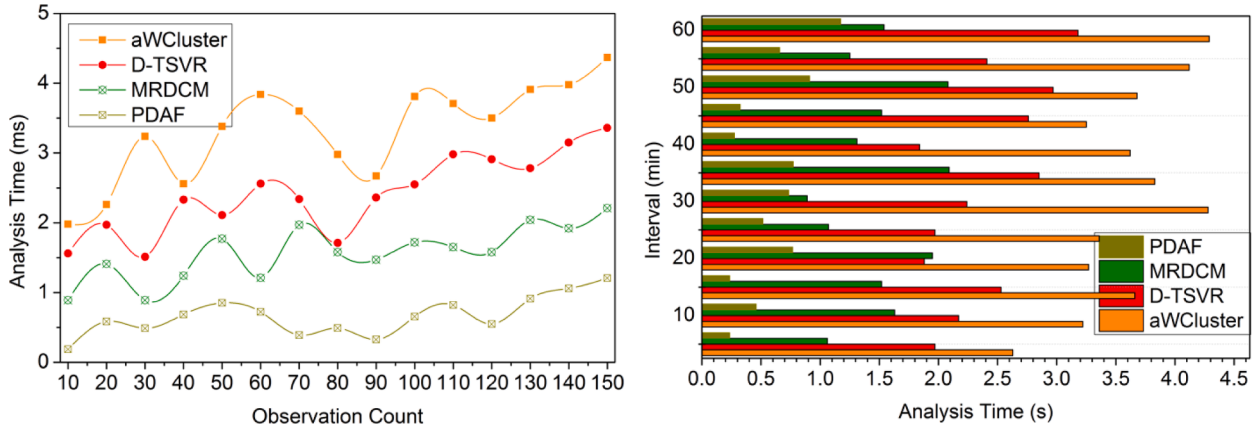

Fig. 12. Analysis Time Comparison.

Table 3
Comparative Analysis Summary for Observation Count.

Metrics	aWCluster	D-TSVR	MRDCM	PDAF
Precision	0.711	0.778	0.851	0.9527
Detection (%)	68.58	74.81	82.79	92.956
Recommendations	30	78	102	132
Error Rate	0.184	0.128	0.081	0.0542
Analysis Time (s)	4.37	3.36	2.21	1.207

Findings: This PDAF improves precision, detection, and recommendations by 8.64%, 8.78%, and 7.83%, respectively. This framework reduces error rate and analysis time by 7.68% and 10.64%, respectively.

Table 4
Comparative Analysis Summary for Intervals.

Metrics	aWCluster	D-TSVR	MRDCM	PDAF
Precision	0.708	0.781	0.851	0.9643
Detection (%)	68.31	74.27	81.33	93.221
Recommendations	32	75	103	134
Error Rate	0.308	0.252	0.184	0.1269
Analysis Time (s)	4.29	3.18	1.54	1.173

Findings: The proposed framework maximizes precision, detection, and recommendations by 9.22%, 9.29%, and 7.96%, respectively. This framework reduces error rate and analysis time by 6.06% and 10.17%, respectively.

the precision is improved during different sequences and observation intervals. A personal or family history of lung cancer is an unchangeable risk factor for the disease. As lung cancer tends to run in families, it is essential to discuss preventative measures with the physician if people notice a trend. Targeted treatments using the Pervasive Data Analytical Framework may be effective against cancers caused by these mutations. Most of the available tools and approaches in use today can detect cancer in its advanced stages, when therapy and a cure may not be efficient enough to control the disease, making early diagnosis of lung cancer difficult. Consequently, despite substantial advances over the past few years, early diagnosis is still inaccurate. The proposed framework maximizes precision, detection, and recommendations by 9.22%, 9.29%, and 7.96%, respectively, and reduces error rate and analysis time by 6.06% and 10.17%, respectively.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Mohamed Shakeel Pethuraj: Conceptualization, Writing – original draft, Writing – review & editing. **Burhanuddin bin Mohd Aboobaider:** Supervision. **Lizawati Binti Salahuddin:** Investigation.

Declaration of Competing Interest

On behalf of all the authors, I declare that there is no conflict of interest to disclose.

Data availability

No data was used for the research described in the article.

Acknowledgement

The authors would like to thank BIOCORE Research Group, Center for Advanced Computing Technology (C-ACT), Fakulti Teknologi Maklumat dan Komunikasi (FTMK) and Centre for Research and Innovation Management (CRIM), Universiti Teknikal Malaysia Melaka (UTeM) for providing the facilities and support for this research.

References

[1] Jiang L, Huang W, Liu J, Harris K, Yarmus L, Shao W, Chen H, Liang W, He J, Group AL. Endosonography with lymph node sampling for restaging the mediastinum in lung cancer: A systematic review and pooled data analysis. J Thor Cardiovasc Sur 2020;159(3):1099–108. Mar 1.
[2] Hardtstock F, Myers D, Li T, Cizova D, Maywald U, Wilke T, Griesinger F. Real-world treatment and survival of patients with advanced non-small cell lung Cancer: a German retrospective data analysis. BMC cancer 2020;20(1):1–4. Dec.
[3] Wu J, Guan P, Tan Y. Diagnosis and data probability decision based on non-small cell lung cancer in medical system. IEEE Access 2019;7:44851–61. Apr 9.
[4] Medbery RL, Fernandez FG, Kosinski AS, Tong BC, Furnary AP, Feng L, Onaitis M, Boffa D, Wright CD, Cowper P, Jacobs JP. Costs associated with lobectomy for lung cancer: an analysis merging STS and Medicare data. Annal Thora Surg 2021;111(6):1781–90. Jun 1.
[5] Wen S, Peng W, Chen Y, Du X, Xia J, Shen B, Zhou G. Four differentially expressed genes can predict prognosis and microenvironment immune infiltration in lung cancer: a study based on data from the GEO. BMC cancer 2022;22(1):1–2. Dec.
[6] Hirsch EA, Barón AE, Risendal B, Studts JL, New ML, Malkoski SP. Determinants associated with longitudinal adherence to annual lung cancer screening: a retrospective analysis of claims data. J Am Coll Radiol 2021;18(8):1084–94. Aug 1.
[7] Lu M, Hu C, Wu F, Shu L, Pan Y, Liu X, Liu P, Ma F, Deng C, Huang M. MiR-320a is associated with cisplatin resistance in lung adenocarcinoma and its clinical value in non-small cell lung cancer: A comprehensive analysis based on microarray data. Lung Cancer 2020 Sep 1;147:193–7.
[8] Tyczynski JE, Potluri R, Kilpatrick R, Mazumder D, Ghosh A, Liede A. Incidence and risk factors of pneumonitis in patients with non-small cell lung cancer: an observational analysis of real-world data. Oncol Therapy 2021;9(2):471–88. Dec.
[9] Tetzlaff F, Epping J, Tetzlaff J, Golpon H, Geyer S. Socioeconomic inequalities in lung cancer–a time trend analysis with German health insurance data. BMC Public Health 2021;21(1):1–12. Dec.
[10] Hüsing A, Kaaks R. Risk prediction models versus simplified selection criteria to determine eligibility for lung cancer screening: an analysis of German federal-wide survey and incidence data. European J Epidemiol 2020;35(10):899–912. Oct.
[11] Liao Z, Xie Y, Hu S, Xia Y. Learning from ambiguous labels for lung nodule malignancy prediction. IEEE Transact Med Imaging 2022. Feb 7.
[12] Chen Y, Yang H, Cheng Z, Chen L, Peng S, Wang J, Yang M, Lin C, Chen Y, Wang Y, Huang L. A whole-slide image (WSI)-based immunohistochemical feature prediction system improves the subtyping of lung cancer. Lung Cancer 2022;165:18–27. Mar 1.
[13] Ninomiya K, Arimura H. Homological radiomics analysis for prognostic prediction in lung cancer patients. Physica Medica 2020;69:90–100. Jan 1.
[14] Tammemägi MC, Darling GE, Schmidt H, Llovet D, Buchanan DN, Leung Y, Miller B, Rabeneck L. Selection of individuals for lung cancer screening based on risk prediction model performance and economic factors–The Ontario experience. Lung Cancer 2021;156:31–40. Jun 1.
[15] Lee HA, Rau HH, Chao LR, Hsu CY. Establishing a survival probability prediction model for different lung cancer therapies. J Supercomput 2020;76(8):6501–14. Aug.
[16] Cao H, Liu H, Song E, Ma G, Xu X, Jin R, Liu T, Hung CC. A two-stage convolutional neural networks for lung nodule detection. IEEE J Biomed Health inform 2020;24(7):2006–15. Jan 3.
[17] Pouryahya M, Oh JH, Javanmard P, Mathews JC, Belkhatir Z, Deasy JO, aWCluster Tannenbaum A. A novel integrative network-based clustering of multiomics for subtype analysis of cancer data. IEEE/ACM Transact Comput Biol Bioinform 2020. Nov 23.
[18] Yang AM, Han Y, Liu CS, Wu JH, Hua DB. D-TSVR recurrence prediction driven by medical big data in cancer. IEEE Transact Indus Inform 2020;17(5):3508–17. Jul 24.

[19] Wang B, Zhang J. Logistic regression analysis for LncRNA-disease association prediction based on random forest and clinical stage data, 8. IEEE Access; 2020. p. 35004–17. Feb 17.
[20] Petousis P, Winter A, Speier W, Aberle DR, Hsu W, Bui AA. Using sequential decision making to improve lung cancer screening performance, 7. Ieee Access; 2019. p. 119403–19. Aug 16.
[21] Liu S, Yao W. Prediction of lung cancer using gene expression and deep learning with KL divergence gene selection. BMC Bioinform 2022;23(1):1–11. Dec.
[22] Jin X, Guan Y, Zhang Z, Wang H. Microarray data analysis on gene and miRNA expression to identify biomarkers in non-small cell lung cancer. BMC cancer 2020;20(1):1–10. Dec.
[23] Witlox WJ, Ramaekers BL, Lacas B, Le Pechoux C, Pignon JP, Sun A, Wang SY, Hu C, Redman M, van der Noort V, Li N. Individual patient data meta-analysis of prophylactic cranial irradiation in locally advanced non-small cell lung cancer. Radiother Oncol 2021;158:40–7. May 1.
[24] Doshita K, Kenmotsu H, Omori S, Tabuchi Y, Kawabata T, Kodama H, Nishioka N, Miyawaki E, Iida Y, Miyawaki T, Mamesaya N. Long-term survival data of patients with limited disease small cell lung cancer: a retrospective analysis. Investigat New Drug 2022;40(2):411–9. Apr.
[25] https://data.world/sta427ceyin/survey-lung-cancer. Accessed July,15, 2022.

Mohamed Shakeel Pethuraj received his M.Sc. in Information Technology in 2007 from Nehru Memorial College, MBA from Bharathi Dasan University, Trichy in 2009, and ME in Computer Science and Engineering from Karpagam University in 2013, respectively. Presently he is doing extended research in Universiti Teknikal Malaysia Melaka, Malaysia. His research interests include Medical Image processing, Networking, and Cloud IoT.

Burhanuddin bin Mohd Aboobaider is currently working as Associate Professor and Deputy Dean of Research and Postgraduate Studies, Faculty of Information Technology and Communication Universiti Teknikal Malaysia Melaka, Malaysia. He has completed his Ph.D. in Industrial Computing from the same University. His area of research interest is related to Decision Support System, Optimization Techniques, Operational Research, AI, and Health Informatics.

Lizawati Binti Salahuddin received Ph.D. in Information System (UTM) and MSc in Bio-system, KAIST, South Korea. She completed BSc in Computer Science (Software Engineering) (UTM) in the Field of Specialization Healthcare and Information Systems.