Prediction of Cancer Disease using Machine learning Approach

Hashim H, Anirudh C Gupta, Dr.N.Saranya,
Department of IoT and AI & ML, Nehru Arts and Science College, Coimbatore, India
Assistant Professor & Head, Department of IoT and AIML, Nehru Arts and Science College

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 lowrisk 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 trans-forms 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.

2. LITERATURE REVIEW

ChaoTan et al 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, 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.[3]Maciej Zie˛ba et al. introduced boosted SVM in 2014 which is dedicated to solving imbalanced results. The solution proposed combined the advantages of using ensemble classifiers with costsensitive 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[4]Worrawat Engchuan . 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. [5]H. Azzawi et al. 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. 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. of the DBNs were designed with a reverse style, and 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 [6]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 [7]M. Lynch et al. 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.

3. METHODOLOGY

1. Overview of Cancer Detection

Lung cancer remains one of the most lethal diseases globally, accounting for a significant percentage of cancer-related deaths. The survival rate for lung cancer is alarmingly low, with fewer than 14% of patients surviving five years after diagnosis. The prognosis of lung cancer significantly improves if the disease is detected in its early stages, which underscores the necessity for developing accurate diagnostic tools that can identify lung cancer at an early and treatable stage.

Machine learning (ML), particularly when applied to big data, has proven to be a promising approach for early cancer detection. Healthcare data, which includes patient history, demographic data, medical records, and imaging information, can be leveraged to develop predictive models. These models can identify risk factors and help clinicians with early diagnosis, potentially saving lives and improving survival rates.

2. Importance of Big Data in Healthcare

Big data refers to large and complex datasets that are generated from a variety of healthcare sources, including patient records, diagnostic results, medical imaging, genomic data, and even wearable health monitors. The size and complexity of these datasets often make it impossible to analyze them manually, which is where machine learning and data mining techniques come into play.

Healthcare is one of the most important sectors benefiting from big data, with machine learning models capable of analyzing vast amounts of historical patient data to uncover patterns and trends that might otherwise go unnoticed. These insights can then be used to inform clinical decisions and predict the onset and progression of diseases such as lung cancer.

The integration of big data analytics into the healthcare domain has led to the development of models that not only predict the likelihood of disease occurrence but also provide an estimation of the survival rates and treatment outcomes for patients.

3. Machine Learning for Cancer Detection

Machine learning algorithms have become instrumental in the field of medical diagnostics, offering several advantages over traditional statistical methods:

Supervised Learning: This involves training a model using labeled data, where the output (e.g., survival rate, presence of cancer) is known. Common supervised learning techniques used in cancer detection include:

Artificial Neural Networks (ANN): ANN models are designed to mimic the way the human brain processes information. They can capture complex, non-linear relationships between input features (such as patient demographics, biomarkers, or imaging data) and the target variable (e.g., cancer prognosis).

Decision Trees (DT): These models work by splitting the dataset into smaller subsets based on feature values, ultimately resulting in a tree structure that helps classify data points. DT is particularly popular because of its interpretability.

Support Vector Machines (SVM): SVM is a powerful classification technique that tries to find the optimal hyperplane to separate different classes in the dataset. It is especially useful for high-dimensional data such as gene expression data or medical imaging.

Logistic Regression (LR): LR is often used for binary classification tasks and helps estimate the probability that a patient will survive or succumb to the disease based on various predictive factors.

Unsupervised Learning: In the context of lung cancer, unsupervised learning techniques can be used to group patients based on similar characteristics or to identify hidden patterns that may not be explicitly labeled in the data.

4. Application of Machine Learning to Cancer Detection

Various studies have applied machine learning techniques to the detection and survival prediction of lung cancer. The SEER (Surveillance, Epidemiology, and End Results) database, a large repository of cancer-related data, is commonly used for training and validating these models. Several findings from previous research show how different classifiers and machine learning algorithms have been used to predict lung cancer survival rates and outcomes:

Decision Tree and Naive Bayes Comparison: Decision Trees (DT) and Naive Bayes (NB) classifiers were tested on lung cancer data from the SEER database. The DT classifier achieved an accuracy of approximately 90%, proving to be a reliable tool for predicting lung cancer survival.

Ensemble Models: Some studies explored the use of ensemble models, where multiple classifiers are combined to improve prediction accuracy. For instance, an ensemble approach involving five Decision Trees and meta-classifiers was shown to outperform individual models in terms of survival prediction, providing higher precision and better AU-ROC (Area Under the Receiver Operating Characteristic Curve) scores.

Comparison with Traditional Statistical Methods: Machine learning approaches consistently outperformed traditional statistical methods, such as linear regression and logistic regression, in terms of accuracy and robustness. While traditional methods rely heavily on assumptions about the data (e.g., normality), ML algorithms can better handle non-linear relationships and complex interactions between features.

5. Challenges in Machine Learning for Cancer Detection

Despite the promise of machine learning in lung cancer detection, several challenges must be addressed:

Data Quality and Preprocessing: Data preprocessing is a critical step in the machine learning pipeline. Raw healthcare data often contain missing values, noise, or irrelevant features, which can degrade the performance of predictive models. Cleaning, normalizing, and handling missing values are essential tasks to ensure that the data is ready for modeling.

Feature Selection: Identifying the most relevant features for predicting lung cancer is a complex task. Features can include demographic information, clinical history, lifestyle factors (e.g., smoking), imaging data, and molecular markers. Feature selection techniques, such as Recursive Feature Elimination (RFE) or mutual information, are often used to reduce dimensionality and improve model efficiency.

Overfitting: In machine learning, overfitting occurs when a model learns the training data too well, including noise and irrelevant patterns, which leads to poor generalization to new, unseen data. This is particularly common in deep learning models, which have a large number of parameters. Regularization techniques, such as L2 regularization or dropout, are used to combat overfitting.

Model Interpretability: For healthcare practitioners to trust and use machine learning models in clinical settings, the models need to be interpretable. Decision Trees are more interpretable compared to deep learning models, which are often considered black-box approaches. However, methods like LIME (Local Interpretable Model-Agnostic Explanations) are being developed to explain predictions from more complex models.

6. Optimization Techniques for Machine Learning Models

To improve the performance of machine learning models for lung cancer detection, various optimization techniques are employed:

Particle Swarm Optimization (PSO): PSO is a population-based optimization technique that mimics the social behavior of birds flocking. It is used to optimize the parameters of machine learning models by finding the best possible solution within the search space. PSO is particularly useful for tuning hyperparameters like learning rates, kernel parameters in SVM, or the number of hidden layers in an ANN.

Sequential Minimal Optimization (SMO): SMO is a specialized optimization technique used in training Support Vector Machines (SVMs). It helps in solving quadratic optimization problems that arise during the

training of SVMs, making it a crucial algorithm for large-scale data.

Genetic Algorithms (GA): Genetic algorithms are another form of evolutionary optimization technique that can be used to fine-tune the hyperparameters of machine learning models. These algorithms simulate the process of natural selection by evolving a population of candidate solutions over multiple generations.

7. The Role of Deep Learning in Cancer Prediction

Deep learning models, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), are becoming more common in healthcare applications. These models can process unstructured data such as medical images (e.g., CT scans or MRIs) and sequential patient data (e.g., time-series data from patient records). By leveraging large volumes of historical data, deep learning models can extract hierarchical features and patterns, making them ideal candidates for lung cancer detection.

8. Future Directions and Research Opportunities

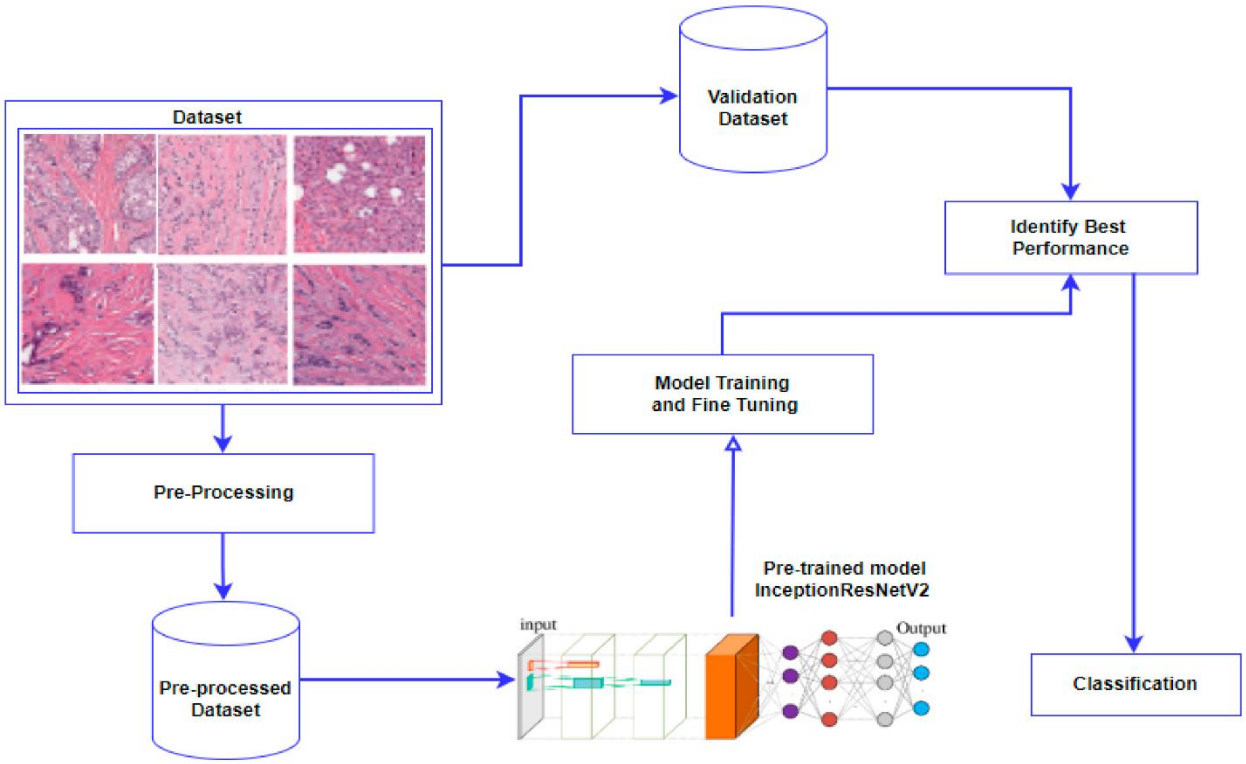
Integration of IoT Data: The advent of IoT devices, such as wearable health monitors, can provide real-time data that can complement traditional diagnostic methods. Machine learning models that integrate IoT data can predict lung cancer progression or help monitor a patient's health after diagnosis.

Hybrid Models: Combining multiple machine learning algorithms (e.g., SVM and ANN) in hybrid models could further improve prediction accuracy. These hybrid models can harness the strengths of different algorithms and compensate for the weaknesses of individual models.

Data Imbalance Solutions: Healthcare datasets often suffer from class imbalance, where the number of non-cancer cases far exceeds the number of cancer cases. Techniques such as SMOTE (Synthetic Minority Over-sampling Technique) can be used to address this issue and improve the performance of classifiers.

Personalized Medicine: The future of lung cancer prediction lies in the development of personalized models that take into account individual genetic profiles and treatment responses. By incorporating genomic data into ML models, predictions can be made more specific to each patient.

FrameWork For Cancer Detection



4. EXPERIMENTAL RESULTS

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 . As discussed in our introductory section, ANNs are widely used for nearly 30 years . 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 decisionmaking.

5. 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, F.J. Shaikh and D.S. Rao Materials Today: Proceedings 50 (2022) 40–47 45 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.

6. FUTURE WORK

Future research can significantly benefit from combining various data types, such as clinical data (e.g., age, smoking history, and family history) and image data (e.g., CT scans, MRIs). By integrating multi-modal data, predictive models can become more robust and accurate, providing a comprehensive view of a patient's health, which could lead to better early detection of lung cancer. Additionally, there is potential for improvement in feature extraction techniques; while Local Binary Pattern (LBP) has proven useful in image analysis, advanced methods like Gray-Level Co-occurrence Matrix (GLCM), Gabor filters, and deep learning-based feature extractors, such as convolutional neural networks (CNNs), can capture finer details in lung images, potentially improving model performance. A hybrid approach, combining models like Support Vector Machines (SVM) with algorithms like Random Forests or Neural Networks, could enhance predictive accuracy and robustness, and ensemble learning techniques like boosting or bagging may further improve performance. Developing real-time monitoring systems for lung nodule progression could also be a key area for future work. These systems could provide continuous updates, enabling timely medical interventions, and integrating wearable devices or telemedicine platforms would allow for dynamic monitoring. In terms of model sophistication, deep learning and transfer learning present exciting possibilities, especially when fine-tuning pre-trained models for specific lung cancer datasets, which is particularly useful when working with limited or imbalanced data. Moreover, the interpretability and explainability of models must be prioritized in medical applications to build trust among healthcare professionals. Techniques like SHAP and LIME could offer valuable insights into how predictions are made, improving transparency. To further refine lung cancer prediction models, conducting longitudinal studies and gathering diverse datasets spanning multiple years would help track the progression of the disease over time and identify early biomarkers, enhancing the model's generalizability. Additionally, extending models to support personalized cancer treatment plans—tailoring strategies based on patient-specific data, including genetic information—could have a major impact on outcomes. As the reliance on patient data grows, ensuring data privacy through compliance with regulations like HIPAA and GDPR is crucial, along with addressing ethical considerations, such as minimizing bias and ensuring fairness in predictions. Finally, collaboration with radiologists and oncologists is vital for validating and refining these models, ensuring they meet clinical needs, and enabling faster adoption and integration into medical practice.

REFERENCES

Books & Journals

[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] Maciej. Zie˛ba, J.M. Tomczak, Marek Lubicz, Jerzy S´wia˛tek, 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.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.