Materials Today: Proceedings 50 (2022) 40–47

Contents lists available at ScienceDirect

Materials Today: Proceedings

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

# Prediction of Cancer Disease using Machine learning Approach

F.J. Shaikh *, D.S. Rao

[a] School of Computer Engineering and Technology, MIT Academy of Engineering, Alandi Road, Pune, India
[b] Computer Science Engineering, Koneru Lakshmaih Education Foundation, Hyderabad Campus, India

ARTICLE INFO

*Article history:*
Available online 16 April 2021

*Keywords:*
Cancer
Deep learning
ML
ANN
SVM
Decision tress

ABSTRACT

Cancer has identified a diverse condition of several various subtypes. The timely screening and course of treatment of a cancer form is now a requirement in early cancer research because it supports the medical treatment of patients. Many research teams studied the application of ML and Deep Learning methods in the field of biomedicine and bioinformatics in the classification of people with cancer across high- or low-risk categories. These techniques have therefore been used as a model for the development and treatment of cancer. As, it is important that ML instruments are capable of detecting key features from complex datasets. Many of these methods are widely used for the development of predictive models for predicating a cure for cancer, some of the methods are artificial neural networks (ANNs), support vector machine (SVMs) and decision trees (DTs). While we can understand cancer progression with the use of ML methods, an adequate validity level is needed to take these methods into consideration in clinical practice every day.

In this study, the ML & DL approaches used in cancer progression modeling are reviewed. The predictions addressed are mostly linked to specific ML, input, and data samples supervision.


## 1. Introduction

The main weight of ailment overall is as Lung malignancy that is the most inescapable disease in the two men and women [1]. A few other reports estimate some 221,200 new cases of pulmonary cancer occur and represent approximately 13% of all cancer diagnoses in 2015. Approximately 27 percent of all cancer deaths are attributed to lung cancer [2]. Lung nodules must therefore be closely examined and monitored when at an early stage. In this study, the ML & DL approaches used in cancer progression modeling are reviewed. The predictive models discussed here are based on different supervised ML techniques, input and data samples.

A Local Binary Pattern (LBP) is an image operator that transforms an image into an array or picture of integer labels that describe the appearance of the small picture. These labels are then used for further image analysis, most frequently in the histogram. The LBP texture operator has become a popular approach to various applications thanks to its discriminative power and computational simplification [2].

A Binary Local Pattern (LBP) is a picture administrator that changes over a picture into a variety of number names speaking to its essence. These markers are then more commonly used in the histogram for further image processing. In the last three decadent years the prevalence of prostate and breast cancer in male and female cancer has been the largest, but lung cancer remains the highest in cancer-patient mortality [3]. One of the main reasons for this is that prostate and breast cancer prognostic models are comparatively more advanced and systemic than pulmonary cancer. Thus, it is urgently necessary to establish an effective early-stage lung cancer forecast model. In linear and non-linear problems, SVM has superior predictor performance and is widely used in various fields including in medical matters. Even if SVM is a superior classifier, the field of cancer prognosis models is relatively immature [4].

The mutation test [5] has become an important tool for deciding the right therapy options for patients in clinical tests. Direct sequencing is an alternative approach for unknown mutations based on screening. The Mutation Test for Epidermal Growth Factor Receptors (EGFR) has been identified for lung cancer genetic mutation testing [4]. A contrast with their non-ensemble variants of two types of categorizing equipment Artificial Neural Network (ANN) and Support Vector Machine (SVM) is published. The

* Corresponding author.
*E-mail address:* farhahashaikh20@gmail.com (F.J. Shaikh).

https://doi.org/10.1016/j.matpr.2021.03.625
2214-7853/

weight of the misjudgment for the majority is higher than that of the minority class and is likely to cause misjudgment. Traditional algorithms of classification are not successful and excellent.

### 1.1. Topology of machine learning & deep learning algorithms

In order to predict the various types of diseases, different deep learning & machine learning algorithms are used , such as Support vector machine (SVM), Neural Network (NN), LR, Nevin biases (NB), Fuzzy logic, transfer learning, ensemble learning, Transduction learning, KNN, and Adaboost are mostly utilized in diverse contributions. Moreover, SVM is categorized into Boosted SVM & MLSVM for predicting distinct diseases in the earlier contributions. Similarly, NN is classified as Dynamic Neural Network (DNN) & Convolution Neural Network (CNN) which are employed for diagnosing different diseases in different contributions. Moreover, GBDT is a modified form of DT, CVIFLR is the modified form of LR that are used for detecting diseases. Moreover, RF and Fuzzy logic is grouped into HRFLM and Fuzzy SVM, respectively in order to predict discrete diseases in various contributions. So, for predicting lung cancer in an efficient manner with the help of improved machine learning techniques can be use.

## 2. Literature review

ChaoTan et al [1] explored the feasibility of using decision stumps as a poor classification method and track element analysis to predict timely lung cancer in a combination of Adaboost (machine learning ensemble). For the illustration, a cancer dataset was used which identified 9 trace elements in 122 urine samples. The sample set partitioning was performed using Kennard and Stone algorithm (KS), combined with alternative samples. The adaboost forecast results were contrasted with the Fisher Biased Analytic (FDA) results. In the test set, 100% of Adaboost's sensitivity for both cases was reached, 93.8% of accuracy was 95.7% and 95.1% respectively for case A and case B 96.7%. The structure of both the test data is less reactive than the FDA and the change is often easier to monitor than the FDA. The Adaboost appeared superior to FDA and proved that combining Adaboost and urine analysis could be a valuable method through clinical practice for the diagnosis of early lung cancer.

Tae-WooKim et al. [2], have developed a decision tree on occupational lung cancer. In 1992–2007, 153 lung cancer cases were reported by the Occupational Safety and Health Researcher's Institute (OSHRI). The objective parameter was to determine if the situation was accepted as lung cancer linked to age, sex, smoking years, histology, industry size, delay, working time and exposure of independent variables. During the whole journey for indicators for word related cellular breakdown in the lungs the characterization and relapse test (CART) worldview is utilized. Presentation to known lungs disease specialists was the best pointer of the CART model. As the CART model is not absolute, the functionality of lung cancer must be carefully determined.

Maciej Zięba et al. [3] introduced boosted SVM in 2014 which is dedicated to solving imbalanced results. The solution proposed combined the advantages of using ensemble classifiers with cost-sensitive support vectors for uneven data. In addition, a method for extracting decisions from the boosted SVM was presented. In the next step, the efficiency of the solution proposed was assessed by comparing the performance of the unbalanced data with other algorithms. Finally, improved SVM was used to estimate after surgery life expectancy in patients with lung cancer.

A multiclass data pathway behavior transformation approach called Analysis-of-Variance Based Feature Set (AFS) was suggested by Worrawat Engchuan [4]. The results of the classification using pathway behavior derived from the proposed approach indicate that all four lung cancer data sets used have high classification capacity in three-fold validity and robustness.

H. Azzawi et al. [5] proposed a GEP (gene expression) model to forecast microarray data on lung cancer in 2016. In order to extract important lung cancer related genes, the authors use two approaches for selecting genes and thus suggest specific GEP prediction models. The validation of the cross-data collection was tested for reliability. The test results show that, considering precision, sensitivity, speciality, and region under the recipient functional property curve, the GEP model using fewer features surpassed other models. The GEP model was a better approach to problems of diagnosis of lung cancer. It has been found.

Panayiotis Petousis et al. [6] created and evaluated a range of dynamic Bayesian Networks (DBN) to assist in informing decisions about lung cancer screening by providing insights into how longitudinal data can be used. The NLST dataset LDCT arm has been used in creating and exploration five DBNs for high-risk people. 3 of the DBNs were designed with a reverse style, and 2 through methods of structural learning. All applications are based on population, smoking status, a history of cancer, family history of lung cancer, risk factors for exposure, lung cancer co-orbidities and information on LDCT screenings. In view of the uncertainty resulting from lung cancer screening, a lung cancer-state model was used to identify the individual's cancer status over time. These models have been tested on balanced cancer and non-cancer research and test sets in order to resolve data disequilibrium and over fitting. Expert judgments contrasted the results. In all three NLST test intervention stages, the average area underneath the curve (AUC) of the receiver operating feature (ROC) was above 0.75. Superior were compared models such as logistic regression and naïve Bay. Lung screening DBNs have demonstrated strong discrimination and predictive strength in both cancer and non-cancer cases.

The SEER database was used by Chip M. Lynch et al. [7] to classify the survival of lung cancer patients as a linear regression, decision trees, gradient boosting machines (GBM), support-vector machines (SVMs) and a custom set. In order to allow the comparisons between the different approaches, the main data attributes for applying these processes includes the tumor level, tumor size, gender, age, stage and number of primaries. Rather of being divided into classes, the prediction has been viewed as a continuous goal as a first step to enhancing survival. Results have indicated that the expected values conform to the actual values, which constitute the majority of the results, for low to moderate survival. The model that was most popular in the custom set was GBM, though Decision Trees did not function, because it consists of some discreet performance. The outcome show that GBM with RMSE value of 15.32 was the most precise of the five individual models produced. While the SVM has an underperformed RMSE of 15.82, the SVM is perhaps the only system delivering a distinctive efficiency in the quantitative tests. The results of the simulations were consistent with a traditional Cox proportional risk model, which is used as a reference point. In order to inform the patient's decision in final analysis of these supervised learning strategies, SEER data were found to be used as a way of assessing the time for patient survival and that the findings of these technologies for this particular dataset may equate to those of conventional methods.

Deep learning is dependent on several covered layers on use of neural "deep" networks, which have understood relations between the top two strata and its down bound ties to all their lower characteristics were tested to determine the malignancy of its Lung nodules despite the computation of the morphology and structure characteristics [9]. For a strong faith network, they had hit the responsiveness rate of 73.40% and the specificity rate of 82.20%.

Deep Convolutional Neural Network CNNs is used to identify or label a medical image in some research papers. Diagnosed lung cancer in 2015 with a multiscal two-layer CNN [13] recorded 86.84% accuracy in [12] the CNN architecture, data set characteristics, and transfer learning factors were exploiting and extensively analyzing three significant and previously under studied factors.

Predictive methods for breast cancer's survival by a large dataset were built in [15] by the computational regression of 2 major data mining methods, artificial neural networks and Decision Trees. The impartial approximation of the three prediction models was measured by ten times the cross-validation methods for comparative analysis. Results indicate that the Decision Tree (C5), second most effective artificial 91.2 percent neural networks, and the 89.2 percent logistic regression models, are the best predictor of 93,6 percent accuracy for the holdout study. A study was conducted with predictive models for the survival of prostate cancer, using vector support machines (SVM) in relation to the three techniques.

In [16] SVM with Artificial Neural Networks and Decision-making Trees is identified in this case as the precision predictor (92,85% accuracy). Prostate cancer survival is also examined in context, including artificial neural networks, decision trees and logistical regression. In the segment, data on patients suffering from colon cancer were compared to predict survival and more accurate neural networks were determined.

In [19] the assembly of the 3 most effective classification methods leads to an ideal forecast and region under the ROC for colon cancer survival. Some studies have evaluated the survival of lung cancer patients by analyzing the SEER database using learning machinery, such as group class-based methods and SVM and logistic regression. Techniques to assess the probability of development of lung cancer in patients with certain symptoms have been analyzed in data classification techniques. Comparisons with the lung cancer data in were made with the C4.5 and Naïve-Bayes graders and 90% of the survival estimates were achieved. In [19,20], there was the establishment of a joint voting process with five Decisions to provide the best predictor of the precision and survival area of ROC lung cancer. In order to identify interesting association rules or correlation among a wide range of items, Association rules mining techniques have been used; different methods for extracting rules and standard criteria have been proposed to indicate the best way to choose the rules and optimize them based on a specific data set.

A pulmonary cancer rulebook was developed using automated technology in, some of which was redundant and manually removed on the basis of domain expertise. There were three factors: maximum branch factor, addition of a new branch, and the factor used to add a new branch. From the very beginning, the authors suggested a tree algorithm that uses the whole dataset and descends in depth to the data with a greedy approach. Each tree node was a segment and therefore a rule of association. The attributes were: age, birthplace, grade of cancer, diagnostic test confirmations, the most remote tumor extension, involvement with lymph nodes; type of operation performed, cause for no operation; order of operation and radiation; area of the lymphatic node surgery, cancer phase. The measurement of treatment effectiveness and surgery is a required result of a SEER dataset analysis, despite a lack of chemotherapy information in the dataset.

Treatment effectiveness was taken into account in [21]. The study investigated whether patients with lung cancer had survival or radiation for longer or for both. A Propensity Score was utilized which represents a dependent likelihood of treatment for a unit given a collection of covariates observed. Two methods were used, known as logistic regression and classification tree, for the assessment of score. As patients can be treated separately or together for surgery or radiation, the score for every class was calculated and the variables were then numbered. A mathematical collection was generated on the life expectancy and radiation combination, together with a grading tree to every cluster. The results demonstrated that consumers that didn't obtain radiation both with and without surgery have the longest survival time.

### 2.1. Analysis of ML applicability's in cancer

A comprehensive search of ML techniques in cancer sensitivity, recurrence and survival predictions was conducted. The PubMed, Scopus, has entered two online databases. Further review was required due to the large no of blogs which were found by the search queries Most of the studies use different input data types: Clinical, genetic, histological, imaging, socioeconomic, epidemiological or mixed details. According to [13] and their survey based on ML use of cancer prediction, we have seen a large rise in documents released in the last decade. Discuss, we selected from the first category of papers a representative list following a well-defined structure. We have selected such studies in order to prevent the desired effects, especially with the implementation of recognisable ML techniques and integrated data through heterogeneous data. Tables 1 indicate some publications in this study. Each suggested approach specifies the type of cancer, the ML treatment, number of patients, type of data and overall precision achieved. Each table corresponds to a particular scenario of study. i.e., prediction of cancer sensitivity, forecast of cancer occurrence and prediction of cancer survival. It should have been acknowledged to introduce the most precise predictive model here in publications which apply more than one ML technique to prediction. Different study projects have attempted to forecast cancer regeneration after remission and have managed to boost predictions correctly in contrast to alternative statistical techniques. In addition, molecular and clinical data have been used to estimate the large bulk of such papers. The implementation of observed behaviors such as input data is a growing phenomenon, based on the growth of HTTs.

### 2.2. Case study 1

Application of machine learning to predict the susceptibility of cancer risk from the 79 papers surveyed in this study are relative limited (only 3). The development of a retrospective methodology to predict the presence of 'spontaneous' breast cancer using single nucleotide polymorphism (SNP) steroid metabolizing enzymes (CYP 450) is among the interesting documents. Close. Sporadic and non-family breast cancers account for 90% (Dumitrescu and Cotarla, 2015). This trial was conducted with the theory that environmental toxins or hormones were accumulated in breast tissue and that some combinations of the SNP gene were at increased risk of breast cancer. The authors have obtained data on 63 breast cancer patients and 74 breast-free (controls) patients from the SNP (98 SNP from 45 cancer-associated Genes). It was vital to the progress of this research that researches used various methods to minimize a sample-per-feature ratio and analyzed several processes of machine training in order to find optimum classification. In particular, the authors rapidly reduced this set from a start set of 98 SNPs to just 2–3 SNPs that appeared to be as informational as possible. Instead of almost 3:2 (with all the 98 SNPs used), the specimen ratios were reduced to 45:1 (for 3 SNPs) and 68:1 (for 2 SNPs). This made it possible to prevent the "dimensionality curse" from being affected (Bellman, 1961; Somorjai et al., 2013). When the testing sample gets minimized, a number of machine learning techniques, consisting of a naïve Bayes model, various decision-making methods and a sophisticated SVM were applied. With just a set of three SNPs the SVM and naïve Bayes classifier were maximum in precision. The decision-tab classifier achieved maximum accuracy with a set of two SNPs. The SVM classification was the optimum, along

**Table 1**
Features and challenges of existing lung cancer prediction models.

| Author [citation] | Methodology | Features | Challenges |
|---|---|---|---|
| Tan et al. [1] | Adaboost | • Has attained high sensitivity and best performance.<br>• It is very simple to implement. | • It is very sensitive to noisy data. |
| Tae-WooKim et al. [2] | Decision Tree (DT) | • These are simple to interpret.<br>• It should be taken as the minimal decision standard of work-relatedness for lung cancer. | • They suffer from over fitting. |
| Maciej Zięba et al. [3] | Boosted SVM | • It is used in medical application for predicting post-operative life expectancy in lung cancer patients.<br>• It is used to solve the imbalanced data problems. | • The running time of training algorithms do not scale well with the size of the training set. |
| Worrawat Engchuan [4] | SVM | • It is used to build n- hyperlanes and n-features for dividing each different class apart from maximal margin. | • Many parameters need to be set accurately for attaining the best results. |
| H. Azzawi et al. [5] | Gene Expression Programming | Gene Expression Programming | • It has better solution for predicting lung cancer difficulties.<br>• Has high accuracy. |
| Panayiotis Petousis et al. [6] | Dynamic Bayesian Networks | Dynamic Bayesian Networks | • Has demonstrated high discrimination and predictive power.<br>• It is used to acquire the probability of positive outcome of a biopsy for the given individual. |
| Chip M. Lynch et al. [7] | DT | • It is the best predictor by attaining high accuracy.<br>• It automatically prunes to a very short three-level depth. | |
| P. Petousis et al. [10] | Partially- Observable Markov Decision Process (POMDP) | Partially-Observable Markov Decision Process (POMDP) | • It optimizes the lung cancer prediction during the improvement of test specificity.<br>• It reduces the false positive rates. |

with a precision of 69%, and 67% and 68%, respectively, were found in the naive Bayes and Decision Tree classification systems. The outputs are about 23–25% better than original. The extensive level of cross validation and confirmation conducted was another notable feature of this study. At least three ways have been validated for each model's predictive power. Firstly, model training with 20-fold cross validation has been evaluated and monitored. A bootstrap resampling approach was used when the cross validation is performed 5 times and the outputs were averaged to keep the stochastic dimension in the division of samples to a minimum. In addition, the selection process was carried out for 100 times in each fold (5 times for each of 20 folds) in order to reduce unequality in function selection (i.e. selecting the most informative SNP sub-ensemble). Thus, the outputs are then matched with an altered permutation test that, had 50 percent predictive precision. While the researchers tried to reduce the stochastics in sample partitioning, it could have been better to use leave-one-out cross-validation that shall have completely deleted this stochastical element. This trial was conducted with the theory that environmental toxins or hormones were accumulated in breast tissue and that some combinations of the SNP gene were at increased risk of breast cancer. The authors have obtained data on 63 breast cancer patients and 74 breast-free (controls) patients from the SNP (98 SNP from 45 cancer-associated Genes). It was vital to the progress of this research that researchers used various models to minimize a sample-per-feature ratio and analyzed several methods of machine training in order to find optimum classification. It also points out the wayin which machine learning can disclose significant information into the biology of spontaneous or non-famile breast cancer and polygenic risk factors.

### 2.3. Case study 2

Cancer Survival Prediction Almost half (or 1 year or 5 years survival rates) of all the machine study studies on cancer prediction. One report of a specific interest (Futschik et al., 2013) was used to predict the outcomes of diffused large-B-cell lymphoma (DLBCL) by hybrid machine learning. In particular, both clinical and genomic (microarray) information was gathered into creating one clinical classification to predict DLBCL survival. The approach is somewhat different from the Listgarten et al. study (2014) in its classification scheme, which only used genomic data (SNP). Futschik et al. hypothesized, rightly, that clinical knowledge may improve data on microarrays to a better output than a microarray-alone or clinical data-based classifier. In addition, different kinds of "Evolving Neural Network" (EFuNN) classifiers have been produced to manage genomic data, separate from the Bayesian classification system. A mixture of 17 genes from the microarray data is used by the best EFuNN classifier. The accuracy of this best EFuNN was 78.5%. In order to achieve consensus prediction, the EFuNN and the Bayesian classificators were mixed in a hierarchic modular structure. This hybrid classifier has a precision of 87.5 per cent, significantly improving both classifications' performance alone. This was 10% good than the excellent-performing classification for machine learning (77.6% by VMS). A cross-validation strategy for the EFuNN classifier was applied. Possibly because the sample was small. No external validation collection was present to check for the overall system, as with Case Study # 1. The Sample per Feature Ratio (SFR) is above 3 with just 56 patients (samples) categorized using 17 gene features. SFR below 5 do not always necessarily ensure a robust classification (Somorjai et al., 2013). Moreover, it is obvious that the researchers were known of this problem and spent enough time explaining in depth the internal functioning of their classifier to justify their approach. This consisted of a summary of how the Bayesian classification was constructed, how the EFuNN works and how the two classification systems cooperate to make one prediction. Also, the researchers tested the independence of the micro-array knowledge from clinical data and subsequently verified it. This eye for detail is particularly outstanding in such a study. The whole research reveals how well the capacity to use both genomic and clinical data significantly improves cancer prediction accuracy.

### 2.4. Case study 3

The Laurentian and the other in a particularly good example, and also discusses some of the inconveniences observed in existing researches. The authors wanted to predict the possibility of recurrence in patients with breast cancer for five years. Seven predictive variables have been combined, comprising of clinical

information like patient age, tumor size and no. of axilla metastases. Protein biomarkers, like oestrogen and progesterone receptor levels, also received information. The focus of the research was to produce an automated, quantitative predictive approach more precise than those of the classical metastasization of the tumor node (TNM). TNM is a group of medical experts that rely majorly on the professional judgment of a pathologist or clinician. The researchers used an ANN model, were using information from 2441 breast cancer patients (each time seven data points). A sample to feature ratio remained significantly higher than the recommended minimum of five (Somorjai et al., 2013). The whole dataset was divided into 3 classes: training (1/3), testing (1/3) and test sets (1/3). Furthermore, the authors have collected 310 separate samples from another organization to carry out an external assessment of breast cancer patients. This helped the researchers to test the generalization of their system out beyond ones institution — a phase not taken in the 2 experiments discussed above. The analysis demonstrates not only the volume of data and the thoroughness of validation, but also the level of quality control for data processing. The information, for example, was decided to enter and collected autonomously in a connection database and was autonomously checked to keep the referring doctors in good standing. The samples of 2441 patients and 17 000 data points were sufficiently large for a typical breast cancer population demographics when subdivided into the data sequence. However, by examining data distributions for patients in each set (training, monitoring, testing and external), the authors explicitly verified this assumption and demonstrated that distributions are quite same. The Authors built an extremely accurate and robust classifier through consistency and attention to detail Since the study's aim is to produce a system that better predicted re-currence of breast cancer than the traditional TNM stalemate method, comparing the ANN model with the TNM stalemate predictions was important. This was done by using an Operator Characteristic (ROC) curve to compare the performance. The ANN model (0.726), calculated by the portion in the ROC curve, exceeded the TNM system (0.677). This research is an brilliant illustration that machinery is well articulated and tested. A large enough set of data was obtained and data was tested for performance assurance and precision for each sample independently. In addition, blinded validation sets were available for assessing the generality of the machine learning system both from the original data set and through an external point. Finally, the precision of the model has been contrasted directly with that of the traditional TNM projection scheme. Thus the only challenge to this analysis was that the researchers evaluated only one form of ANN algorithm. Because of the type and the amount of data used, another machine learning technique can well have exceeded their ANN model.

## 3. Research gap

Lung cancer is the second largest human illness, which refers to deaths from cancer worldwide. The average survival rate of 5 years for patients with lung cancer in other organs such as the breast, cervix, bladder, prostate or colon does not exceed 14 percent, which is significantly less than the rate of patients with cancer [18]. Thus, early prediction of lung cancer is very important for the appropriate treatments for decreasing the deaths. In big data, healthcare is one of the significant sources. Accurate examination of healthcare information is mostly demanded for detecting lung cancer in an early stage. Multiple researches are being designing newly to recognize lung cancer with more quality using big data. As there is a necessity to classification approach for improving the detection accuracy with respect to time. In addition, machine learning techniques are modelled for enhancing the detection accuracy in big data. Specifically, lung cancer is not well known that means which kind of approaches will give high detection data and which data attributes must be employed for the detection purpose.

With the help of huge datasets, prediction methods for breast cancer survivability were introduced by implementing two famous data mining techniques such as Artificial Neural Network (ANN) and Decision Tree (DT) and also utilized a common statistical approach. For measuring the unbiased assessment of three detection models, ten-fold cross validation mechanisms were used for the performance comparison. The outcomes have proven that DT was the well performing classifier for predicting the disease with an accuracy of 93.6% on the holdout sample; ANN was standing the second best position with an accuracy of 91.2%. Similarly, logistic regression has attained the accuracy of 89.2%. A research was done in [20] for developing detection techniques to know the survivability of prostate cancer, using SVM along with that three methods that were mentioned earlier. Here, the outputs have revealed that the singled out SVM acquired high accuracy, next to that ANN and DT attained more accuracy. Moreover, prostate cancer survivability was examined by ANNs, DTs, and LR methods in [21]. Multiple methods were contrasted in [22] by SEER colon cancer patient dataset for predicting survival rate, and recognized that NNs were best for predicting the survival rate. Ensemble voting of three outperformed classifiers present in [23] was resulted in optimal prediction, and AU-ROC curve to colon cancer survival rate. In some researches, the survival of lung cancer patient was examined by evaluating the SEER database using machine learning algorithms, consisting of SVM, LR [24], unsupervised approaches [25], and clustering-based techniques [26]. In [27], data classification approaches were assessed for finding the chances of patients with definite indications for the growth of lung cancer. The performance of DT and NB classifiers were compared in [28], and implemented for lung cancer data acquired from SEER database. This attained approximately 90% precision in detecting the survival of patient. Ensemble voting of five DTs and meta-classifiers existing in [29] was resolute for acquiring the best prediction survival rate of lung cancer regarding precision and AU-ROC curve. Many challenges related to the machine learning algorithms are associated with manual training. The significant thing is complexity in accurate recognition of nature for pre- processing them correspondingly before subjected to machine learning algorithms. The time and the experts linked with this job were majorly high. According to the research, it was manifested that there is lack of consistency in detection accuracy of machine learning techniques over classical prediction techniques. With the present literature, this was made reliable. Many investigations that compared the machine learning models with classic statistical model have been confirmed that their outcomes were different.

Even though multiple strategies were utilized for predicting different types of diseases, the predictive models using the machine learning algorithms reported in the literal works are less for lung cancer detection with IoT integration. There is a high scope to implement more well- performing deep learning models that might produce best prediction outcome. Moreover, the enlarged availability of adequate historical data of patients has paved the way for the development of novel deep learning algorithms for lung cancer prediction. In addition, the optimization algorithms have the ability to improvise the deep learning models. However, there are few disadvantages such as it is not able to find the optimal solution to the problem defined, and it is complex to select parameters. Moreover, the benefit of PSO is its ability to solve the complex optimization problem. But, the convergence concept is not applicable. Some of the positives of SMO are useful for solving quadratic problems that occurs in the training of SVM, and also it reduces the memory storage. Yet it has to improve by introduc-

ing a new variant. The ability of machine learning to solve composite tasks with dynamic environment and knowledge has contributed to its success in prediction research especially lung cancer, enabled with novel met-heuristic algorithms.

Although there are many advantages for predicting the lung cancer, but still there are few defects with the existing methodologies so that a new method needs to be implemented. Adaboost [31] has attained high sensitivity and best performance, and it is very simple to implement. But, it is very sensitive to noisy data. DT [32] is simple to interpret, it should be taken as the minimal decision standard of work-relatedness for lung cancer, is the best predictor by attaining high accuracy, and it automatically prunes to a very short three-level depth. However, the running time of training algorithms do not scale well with the size of the training set. SVM [34] is used to build n-hyperlanes and n-features for dividing each different class apart from maximal margin, and it improves the classification power and robustness. Yet, many parameters need to be set accurately for attaining the best results. Gene Expression Programming [35] has the better solution for predicting lung cancer difficulties, and has high accuracy. However there are some disadvantages such as if they are easy to manipulate, they lose in functional complexity. Dynamic Bayesian Networks [36] has demonstrated high discrimination and predictive power, and it is used to acquire the probability of positive outcome of a biopsy for the given individual. Though there are few challenges like if there is longer search time, the performance might be affected. POMDP [38] optimizes the lung cancer prediction during the improvement of test specificity, and it reduces the false positive rates. But, the performance needs to be improved. Hence, the new model needs to be introduced for providing best performance so that the above conflicts are useful for the new development method.

## 4. Research objectives

The objective of this research work is discussed as follows.

1. To review on various state-of-the-art lung cancer prediction models and develop a new feature extraction model.
2. Compare the symptoms of cancer for early notification.
3. To design and develop a deep learning model to predict the lung cancer.
4. To validate the proposed model by comparing it with other conventional models.
5. Sending Automatic notification for detecting the cancer.

## 5. Discussion

In The latest research on predicting cancer using ML & DL techniques are discussed in this study. Further through the short details of the ML & DL field and the preprocessing data techniques, the selection techniques and the classification algorithms were employed, we discussed three specific case studies based on popular ML tools, concerning foretell of the susceptibility of cancer, cancer recurrence and cancer survival. Clearly, a huge number of ML & DL concepts released over the past decade produce precise outputs regarding particular cancer predictions. Moreover, it is crucial for the separation of clinical decisions to identify potential problems including experimental design, collecting suitable samples of data and validating classified results. Moreover, despite claims to have contributed to appropriate and efficient decision-making by the ML classification methods, very few have in fact entered clinical practice. Recent advances in omits technology have led us further to better understand a wide range of diseases, but validation results need to be accurate before signatures of gene expression shall be used in hospitals. Only a few marked samples in general. The small amount of data samples is a majorly frequent drawback observed in the research surveyed in this article. The size of training data sets that need to be large enough is a basic requirement in the use of classification schemes to model a disease. A relatively large dataset makes it possible to divide enough into training and trial sets and therefore to validate the calculators reasonably. A small training sample can result in misclassifications compared with the dimension of the data, while estimators can develop unstable and partial techniques. It's clear that a more wealthy group of patients could predict their survival may improve predictive model capacity. The quality of the dataset and the selection schemes are important for efficient ML and DL and then for precise cancer foretell except for data size. Using feature selection methods to select the maximum informative characteristics subset for training the technique could lead to sturdy models. Reproducible values are also characterized as characteristic sets consisting of histology and pathology studies. Given the lack of static entities, it is essential that a multiple feature sets are adapted to the ML & DL technology over time. We also discovered which SVM and ANN classifiers are commonly utilized for cancer forecasting results as one of the most frequently used ML algorithms [35]. As discussed in our introductory section, ANNs are widely used for nearly 30 years [40]. SVMs are also a newer method to cancer prediction but have already been widely included in their trustworthy predictive results. However, the selection of the best algorithm is dependent on a large number of parameters, which include data types collected, sample size, time limits and the type of prediction results. New methods for overcoming the above-mentioned limitations should be explored regarding the future of cancer modeling. More accurate results and reasoned conclusions would be obtained through efficient quantitative research of the heterogeneous data sages used. Further research on the basis of more public databases, which gather valid cancer data for all diagnosed patients, is needed. Their use by scholars will allow their modeling studies to generate relevant outputs and integrated clinical decision-making.

## 6. Conclusion

The whole study explains and compares the findings of various machine learning and in-depth learning implemented to cancer prognosis. Specifically, several trends related to those same kinds of machines techniques to be used, the kinds of training data to be incorporated, the kind of endpoint forecasts to be made, sorts of cancers being investigated, and the overall performance of cancer prediction or outcome methods have been identified. While the ANNs are common, it is clear that a broader variety of alternative learning approaches is also used to predict at least three different cancer types. ANNs continue to be prevalent. Furthermore, it is clear that machine training methods typically increase the efficiency or predictable accuracy of most pronostics, in particular when matched with conventional statistical or expert systems. Although most researches are usually excellently-designed and fairly validated, more focus is quite desirable for the planning and implementation of experiments, in particular with regard to quantity and quality of biological data. Improving the experimental design and the biological validation of several device classification systems would undoubtedly increase the general Quality, replicability and reproductivity of many systems. In total, we believe that the usage of the devices education & deep learning classificatory will probably be quite common in many clinical and hospital settings if the quality of study continues to improve.

The assimilation of multifaceted heterogeneous data, which can offer a promising tool for cancer infection and foresee the disease,

also demonstrates the incorporation in the application of different analytical and classification methods.

In future, by using the proposed framework, we would like to use other state of the art machine learning algorithms and extraction methods to allow more intensive comparative analysis.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References


[1] Chao Tan, Hui Chen, Chengyun Xia, Early prediction of lung cancer based on the combination of trace element analysis in urine and an Adaboost algorithm, J. Pharm. Biomed. Anal. 49 (3) (2009) 746–752.

[2] D.-H. Tae-WooKim, Chung-Yill Park, Decision tree of occupational lung cancer using classification and regression analysis, Safety Health Work 1 (2) (2010) 140–148.

[3] M. Zięba, J.M. Tomczak, Marek Lubicz, Jerzy Świątek, Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients, Appl. Soft Comput. 14 (2014) 99–108.

[4] Worrawat Engchuan, Jonathan H. Chan, Pathway activity transformation for multi-class classification of lung cancer datasets, Neurocomputing 165 (2015) 81–89.

[5] H. Azzawi, J. Hou, Y. Xiang, R. Alanni, Lung cancer prediction from microarray data by gene expression programming, IET Syst. Biol. 10 (5) (2016) 168–178.

[6] P. Petousis, S.X. Han, Denise Aberle, Alex A.T. Bui, Prediction of lung cancer incidence on the low-dose computed tomography arm of the National Lung Screening Trial: a dynamic Bayesian network, Artif. Intell. Med. 72 (2016) 42–55.

[7] C.M. Lynch, J.D. Behnaz Abdollahi, A. Fuqua, R. de Carlo, James A. Bartholomai, Rayeanne N. Balgemann, Victor H. van Berkel, Hermann B. Frieboes, Prediction of lung cancer patient survival via supervised machine learning classification techniques, Int. J. Med. Inf. 108 (2017) 1–8.

[9] D.S. Rao, D.P. Tripathy, Optimization of machinery noise using Genetic Algorithm. Noise Conference 2017. Michigan, 2017; 527–537.

[10] P. Petousis, A. Winter, W. Speier, D.R. Aberle, W. Hsu, A.A.T. Bui, Using sequential decision making to improve lung cancer screening performance, IEEE Access 7 (2019) 119403–119419.

[12] V. Krishnaiah, G. Narsimha, C. Subhash, Diagnosis of lung cancer prediction system using data mining classification techniques, Int. J. Comp. Sci. Inf. Technol. 4 (1) (2013) 39–45.

[13] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.

[15] L. Demidova, I. Klyueva, Y. Sokolova, N. Stepanov, N. Tyart, Intellectual approaches to improvement of the classification decisions quality on the base of the SVM classifier, Procedia Comput. Sci. 103 (2017) 222–230.

[16] N. Picco, R.A. Gatenby, A.R.A. Anderson, Stem cell plasticity and niche dynamics in cancer progression, IEEE Trans. Biomed. Eng. 64 (3) (2017) 528–537.

[18] Paweł Krawczyk, Tomasz Kucharczyk, Kamila Wojas-Krawczyk, Screening of Gene Mutations in Lung Cancer for Qualification to Molecularly Targeted Therapies, INTECH Open Access Publisher, 2012.

[19] A. Colquhoun, L. McHugh, E. Tulchinsky, M. Kriajevska, J. Mellon, Combination treatment with ionising radiation and Gefitinib ('Iressa', ZD1839), an epidermal growth factor receptor (EGFR) inhibitor, significantly inhibits bladder cancer cell growth in vitro and in vivo, J. Radiat. Res. 48 (5) (2007) 351–360.

[20] E. Adetiba, O.O. Olugbara, Lung cancer prediction using neural network ensemble with histogram of oriented gradient genomic features, Sci. World J. (2015).

[21] S.S. Alahmari, D. Cherezov, D.B. Goldgof, L.O. Hall, R.J. Gillies, M.B. Schabath, Delta radiomics improves pulmonary nodule malignancy prediction in lung cancer screening, IEEE Access 6 (2018) 77796–77806.

[22] S. Park, S.J. Lee, E. Weiss, Y. Motai, Intra- and inter-fractional variation prediction of lung tumors using fuzzy deep learning, IEEE J. Transl. Eng. Health Med. 4 (2016) 1–12.

[23] A. Raweh, M. Nassef, A. Badr, A hybridized feature selection and extraction approach for enhancing cancer prediction based on DNA methylation, IEEE Access 6 (2018) 15212–15223.

[24] J. Pati, Gene expression analysis for early lung cancer prediction using machine learning techniques: an eco-genomics approach, IEEE Access 7 (2019) 4232–4238.

[25] B. Zhang et al., Ensemble learners of multiple deep CNNs for pulmonary nodules classification using CT images, IEEE Access 7 (2019) 110358–110371.

[26] C. Arunkumar, S. Ramakrishnan, Prediction of cancer using customised fuzzy rough machine learning approaches, Healthcare Technol. Lett. 6 (1) (2019) 13–18.

[27] H. Guo, U. Kruger, G. Wang, M.K. Kalra, P. Yan, Knowledge-based analysis for mortality prediction from CT images, IEEE J. Biomed. Health. Inf. 24 (2) (2020) 457–464.

[28] J. Yang, N. Li, S. Fang, K. Yu, Y. Chen, Semantic features prediction for pulmonary nodule diagnosis based on online streaming feature selection, IEEE Access 7 (2019) 61121–61135.

[29] Raja Mohammad Taisir Masadeh, Basel A. Mahafzah, Ahmad Abdel-Aziz Sharieh, Sea lion optimization algorithm, Int. J. Adv. Comp. Sci. Appl. 10 (5) (2019) 388–395.

[31] A. Jemal, F. Bray, M.M. Center, J.J. Ferlay, E. Ward, D. Forman, CA A Cancer J. Clin., 61 (2), 69–90, 2011.

[32] D. Delen, G. Walker, A. Kadam, Predicting breast cancer survivability: a comparison of three data mining methods, Artif. Intell. Med. 34 (2) (2005) 113–127.

[34] D. Delen, N. Patil, Knowledge Extraction from Prostate Cancer Data, Proceedings of the 39th Annual Hawaii International Conference on, vol. 5, 2006.

[35] M. Hoogendoorn, L.M.G. Moons, M.E. Numans, R.-J. Sips, Utilizing data mining for predictive modeling of colorectal cancer using electronic medical records, in: International Conference on Brain Informatics and Health BIH 2014: Brain Informatics and Health, 2014, pp. 132–141.

[36] R. Al-Bahrani, A. Agrawal, A. Choudhary, Colon cancer survival prediction using ensemble data mining on SEER data, 2013 IEEE International Conference on Big Data, Silicon Valley, CA, pp. 9–16, 2013.

[38] C.M. Lynch, V.H.V. Berkel, H.B. Frieboes, Application of unsupervised analysis techniques to lung cancer patient data, PLoS One, 12 (9), 2017.

[40] N. Arshadi, I. Jurisica, Data mining for case-based reasoning in high-dimensional biological domains, IEEE Trans. Knowl. Data Eng. 17 (8) (2005) 1127–1137.


## *Further Reading*


[8] D.S. Rao, D.P. Tripathy, Optimization of machinery noise using Differential Evolution algorithm, Int. J. Min. Mineral Eng. 8 (4) (2017) 294–309.

[11] D.S. Rao, D.P. Tripathy, A Genetic Algorithm approach for optimization of machinery noise calculations. Noise Vibr. Worldwide. 2019 50(4): 112–123.

[14] David Meyer, Friedrich Leisch, Kurt Hornik, The support vector machine under test, Neurocomputing 55 (s 1–2) (2003) 169–186.

[17] W. Kim, K.S. Kim, J.E. Lee, D.Y. Noh, S.W. Kim, Y.S. Jung, M.Y. Park, R.W. Park, Development of novel breast cancer recurrence prediction model using support vector machine, J. Breast Cancer 15 (2) (2012) 230–238.

[30] Z.W. Huang, A. Mcwilliams, H. Lui, D. Mclean, S. Lan, H.S. Zeng, Near-infrared Raman spectroscopy for optical diagnosis of lung cancer, Int. J. Cancer 107 (6) (2003) 1047–1052.

[33] D. Delen, Analysis of cancer data: a data mining approach, Expert Syst. 20 (1) (2009) 100–112.

[37] D. Fradkin, I. Muchnik, D. Schneider, Machine Learning Methods in the Analysis of Lung Cancer Survival Data, DIMACS Technical Report, 2005.

[39] D. Chen, K. Xing, D. Henson, L. Sheng, A.M. Schwartz, X. Cheng, Developing prognostic systems of cancer patients by ensemble clustering, J. Biomed. Biotechnol. (2009).

[41] G. Dimitoglou, J.A. Adams, C.M. Jim, Comparison of the C4.5 and a naive bayes classifier for the prediction of lung cancer survivability, J. Comput. (2012).

[42] A. Agrawal, S. Misra, Ramanathan Narayanan, Lalith Polepeddi, Alok Choudhary, Lung cancer survival prediction using ensemble data mining on seer data, Sci. Program. 20 (1) (2012) 29–42.

[43] S.M. Agrawal, R. Narayanan, L. Polepeddi, A. Choudhary, A lung cancer outcome calculator using ensemble data mining on SEER data, Proceedings of the Tenth International Workshop on Data Mining in Bioinformatics, 2011.

[44] D.L. Tong, A.C. Schierz, Hybrid genetic algorithm-neural network: feature extraction for unpreprocessed microarray data, Artif. Intell. Med. 53 (1) (2011) 47–56.

[45] M.R. Mohebian, H.R. Marateb, M. Mansourian, Miguel Angel Mañanas, Fariborz Mokarian, A hybrid computer-aided-diagnosis system for prediction of breast cancer recurrence (HPBCR) using optimized ensemble learning, Comput. Struct. Biotechnol. J. 15 (2017) 75–85.

[46] M. Zięba, J.M. Tomczak, M. Lubicz, J. Świątek, Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients, Appl. Soft Comput. 14 (January 2014) 99–108.

[47] L.-J. Tang, J.-H. Jiang, H.-L. Wu, G.-L. Shen, R.-Q. Yu, Variable selection using probability density function similarity for support vector machine classification of high-dimensional microarray data, Talanta 79 (2) (2009) 260–267.

[48] H.-L. Chen, B. Yang, J. Liu, D.-Y. Liu, A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis, Expert Syst. Appl. 38 (7) (2011) 9014–9022.

[49] A.H. Chen, C.-H. Lin, A novel support vector sampling technique to improve classification accuracy and to identify key genes of leukaemia and prostate cancers, Expert Syst. Appl. 38 (4) (2011) 3209–3219.

[50] W. Zhong, R. Chow, J. He, Clinical charge profiles prediction for patients diagnosed with chronic diseases using multi-level Support Vector Machine, Expert Syst. Appl. 39 (1) (2012) 1474–1483.

[51] H. Choi, D. Yeo, S. Kwon, Y. Kim, Gene selection and prediction for cancer classification using support vector machines with a reject option, Comput. Stat. Data Anal. 55 (5) (2011) 1897–1908.
[52] W.H. Delashmit, M.T. Manry, Recent developments in multilayer perceptron neural networks, in: Proceedings of the 7th Annual Memphis Area Engineering and Science Conference (MAESC'05), pp. 1–15, 2005.
[53] G. Cybenko, Approximation by superpositions of a sigmoidal function, Math. Control Signals Syst. 2 (4) (1989) 303–314.
[54] M.C. Popescu, V.E. Balas, L. Perescu-Popescu, N. Mastorakis, Multilayer perceptron and neural networks, WSEAS Trans. Circuits Syst. 8 (7) (2009) 579–588.
[55] M.Z. Rehman, N.M. Nawi, Improving the accuracy of gradient descent back propagation algorithm (GDAM) on classification problems, Int. J. New Comp. Archit. Their Appl. 1 (4) (2011) 838–847.
[56] Z.-G. Che, T.-A. Chiang, Z.-H. Che, Feed-forward neural network straining: a comparison between genetic algorithm and back-propagation learning algorithm, Int. J. Innov. Comp., Inf. Control, 7 (10) 2011.
[57] M.F. Møller, A scaled conjugate gradient algorithm for fast supervised learning, Neural Networks 6 (4) (1993) 525–533.
[58] K. Gopalakrishnan, Effect of training algorithms on neural networks aided pavement diagnosis, Int. J. Eng., Sci. Technol. 2 (2) (2010) 83–92.
[59] I.B. Othman, F. Ghorbel, Stability evaluation of neural and statistical classifiers based on modified semi-bounded plug in algorithm, Int. J. Neural Networks Adv. Appl. 1 (2014) 37–42.
[60] L. Breiman, Arcingclassifiers, Ann. Stat. 26 (3) (1998) 801–849.
[61] G.P. Zhang, Neural networks for classification: a survey, IEEE Trans. Syst., Man Cybernetics Part C 30 (4) (2000) 451–462.
[62] Y. Freundand, R.E. Schapire, Experiments with a new boosting algorithm, in: Proceedings of the 13th International Conference on Machine Learning (ICML '96), pp. 148–156, Morgan Kaufmann, San Francisco, Calif, USA, 1996.