The Analysis of Anticancer Drug Sensitivity of Lung Cancer Cell Lines by using Machine Learning Clustering Techniques

Chandi S. Wanigasooriya, Malka N. Halgamuge, Azeem Mohammad
School of Computing and Mathematics
Charles Sturt University
Melbourne, Victoria 3000, Australia

***Abstract*—Lung cancer is the commonest type of cancer with the highest fatality rate worldwide. There is continued research that experiments on drug development for lung cancer patients by assessing their responses to chemotherapeutic treatments to select novel targets for improved therapies. This study aims to analyze the anticancer drug sensitivity in human lung cancer cell lines by using machine learning techniques. The data for this analysis is extracted from the National Cancer Institute (NCI). This experiment uses 408,291 human small molecule lung cancer cell lines to conclude. The values are drawn from describing the raw viability values for 91 human lung cancer cell lines treated with 354 different chemical compounds and 432 concentration points tested in each replicate experiments. Our analysis demonstrated the data from a considerable amount of cell lines clustered by using Simple K-means, Filtered clustering and by calculating sensitive drugs for each lung cancer cell line. Additionally, our analysis also demonstrated that the Neopeltolide, Parbendazole, Phloretin and Piperlongumine anti-drug chemical compounds were more sensitive for all 91 cell lines under different concentrations (p-value < 0.001). Our findings indicated that Simple K-means and Filtered clustering methods are completely similar to each other. The available literature on lung cancer cell line data observed a significant relationship between lung cancer and anticancer drugs. Our analysis of the reported experimental results demonstrated that some compounds are more sensitive than other compounds; Phloretin was the most sensitive compound for all lung cancer cell lines which were nearly about 59% out of 91 cell lines. Hence, our observation provides the methodology on how anticancer drug sensitivity of lung cancer cell lines can be analyzed by using machine learning techniques, such as clustering algorithms. This inquiry is a useful reference for researchers who are experimenting on drug developments for the lung cancer in the future.**

Keywords—Data analysis; clustering; filtered clustering; simple k-means clustering; cancer; lung cancer; cancer cell lines; drug sensitivity

I. INTRODUCTION

All around the world, cancer is the second leading cause of death. However, there is a significant challenge to prescribe the right drug for the right cancer patient. Using a large number of cancer patient reviews to prescribe anti-cancer drugs is neither effective nor practical. Therefore, several pharmaceutical companies, non-profit organizations, and non-government organizations have invested huge funds for the prevention, diagnosis, and treatment of cancers. For instance, the United States National Cancer Institute (NCI) [1], British Cancer Research Campaign (CRC) [2] and the European Organization for Research and Treatment of Cancer (EORTC) [3]. Besides, the melatonin has also been known as an effective agent that avoids both the initiation and promotion of cancer. Previous studies [4], [5] demonstrate the importance of disruption of melatonin due to exposure to weak electromagnetic fields, which may possibly lead to long-term health effects in humans.

A major goal of cancer researchers measures the effectiveness of anti-cancer drugs in pursuance to select the correct drug combinations based on their genetic and cell line structure of each patient, such as customizing medicinal products for each patient. Hence, to get a better understanding of the underlying cell lines with various cancer types are important. However, the methodology for converting the genetic measurements into predictive models to assist with therapeutic decisions remains a challenge.

Cancer can be developed anywhere in the human body. Human cells grow and break up to form novel cells when the body needs them [3]. Then the cells mature or turn into damaged ones, and die out, and novel cells get their position [6]. Cancer develops when this cycle breaks down. As cells become increasingly abnormal, matured or damaged cells stay alive as they normally should die; also, novel cells unnecessary develop as they are not required [1]. These additional cells can split without stopping that forms tumors and cysts. Normal cancer cells are different from standard cells in numerous ways. The abnormal cancer cell growth cannot be controlled. One major characterization is that they are less specialized than regular cells. While normal cells developed into very different cell types with detailed functions, cancer cells do not [2].

Most lung cancers originate in the lung carcinomas (epithelial tissue of the internal organs) and divide into non-small-cell lung cancer (NSCLC) [7], [8] and small-cell lung cancer (SCLC) [9]. SCLC is a critical type of lung cancer, caused by smoking and also responsible for diagnosing cases [10]. NSCLC records as the most common type as 85% of all lung cancers are this type [11]. There are three different subtypes of NSCLC [10], Adenocarcinomas (ADCA), Squamous Cell Carcinomas (SQ), and Pulmonary Carcinoids

(COID) [12]. ADCA is mostly described by the major production of mucus and SQ that usually occurs in larger bronchi [13].

In the United States (around 19.4%) [14]; in 2012, 1.56 million people died due to lung cancer [15], and 1.8 million related cases are reported [10]. In general, lung cancer does not build up on its own; however, it is caused by several factors. The environmental pollution also significantly contributes to the growth of this particular cancer. Smoking cigarettes are the most common and a major reason for lung cancer. By various approximations, smoking cigarettes causes around 86% of lung cancer, as well as caused by passive smoking (exposed to smoke exhaled by other smokers). The risks are even higher if a patient has started smoking tobacco at a young age. Passive smoking is not that dangerous; however, passive smokers have a 25% increased risk of lung cancer compared with people who are not exposed to the smoke of cigarettes [16]. Albeit, circumstances increase if a person is genetically disposed of or has exposure to asbestos materials, and past lung illnesses contribute to the risks as well. All these instances and circumstances can help the recent global growth of lung cancer. There is still no cure nor a suitable treatment for lung cancer confirmed, but there are ways to restore a patients' health [16].

Currently, lung cancer patients are treated with surgical and chemotherapy treatments. These treatments have made great aid in lung cancer; however, these treatments may bring serious long-term side effects. The main difficulty of the chemotherapeutic management of cancer is drug resistance. Anticancer drug resistance decreases the effectiveness of the drug and helps disease development [17]. This reason requires the development of new drug targeting strategies that can be used to improve the effects of drug resistance. The main purpose of cancer research is selecting the most effective drug combinations for each cancer patient based on their genetic structure and history. In recent cancer research, drug sensitivity prediction is mostly based on the genetic profile (gene expression measurements and genetic mutations). The advance of using genetic mutations is for expecting the cancer sensitivity is controlled by the present non-functional mutations as well as other hidden variables [18].

In late 1980's, the United States National Cancer Institute developed human cancer cell line anticancer drug screening. This screening model was rapidly recognised as a rich source of information about cancer cell line sensitivity [19]. A profile of cell line sensitivity offers data about the mechanisms of growth inhibition with cancer cell killing [11]. In current studies, genetic profiles of human cancer cell lines were treated with different drugs to allow predictive modeling of cancer drug sensitivity [18]. These cells are continuously divided and grow over time, under particular laboratory conditions [1]. Cancer cell lines (CCL) are used in many biomedical researchers to learn the biology of cancer as well as to ensure cancer treatments [20], [21]. Those are additionally used for different high-throughput applications and international mechanistic studies [22].

Discovering genetic modifications that aim to react to a particular therapeutic agent can help to improve cancer cell to produce a perfect cancer medicine. Cancer Cell Line profiling of small-molecule sensitivity has appeared as a balanced method to measure the connections between genetic or cellular features of CCLs and small-molecule reaction [23]. The Cancer Therapeutics Response Portal (CTRP) [24] analyzed a recognized pathway with major transmissions between degrees of difference gene dependency, and sensitive and non-sensitive cell lines. Recognized pathways and their parallel differential dependence networks are more considered to discover an important and precise mediator of cell line reaction to drugs or compounds [25]. They used a new and popular method that is the characterization of human cancer samples aligned with a series of cancer drug results that compare with genetic changes. It developed mainly from the attempts of the Cancer Cell Line Encyclopedia (CCLE) and Cancer Genome Project (CGP). Currently, different data mining and statistical methods will be used to evaluate drug responses of compounds with cancer cell lines [26].

Data Mining (DM) in medical research is an emerging application to observe the useful information and interesting patterns associated with different diseases. A professional DM method could be accepted as an analytical tool for efficient decision making [27], [28]. In DM, the clustering of dataset is more popular, and it has a broad range of applications. There are two types of clustering algorithm; descriptive (patterns and relationship with the available data) and predictive (calculate future aspect data values using the given data) clustering algorithms. Generally, in DM clusters and the analytical method [29] that discovers the unknown structures are fixed in dataset. Clustering is the process of creating groups of general objects into groups of similar objects. The application of DM, information discovery, machine learning techniques for health and medical data is challenging and exciting. The dataset is very complex, large, diverse and hierarchical and different in quality. The character of the data sometimes may not be the greatest for mining process, as the challenge is converting data into a suitable form.

In 2012, Roozgard, et al. suggested sufficient technique for early lung cancer detection and developed new predictive models for early detection of Non-Small Cell Lung Cancer (NSCLC) [30]. There is similar work that has been made to the genetic data about lung cancer. For instance, Cabrera, et al. identifies new molecular targets for drug design and chemotherapy. Lately, the success of this could be noted to increase or save the life of lung cancer patients [31]. Another study carried out in India (Dharmarajan and Velmurugan) has applied with two different lung cancer datasets with two different clustering algorithms. This study helps to develop the cluster analysis performed in the development of general medical application [32]. Palanisamy, et al. have analyzed the gene expression profile of leukemia dataset using the Weighted K-Means (WKM) algorithm [29], [33]. Information about the previous work done by different researchers in the relevant analysis between clustering algorithms and the review was described. The performance statistics of the different dataset for medical and some other related applications were discussed. The main focus of this research is to analyze lung cancer by using big data and DM clustering methods to find suitable medical applications in future.

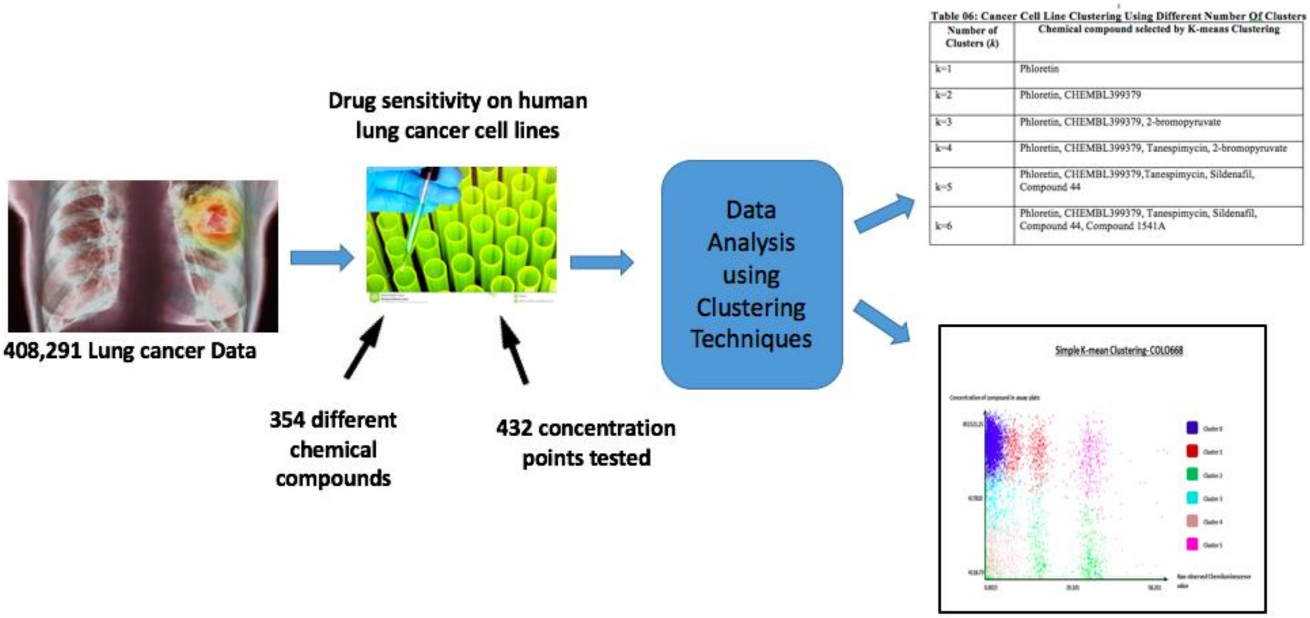
Drug sensitivity on human lung cancer cell lines
408,291 Lung cancer Data
354 different chemical compounds
432 concentration points tested
Data Analysis using Clustering Techniques
Simple K-mean Clustering-COLO668
Concentration of compound in assay plate
Raw observed Chemiluminescence value
Cluster 0
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5

Table 06: Cancer Cell Line Clustering Using Different Number Of Clusters

Number of Clusters (*k*)	Chemical compound selected by K-means Clustering
k=1	Phloretin
k=2	Phloretin, CHEMBL399379
k=3	Phloretin, CHEMBL399379, 2-bromopyruvate
k=4	Phloretin, CHEMBL399379, Tanespimycin, 2-bromopyruvate
k=5	Phloretin, CHEMBL399379,Tanespimycin, Sildenafil, Compound 44
k=6	Phloretin, CHEMBL399379, Tanespimycin, Sildenafil, Compound 44, Compound 1541A

Fig. 1. Graphical abstract (micro abstract).

This paper presents the application of Simple K-means clustering and filtered clusters to predict anticancer drug sensitivity in Small-Molecule Cancer Cell-Line Sensitivity Profiling Data. This research helps to develop the performance of cluster analysis in the general medical application development. The major purpose of this is to support the important method in finding the cluster of the lung cancer dataset. Moreover, this analysis shows the flexibility of dataset for cluster analysis in the medical field.

The paper is organized as follows (Fig. 1): Section II describes materials and methods and introduces the selection criterion of choosing dataset of simulation of the experiments. Then it follows with the data analysis with two types of cluster g techniques of Simple K-means clustering and Filters clustering techniques. In Section III the collection of results from data clustering finalized by the presentation of all clustered data is displayed. Section IV includes a discussion of the results and findings of drug sensitivity for each cell line. Section V, in brief, concludes the analysis of simulated test and opens up limitations for possible future work in this direction on the same topic.

II. Material and Methodology

This framework includes five major steps: Raw dataset collection, Data inclusion criteria, Dataset preparation, Data analysis, and Statistical analysis.

A. Raw Dataset Collection

The raw dataset chosen for this experimental simulation test was obtained from the National Cancer Institute in USA government and the dataset published in 2013 [13]. The dataset contains details about Small-Molecule Cancer Cell-Line Sensitivity Profiling Data used to identify cancer genes and lineage dependencies targeted by small molecules. This dataset is the combination of raw viability values for each cancer cell line treated with different compounds for each concentration point tested for each replicate is tested.

B. Data Inclusion Criteria

This analysis only used lung cancer raw viability data (Instances 408,291), and it filtered it by the use of contextual cancer cell line information and annotation data file.

TABLE I. Raw Viability Data Description for Selected Attribute

Attribute Name	Data Type	Description
ccl_name	Nominal	Primary name of cancer cell line
cpd_name	Nominal	Name of compound (INN preferred; best available otherwise)
cpd_conc_umol	Numeric	Final micromolar concentration of compound in assay plate
raw_value	Numeric	Raw observed chemiluminescence value

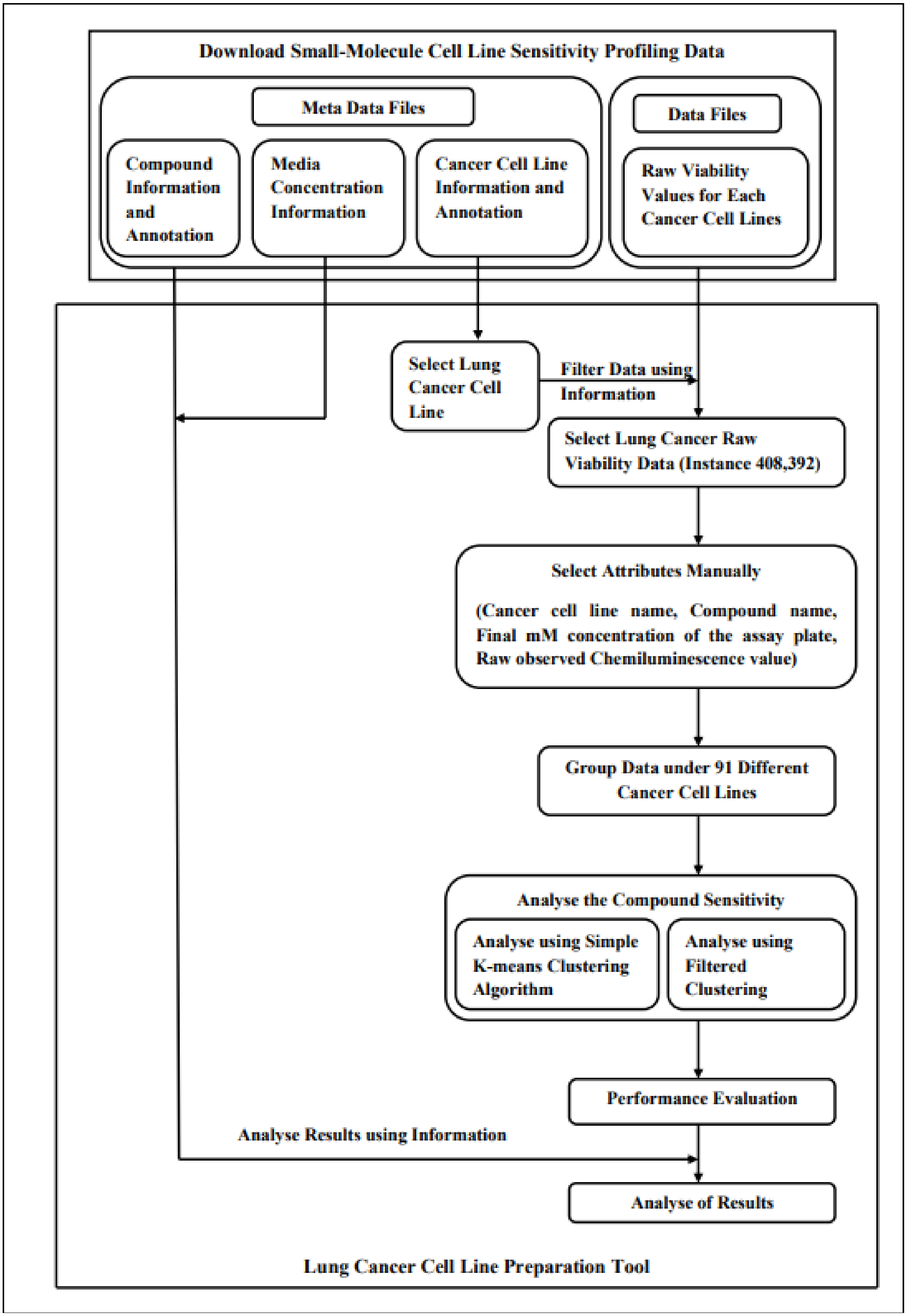
Download Small-Molecule Cell Line Sensitivity Profiling Data
Meta Data Files
Compound Information and Annotation
Media Concentration Information
Cancer Cell Line Information and Annotation
Data Files
Raw Viability Values for Each Cancer Cell Lines
Select Lung Cancer Cell Line
Filter Data using Information
Select Lung Cancer Raw Viability Data (Instance 408,392)
Select Attributes Manually
(Cancer cell line name, Compound name, Final mM concentration of the assay plate, Raw observed Chemiluminescence value)
Group Data under 91 Different Cancer Cell Lines
Analyse the Compound Sensitivity
Analyse using Simple K-means Clustering Algorithm
Analyse using Filtered Clustering
Performance Evaluation
Analyse Results using Information
Analyse of Results
Lung Cancer Cell Line Preparation Tool

Fig. 2. Lung cancer cell line preparation tool.

Filtered data include the primary name of cancer cell line, the name of the compound, replicate serial number, identifier for compound stock plate map in Broad Institute (LIMS), good location on assay plate, compound or vehicle or positive control, final micromolar (mM) concentration of the compound in assay plate, raw observed Chemiluminescence value and logarithm (base 2) of raw observed Chemiluminescence value [6] (Table 1). The selected lung cancer dataset contains 91 cancer cell lines and 354 different concentration points.

C. Dataset Preparation

This analysis, only considered Lung cancer raw viability data from NCI. Once the data is downloaded, the dataset was fully unreadable, and it was prepared to determine meaningful result to observe a drug for lung cancer that can be used in future medical applications. Data preparation depends on the dataset that is important to get a correct result. For this analysis, we used Lung Cancer Cell Line Preparation Tool (LCCLPT), which is shown in Fig. 2. This tool is composed of six main processes, namely, 1) select lung cancer raw viability data; 2) select attributes manually; 3) group under 91 different cells lines; 4) analyze the compound sensitivity using Simple K-means and Filtered clustering algorithms; 5) performance evaluation; and 6) analyzed through information given from NCI. Firstly, the attributes selected from raw datasets; therefore, some attributes were removed because they were not related to the further analysis. Only the used attributes were cell line name, compound name, compound concentration, and raw value. In the next group, the lung cancer data are under 91 different cancer cell lines. Each cell line is treated with 354 numbers of different chemical compounds.

According to Fig. 2 of LCCLPT, there are three main steps for the data analysis. These three steps are: Data Selection, Data Preparation and Analyze Compound Sensitivity using K-means Clustering. Therefore, following three different algorithms has written for those main steps. All these three algorithms are input patterns in the LCCL data analysis using K-means Clustering.

Algorithm 1: Data Selection

```
 string [] SelectAttribute = Select Attribute for the Data Selection
  string [] SelectLCCLNames = Select Lung Cancer Cell Line Names
  load a Meta Data of Cancer Cell Lines Information and Annotation
  select Lung Cancer using Filter Algorithm
  determine SelectAttribute for Select LCCL Names manually
  compute the SelectLCCLNames performing Data Selection using SelectAttribute
  save SelectedLCCLNames [n=91]

then
  string [] FilterAttribute = Filter Attribute for the Data Seperation
  string [] FilterLCCLRawViabilityData = Filter LCCL Raw Viability Data
  load a Data File of Raw Viability Values for CCL
  filter LCCL using Data Selection Algorithm [SelectedLCCLNames]
  determine FilterAttribute for Filter LCCL Raw Viabiity Data
  save FilteredLCCLRawViabilityData [n=408,392]
```

Algorithm 2: Data Preparation

```
 string [] SelectAttribute = Select Attribute for the Data Seperation
  string [] SelectAttriNames = SelectLCCLName,CpdName,CpdConcUmol,RawValue
  load a FilteredLCCLRawViabilityData File
  select SelectAttriNames for Seperate LCCL Raw Viability Data manually
  save SelectedAttriNames

then
  divide FilteredLCCLRawViabilityData using SelectedLCCLNames
  seperate FilteredLCCLRawViabilityData under SelectedLCCLNames
  save SeperatedFilteredLCCLRawViabilityData
```

Algorithm 3: Compound Sensitivity Analysis using K-means Clustering

```
  string [] ClusterAttribute = Cluster Attribute for the Data Analysis
  string [] CpdSensitivitySelectbyClustering = Compound Sensitivity Select by Clustering
  string [] ClusterCpdName = The most sensitive compond for the LCCL
  int k = Counter for number of attributes
  int MostSensitiveCpdSelectbyClustering = Counter for Most Sensitive Compound Selected by Clustering

  load a SaperatedFilteredLCCLRawViabilityData
  compute Sensitive Compound Clusters using K-means Algorithm
  determine Attributes for Compound Name Clustering using Attribute Selected LCCL
  else
    ClusterAttribute = Attribute selected manually
  end if
  while (k=NumberofAttributes) do
    if k=1 then
    while (k=NumberofAttributes) do

MostSensitiveCpdSelectedbyClustering =
MostSenstiveCpdSelectedbyClustering+1
    end while
  define string
[MostSensitiveCpdSelectedbyClustering]Cpd
SelectedbyClustering
  end if
    string
[SensitiveCpdSelectedbyClustering]
CpdSelectedbyClustering = Compound
Selected by K-   means Clustering
Algorithm using ClusterAttribute
  end while
    k=0 then
    for (k=Numberof Attributes)
        get the Most Sensitive Compound
in CpdSelectedbyClustering
  end for
    ClusteredCpdName = Most Sensitive
Compound of each LCCL
  Save ClusteredCpdName
```

D. Analysis of Raw Data

After dividing, the lung cancer data follow two clustering methods (Simple K-mean and Filtered) to calculate final cluster centroids using a changing number of clusters (k=*1* to *k=6*). To analyze both Filtered and Simple K-mean clustering results, one needs to tabulate data separately.

Waikato Environment for Knowledge Analysis, version 3.8 has been used to carry out the analysis on a computer with an AMD Quad-Core A4-6210 APU with AMD Radeon R3 Graphics, 1.8 GHz, and Random Access Memory (RAM) was 4GB. It allows users to analyze the data from many different dimensions or angles, categorizes, and summarizes the relationships identified. It contains Clusters for finding groups of similar instances in a dataset. In this paper, we used lung cell line data and analyzed the data with Filtered and Simple k-means clustering scheme.

1) Simple K-mean Clustering

The Simple K-means algorithm is one of the simplest unsupervised learning algorithms that answer the well-known clustering problem [12]. The procedure follows a simple and the calm method to classify a given dataset. Through some clusters (assume k clusters) static a priori. The K-means algorithm can run multiple times to decrease the complexity of grouping data.

2) Filtered clustering

The Filtered Cluster algorithm is using K-means analyzes algorithm data [12]. This procedure also follows a simple method to run multiple times to decrease the complexity of grouping data. By Using Filtered clustering and Simple K-means clustering, we analyzed all lung cancer cell lines separately. The clustering was performed based on selected preparation parameters. Each clustering method used a different number of clusters; however, it used the same number of attributed for both clusters.

E. Statistical Analysis

It is essential to accomplish a statistical hypothesis testing by calculating the probability value (*p*-value) to statistically prove that the selected chemical compounds are sensitive to lung cancer cell lines and this value should be less than 0.05 (p-value < 0.05). *P*-value is the probability of gaining an outcome similar to or extreme than what was observed when the null hypothesis is true. It was calculated by using the application IBM® SPSS® Statistics 20 which were designed for hypothesis testing.

III. Results

The clustering aims to identify cancer cell line. The most sensitive compound for each cell line is to figure out this way and also, study the connection between compound concentration and drug dosage. After clustering data, the results show that some compounds are more sensitive than other compounds.

In the first step of the LCCLPT, the selection required attributes of the dataset for the clustering, such as cancer cell line name, compound name, final micromolar (mM) concentration of the compound assay plate, raw observed chemiluminescence value. In the section of this analysis, we also analyzed data using the contextual compound information and annotation and list of all media components and concentrations of data files. The combination of the data for this study shows a statistically significant difference in various parameters for different lung cancer cell lines. This analysis showed a statistically significant difference in 91 different lung cancer cell lines.

The important parameters considered for the selection, clustering methods are a number of iterations, within the cluster sum of squared errors, and the time taken to build the model. Table 2 shows the number of iterations and the sum of squared errors that were computed using K-Mean clustering algorithm (*k=6)* and Filtered clustering algorithm for COLO668.

Using these two clustering algorithms (Simple K-means (SKM) and Filtered Cluster) cancer cell line dataset was clustered. The Clustering analysis and the results are illustrated in Table 3. These two analyzes of clustered compounds resulted in totally similar with each other, such as final selected compound name, final concentration, and raw value. Therefore, further analysis was carried out as analyzing these clustering results were based on compound name and final micromolar concentration.

TABLE II. COLO668 CLUSTERING USING DIFFERENT ALGORITHM

Clustering Algorithm	**Time Taken To Build Model**	**Number of Iterations**	**Within Cluster Sum Of Squared Errors**
Simple K-mean (*k=6*)	0.5 seconds	37	4557.349655
Filtered Cluster	0.06 seconds	9	4778.215454

TABLE III. FINAL ANALYSIS RESULTS FOR SIMPLE K-MEAN CLUSTERING

Cancer Cell Line Name	**Number of Instances**	**Final Compound Name**	**Final Concentration**	**Raw Value**
A549	9504	Phloretin	8.5996	1558371.863
BEN	3216	Phloretin	8.6047	1258215.255
CAL12T	4592	Piperlongumine	29.6638	1690291.675
CALU3	3360	Phloretin	8.5903	1648032.229
CALU6	10080	Phloretin	8.5903	1504356.098
CHAGOK1	3360	Phloretin	8.5903	751170.5283
COLO668	4640	Phloretin	7.9464	521638.3879
COLO669	4592	Piperlongumine	29.6638	1258651.038
CORL23	4704	Piperlongumine	29.4864	1695297.647
CORL279	3360	Phloretin	8.5903	1798751.68
CORL51	4254	Parbendazole	31.2378	421143.9187
CORL88	3360	Phloretin	8.5903	571504.6964
DMS273	3360	Phloretin	8.5903	2449319.152
DV90	3216	Phloretin	8.6047	750161.187
ECB1	3072	Phloretin	8.6046	1180258.701
EPLC272H	3360	Phloretin	8.5903	982414.5714
HARA	3216	Phloretin	8.6047	1529582.657
HCC1195	4592	Piperlongumine	29.6638	1352283.998
HCC1359	4592	Piperlongumine	29.6638	1034469.312
HCC15	5104	Piperlongumine	28.1711	1171832.382
HCC1833	4592	Piperlongumine	29.6638	696267.3389
HCC2108	4592	Piperlongumine	29.6638	2145425.478
HCC2935	4592	Piperlongumine	29.6638	844840.1045
HCC33	3216	Phloretin	8.6047	777908.6015
HCC4006	4592	Piperlongumine	29.6638	1377445.013
HCC44	3216	Phloretin	8.6047	1965327.553
HCC78	3072	Phloretin	8.6046	1508638.626
HCC827	4592	Piperlongumine	29.6638	2018265.432
KNS62	3360	Phloretin	8.5903	1622770.036
LC1SQSF	3216	Phloretin	8.6047	1055932.551
LCLC103H	3216	Phloretin	8.6047	2032721.894
LCLC97TM1	3360	Phloretin	8.5903	1103119.741
LU65	3216	Phloretin	8.6047	4977184.126
LU99	3072	Phloretin	8.6046	1446650.282
LXF289	4540	Phloretin	7.9464	1130450.536

NCIH1048	3360	Phloretin	8.5903	372769.0863
NCIH1105	4592	Piperlongumine	29.6638	367399.399
NCIH1299	3360	Phloretin	8.5903	1804504.139
NCIH1355	3360	Phloretin	8.5903	1277084.392
NCIH1373	3072	Phloretin	8.6046	1277897.656
NCIH1435	4640	Phloretin	7.9464	946515.0474
NCIH1437	6432	Phloretin	8.5972	1408976.654
NCIH1568	6432	Phloretin	8.5972	472604.4108
NCIH1573	3360	Phloretin	8.5903	785638.7135
NCIH1666	3360	Phloretin	8.5903	930566.119
NCIH1694	4592	Piperlongumine	29.6638	1105512.807
NCIH1755	5056	Piperlongumine	28.3597	1267151.681
NCIH1781	5056	Piperlongumine	28.3597	929863.4207
NCIH1792	6432	Phloretin	8.5972	1911963.046
NCIH1793	3072	Phloretin	8.6046	901567.8776
NCIH1836	4704	Piperlongumine	29.4864	272683.3759
NCIH1876	4592	Piperlongumine	29.6638	544708.1043
NCIH1915	4640	Phloretin	7.9464	1661567.424
NCIH1930	4640	Phloretin	7.9464	733179.0226
NCIH1944	6432	Phloretin	8.5972	655568.961
NCIH1963	4254	Parbendazole	31.2378	84295.3473
NCIH1975	6432	Phloretin	8.5972	1670169.267
NCIH2009	15392	Neopeltolide	7.4405	997846.0259
NCIH2023	4592	Piperlongumine	29.6638	1719536.76
NCIH2029	3360	Phloretin	8.5903	361672.1972
NCIH2030	4592	Piperlongumine	29.6638	1823878.366
NCIH2073	4704	Piperlongumine	29.4864	1863353.003
NCIH2081	4592	Piperlongumine	29.6638	968291.4362
NCIH2110	4704	Piperlongumine	29.4864	20503.3923
NCIH2122	5104	Piperlongumine	28.1711	1171498.339
NCIH2126	3216	Phloretin	8.6047	863556.1684
NCIH2141	4254	Parbendazole	31.2378	968225.1034
NCIH2172	4592	Piperlongumine	29.6638	1824119.729
NCIH2286	4640	Phloretin	7.9464	1493044.191
NCIH23	6432	Phloretin	8.5972	731529.207
NCIH2342	6207	Neopeltolide	25.5768	918200.7836
NCIH2405	3360	Phloretin	8.5903	1676250.926
NCIH3255	4704	Piperlongumine	29.4864	131478.6038
NCIH358	3072	Phloretin	8.6046	1520317.113
NCIH441	5104	Piperlongumine	28.1711	1341018.991
NCIH460	5104	Piperlongumine	28.1711	2059679.053
NCIH522	3072	Phloretin	8.6046	1048625.351
NCIH596	5104	Piperlongumine	28.1711	1538213.902
NCIH650	3360	Phloretin	8.5903	2372745.501

NCIH661	3216	Phloretin	8.6047	2055989.346
NCIH727	3216	Phloretin	8.6047	939856.79
NCIH810	6576	Phloretin	8.5973	1374307.438
NCIH82	3360	Phloretin	8.5903	2982554.595
NCIH841	4592	Piperlongumine	29.6638	1828496.5
RERFLCKJ	4640	Phloretin	7.9464	1613427.183
SCLC21H	4254	Parbendazole	31.2378	588623.6477
SHP77	3216	Phloretin	8.6047	1008581.497
SKLU1	5056	Piperlongumine	28.3597	1254689.585
SQ1	4592	Piperlongumine	29.6638	1469531.774
T3M10	5104	Piperlongumine	28.1711	1057457.328
VMRCLCD	3360	Phloretin	8.5903	986669.3973

Table 4 shows the final cluster results for Simple K-mean and Filtered cluster analysis for COLO668 cancer cell line. Both analyzed results were similar. Fig. 3 illustrates clustering visualizations of the Simple K-means clustering algorithm for COLO668 lung cancer cell line. We also observed the similar cluster visualization using the same cell line for Filtered Clustering. Both visualizations show similar results for two different clustering methods.

TABLE IV. COLO668 CELL LINE CLUSTERING FOR DIFFERENT ALGORITHMS

Clustering Algorithm	Simple K-mean	Filtered Cluster
Number of Instances	4640	4640
Final Compound Name	Phloretin	Phloretin
Final Micromolar Concentration	7.9464	7.9464
Final Raw Value	521638.3879	521638.3879

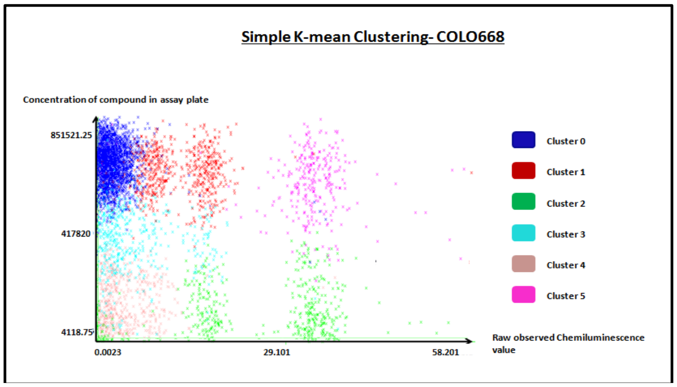

Fig. 3. Simple K-means Cluster Visualization *(K=6)* Lung Cancer Cell Line COLO668

Table 5 shows the analyzed cluster results according to the name of the final chemical compound for all cancer cell lines. It is clearly visible that a huge number of cell lines were most sensitive for Phloretin, it is about 53 out of 91 cancer cell lines resulted in Lung cancer (p-value < 0.001). Other three compounds are less than 33 cell lines. Therefore, according to the cluster results, it shows Phloretin is at the top of the compound list.

Most sensitive compounds for particular cancer cell lines, K-means clustering algorithm was used in the cell line dataset. Therefore, the numbers of clusters (k) were changed from 1 to 6 as there are six attributes in the dataset as seen in Table 6 (p < 0.001). According to those results, Phloretin is seen in all the clusters.

According to this information (Table 3), each cancer cell line shows significant information about the amount of final micromolar (mM) concentration of a particular compound. As shown in each compound it had a particular range of the concentration amount for each cell line. It is shown in Table 7.

TABLE V. NUMBER OF CELL LINES IN EACH COMPOUND

Compound Name	Number of cancer cell lines
Neopeltolide	2
Phloretin*	53
Parbendazole	4
Piperlongumine	32

TABLE VI. CANCER CELL LINE CLUSTERING USING DIFFERENT NUMBER OF CLUSTERS

Number of Clusters (*k*)	Chemical compound selected by K-means Clustering
k=1	Phloretin
k=2	Phloretin, CHEMBL399379
k=3	Phloretin, CHEMBL399379, 2-bromopyruvate
k=4	Phloretin, CHEMBL399379, Tanespimycin, 2-bromopyruvate
k=5	Phloretin, CHEMBL399379, Tanespimycin, Sildenafil, Compound 44
k=6	Phloretin, CHEMBL399379, Tanespimycin, Sildenafil, Compound 44, Compound 1541A

TABLE VII. ANALYZED FINAL CONCENTRATION RANGES FOR PARTICULAR CELL LINES

Compound Name	Range of micromolar (Mm) Concentration
Parbendazole	31.2378
Phloretin	7.9464 ~ 8.6047
Piperlongumine	28.1711 ~ 29.6638
Neopeltolide	7.4405 ~ 25.5768

Our analysis of the clustered result suggests that significant studies on lung cancer cell lines indicate that biologically each cell line is sensitive to a particular compound, as this is considered in both figures as they can overlap with each other. Also, studies on lung cancer cell lines show biologically or genetically changes due to the changes of an anticancer drug observation which should be further analyzed with more studies in the future.

IV. Discussion

A human cancer cell line mainly represents cancer biology. The anticancer drug discovery in basic experimental directions worldwide and detailed research studies had different results for high-throughput applications [34]. The Cancer Cell Line (CCL) sensitivity profiling can develop a new patient-matched therapy, that and only needs to be confirmed. Several types of medical researchers reviewed for small-molecule treatment in CCL models react differently to cancer cells. As believed, small-molecule and CCL models can be fully controlled through cancer cells by effective analysis methods and sensitivity profiling studies [23]. With this reason, we measured the Small-Molecule Cancer Cell-Line Sensitivity Profiling Data to identify sensitive drug or compound for each CCL.

This study has used Small-Molecule Lung Cancer Cell-Line Sensitivity Profiling Data datasets and input dataset that contains an experimental observation of 408,291 instances (or) records, and it grouped them under 91 different CCLs (shown in micro abstract Fig. 1). When we considered the attributes of data samples, there is a connection between the concentration of the compound assay plate and raw observed Chemiluminescence value of the lung cancer data. Researchers have measured Chemiluminescence raw observed value with different concentrations of anticancer drugs in lung cell line plate. One lung cancer cell line is treated with more than 10 different concentrations to increase the accuracy of the research. The raw value of Chemiluminescence might vary with the concentration of the compound assay plate of the lung cancer cell line [1]. According to the results, every compound had a particular concentration range. The lung cancer cell line has a particular concentration value, and each cell line has particular cancer cell line histology type. Therefore, critical histology needs of different concentration of anticancer drugs are observed [35]. The outcome of this research is useful for the Department of Medical application development and especially for lung cancer dataset analysis at the National Cancer Institute [1]. This research can be used in the future for similar types of analysis of lung cancer data in cancer institutions [32].

This study uses machine learning, clustering technique to cluster algorithms of Simple K-means and Filtered clustering. Using two clustering algorithms measured, anticancer drug sensitivity of small molecule Lung cancer cell line. There are three types of machine learning patterns found in this technique; supervised, semi-supervised and unsupervised [36]. Usually, a machine learning technique has been used for medical and many scientific studies to predict drugs [3], [6], [8], [9]. Reason for mostly used K-means clustering technique is data reduction and has better media accuracy [14]. Usually, the time taken will be different from the type of processor used. This research was proposed in grouping the requirements were a large number of requirements are divided into small groups which can be easily analyzed and grouped. The performance of the separation based on algorithms was analyzed using only the selected four attributes from the total number of attributes of the input dataset. For instance, K-means algorithm has been used for leukemia gene expression datasets, to predict the disease [3], [12].

This study is also supported by other studies [18], [20], [22] as they also have used Simple K-means clustering and other clustering methods. However, some studies [19], [21] did not support this method and recommended to use Foggy K-means. Most of the researchers completed their analysis by alternative use of Simple K-means clustering algorithms [16], [18]. Simple K-means clustering algorithm technique has the major advantage of Simple K-means clustering algorithms [23]. Therefore, we selected partition based approaches for implementing this work. In contrast, according to our analysis, Simple K-means clustering algorithm provides correct analysis results for our dataset (Phloretin is one of the most suitable drugs for cancers). The research outcome could be used by the Department of Medical application development and especially for lung cancer dataset analysis in the department of molecular oncology in cancer institution [19].

Natural herbal products are used in traditional medicine, and it is currently considered in anticancer activities [36], [37]. These activity indexes, apoptosis stimulation and antiproliferative activities [38], as research has shown that these natural healing products had no side effects, or as minimum side effects were much more reasonable compared with that chemotherapeutic [39]. Our analysis results indicated that there are four types of compounds more sensitive for all lung cancer cell lines, which are Neopeltolide, Parbendazole, Phloretin, and Piperlongumine. All these compounds can be found in natural sources, including sea sponge, sheep intestine, pepper and apple fruits and apple leaves respectively. However, Phloretin is the most sensitive drug for CCL than other three compounds, according to the results of 53 cell lines (59%). Phloretin (Ph) is a natural polyphenolic compound that exists in apples, pears as well as various vegetables are known to have anticancer activities in numerous cancer cell lines [40]. Ph has also been made known to have anticancer activities by “inducing apoptosis in human bladder cancer cells, leukemia cells, and human colon cancer cells, and inhibiting the growth, invasiveness, and migration of human liver cancer cells” [41].

The results obtained from research done by the National Natural Science Foundation of China confirmed that Phloretin treatment could contain cell production, induce apoptosis and inhibit the persistent and migrant ability of Lung Cancer Cell Lines probably through the variable expression of apoptosis regulators and downstream molecules. Especially, according to a study, Phloretin enhanced the anti-cancer ability of the human body [40]. Also, the research supported by the Key Program of the Shanghai Committee of Science and Technology [42] proves that Ph-induces apoptosis in Non-Small Cell Lung Cancer A549 cells [39]. Additionally, the Cancer Therapeutic Response Portal [24], National Center for

Biotechnology Information [43] and the Genomics of Drug Sensitivity in Cancer Project [44] describe the biological activity of Ph as it inhibits the Glucose uptake. According to this information, Ph is one of the best anticancer drugs for future medical inventions as well as for Lung cancers.

Finally, pharmacological companies and medical institutes continuously develop technology and bioinformatics. This research enhances the ability of the mechanism of action drugs and the interaction with the genetic background of cancer genes as well as clinicians to use anticancer agents more safely and effectively. With the use of Ph, one can invent biologically active anticancer drugs for lung cancer cell lines in the future.

V. Conclusion

Experimental observation using human small molecule lung cancer sensitivity profiling data, and analyzed the anticancer drug sensitivity by machine learning algorithm (Simple k-means and Filtered cluster). The results indicated that k-means clustering algorithm could be used to identify sensitive drug for lung cancer cell lines used in particular concentrations. Furthermore, our analysis confirmed that the Neopeltolide, Parbendazole, Phloretin and Piperlongumine anticancer drug compounds are more sensitive to all 91 human lung cancer cell lines under different concentrations (p-value < 0.001). All these compounds can be found in natural sources, including sea sponge, sheep intestine, pepper and apple fruits and apple leaves, respectively. The performance of the partitioning based algorithms was analyzed by using only selected three attributes from the total number of attributes of the input dataset. It is evident that the results show the computational complexity of the Simple K-Means algorithm with the lung cancer dataset that is better than Filtered clustering algorithm for the dataset. The K-Means algorithm is efficient for the lung cancer dataset. It is well suited for requirement clustering of cancer-related medical applications. This study is beneficial as a reference for researchers who are experimenting drug developments for cancers such as human small molecule lung cancer.

Author Contribution

C.W. and M.N.H. conceived the study idea and developed the analysis plan. C.W. analyzed the data and wrote the initial paper. M.N.H. helped preparing the figures and tables, and in finalizing the manuscript. All authors read the manuscript.

References

[1] "Comprehensive Cancer Information", *National Cancer Institute*, 2017. [Online]. Available: https://www.cancer.gov/.

[2] "Campaign victories", *Cancer Research UK*, 2017. [Online]. Available: http://www.cancerresearchuk.org/support-us/campaign-for-us/our-campaigning-successes.

[3] A. Marinus, "European organisation for research and treatment of cancer", *The Quality Assurance Journal*, vol. 6, no. 4, pp. 251-257, 2002.

[4] M. N. Halgamuge, "Critical Time Delay of the Pineal Melatonin Rhythm in Humans due to Weak Electromagnetic Exposure", Indian Journal of Biochemistry and Biophysics (IJBB), Volume 50, Number 4, pp 259-265, Aug 2013.

[5] M. N. Halgamuge, "Pineal Melatonin Levels Disruption on Human Due to Electromagnetic Fields and ICNIRP Limits", Radiation Protection Dosimetry, Oxford Journals, Volume 154, Issue 4, pp 405-416, June 2013.

[6] Y. Zhao, E. Butler and M. Tan, "Targeting cellular metabolism to improve cancer therapeutics", *Cell Death and Disease*, vol. 4, no. 3, p. e532, 2013.

[7] I. Bahce, M. Yaqub, E. Smit, A. Lammertsma, G. van Dongen and N. Hendrikse, "Personalizing NSCLC therapy by characterizing tumours using TKI-PET and immuno-PET", *Lung Cancer*, vol. 107, pp. 1-13, 2017.

[8] J. Young, M. Peyton, H. Seok Kim, E. McMillan, J. Minna, M. White and E. Marcotte, "Computational discovery of pathway-level genetic vulnerabilities in non-small-cell lung cancer", *Bioinformatics*, vol. 32, no. 9, pp. 1373-1379, 2016.

[9] J. George, J. Lim, S. Jang, Y. Cun, L. Ozretić, G. Kong, F. Leenders and L. Xin, "Comprehensive genomic profiles of small cell lung cancer", *Nature*, vol. 524, no. 7563, pp. 47-53, 2015.

[10] J. Cabrera, A. Dionisio and G. Solano, "Lung Cancer Classification Tool Using Microarray Data and Support Vector Machines", *Information, Intelligence, Systems and Applications (IISA), 2015 6th International Conference*, pp. 1-6, 2015.

[11] S. Peters, A. Adjei, C. Gridelli, M. Reck, K. Kerr and E. Felip, "Metastatic non-small-cell lung cancer (NSCLC): ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up", *Annals of Oncology*, vol. 23, no. 7, pp. vii56-vii64, 2012.

[12] A. Bhattacharjee, W. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E. Mark, E. Lander, W. Wong, B. Johnson, T. Golub, D. Sugarbaker and M. Meyerson, "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses", *Proceedings of the National Academy of Sciences*, vol. 98, no. 24, pp. 13790-13795, 2001.

[13] M. Yousef and E. Tsiani, "Metformin in Lung Cancer: Review of in Vitro and in Vivo Animal Studies", *Cancers*, vol. 9, no. 5, p. 45, 2017.

[14] D. Gomez and Z. Liao, *Non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC). In Target Volume Delineation and Field Setup*. Springer Berlin Heidelberg, 2013, pp. 87-103.

[15] M. Yousef and E. Tsiani, "Metformin in Lung Cancer: Review of in Vitro and in Vivo Animal Studies", *Cancers*, vol. 9, no. 5, p. 45, 2017.

[16] J. George, J. Lim, S. Jang, Y. Cun, L. Ozretić, G. Kong, F. Leenders, X. Lu, L. Fernández-Cuesta, G. Bosco and C. Müller, "Comprehensive genomic profiles of small cell lung cancer.", *Nature*, vol. 524, no. 7563, pp. 47-53, 2015.

[17] "Radiation sensitivities in various anticancer drug-resistant human lung cancer cell lines and mechanism of radioresistance in cis-Diamminedichloroplatinum (II) resistant lung cancer cell line", *Lung Cancer*, vol. 10, p. S333, 1994.

[18] R. Pal, R. Rahman, S. Haider and S. Ghosh, "Design of Probabilistic Random Forests with Applications to Anticancer Drug Sensitivity Prediction", *Cancer Informatics*, p. 57, 2016.

[19] R. Shoemaker, "The NCI60 human tumour cell line anticancer drug screen", *Nature Reviews Cancer*, vol. 6, no. 10, pp. 813-823, 2006.

[20] N. Malviya, N. Choudhary and K. Jain, "Content Based Medical Image Retrieval and Clustering Based Segmentation to Diagnose Lung Cancer.", *Advances in Computational Sciences and Technology*,, vol. 10, no. 6, pp. 1577-1594, 2017.

[21] J. Masters, "Human cancer cell lines: fact and fantasy", *Nature reviews Molecular cell biology*, vol. 1, no. 3, pp. 233-236, 2000.

[22] R. Subbaiya and M. Selvam, "Synthesis and Characterization of Silver Nanoparticles from Streptomyces olivaceus sp-1392 and its Anticancerous Activity Against Non-Small Cell Lung Carcinoma Cell Line (NCI-H460)", *Current Nanoscience*, vol. 10, no. 2, pp. 243-249, 2014.

[23] B. Seashore-Ludlow, M. Rees, J. Cheah, M. Cokol, E. Price, M. Coletti, V. Jones, N. Bodycombe, C. Soule, J. Gould, B. Alexander, A. Li, P. Montgomery, M. Wawer, N. Kuru, J. Kotz, C. Hon, B. Munoz, T. Liefeld, V. Dan ik, J. Bittker, M. Palmer, J. Bradner, A. Shamji, P. Clemons and S. Schreiber, "Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset", *Cancer Discovery*, vol. 5, no. 11, pp. 1210-1223, 2015.

[24] "Cancer Therapeutics Response Portal", *Portals.broadinstitute.org*, 2017. [Online]. Available: https://portals.broadinstitute.org/ctrp/.

[25] G. Speyer, D. Mahendra, H. J. Tran, J. Kiefer, S. L. Schreiber, P. A. Clemons, H. Dhruv, M. Berens, and S. Kim, "Differential Pathway Dependency Discovery Associated with Drug Response Across Cancer Cell Lines," Pacific Symposium on Biocomputing. 2017, 2016. Vol. 22, p. 497, NIH Public Access.

[26] D. G. Covell, "Data Mining Approaches for Genomic Biomarker Development: Applications Using Drug Screening Data from the Cancer Genome Project and the Cancer Cell Line Encyclopedia," Plos One, vol. 10, no. 7, Jan. 2015.

[27] A. K. Yadav, D. Tomar, and S. Agarwal, "Clustering of lung cancer data using Foggy K-means," 2013 International Conference on Recent Trends in Information Technology (ICRTIT), pp. 13–18, 2013.

[28] A. Gupta, A. Mohammad, A. Syed, and M. N. Halgamuge. A Comparative Study of Classification Algorithms using Data Mining: Crime and Accidents in Denver City the USA. International Journal of Advanced Computer Science and Applications (IJACSA), 7(7): 374 – 381, 2016.

[29] P. Palanisamy, P. Perumal, K. Thangavel, and R. Manavalan, "Informative Gene Selection for Leukemia Cancer Using Weighted K-Means Clustering," Pharmacy and Biological Sciences, vol. 9, no. 4, pp. 12–16, Jul. 2014.

[30] A. Roozgard, S. Cheng, and H. Liu, "Malignant nodule detection on lung CT scan images with kernel RX-algorithm," Proceedings of 2012 IEEE-EMBS International Conference on Biomedical and Health Informatics, pp. 499–502, 2012.

[31] J. Wu, M. Gensheimer, X. Dong, D. Rubin, S. Napel, M. Diehn, B. Loo and R. Li, "Robust intratumor partitioning to identify high-risk subregions in lung cancer: a pilot study.", *International Journal of Radiation Oncology* Biology* Physics,*, vol. 95, no. 5, pp. 1504-1512, 2016.

[32] A. Dharmarajan and T. Velmurugan, "Lung Cancer Data Analysis by k-means and Farthest First Clustering Algorithms," Indian Journal of Science and Technology, vol. 8, no. 15, Apr. 2015.

[33] J. Lu, L. Chen, J. Yin, T. Huang, Y. Bi, X. Kong, M. Zheng and Y. Cai, "Identification of new candidate drugs for lung cancer using chemical–chemical interactions, chemical–protein interactions and a K-means clustering algorithm.", *Journal of Biomolecular Structure and Dynamics*, vol. 34, no. 4, pp. 906-917, 2016.

[34] W. Yang, H. Lightfoot, G. Bignell, F. Behan, T. Cokelear, D. Haber, J. Engelman, M. Stratton, C. Benes, U. Mcdermott, and M. Garnett, "Genomics of Drug Sensitivity in Cancer (GDSC): A resource for biomarker discovery in cancer cells," European Journal of Cancer, vol. 69, 2016.

[35] A. Jemal, M. M. Center, C. Desantis, and E. M. Ward, "Global Patterns of Cancer Incidence and Mortality Rates and Trends," Cancer Epidemiology Biomarkers & Prevention, vol. 19, no. 8, pp. 1893–1907, Aug. 2010.

[36] J. Brownlee, "Supervised and Unsupervised Machine Learning Algorithms", *Machine Learning Mastery*, 2016.

[37] M. Rees, B. Seashore-Ludlow, J. Cheah, D. Adams, E. Price, S. Gill, S. Javaid, M. Coletti, V. Jones, N. Bodycombe, C. Soule, B. Alexander, A. Li, P. Montgomery, J. Kotz, C. Hon, B. Munoz, T. Liefeld, V. Dančík, D. Haber, C. Clish, J. Bittker, M. Palmer, B. Wagner, P. Clemons, A. Shamji and S. Schreiber, "Correlating chemical sensitivity and basal gene expression reveals mechanism of action", *Nature Chemical Biology*, vol. 12, no. 2, pp. 109-116, 2015.

[38] A. Alamgir, "Drugs: Their Natural, Synthetic, and Biosynthetic Sources.", *In Therapeutic Use of Medicinal Plants and Their Extracts*, vol. 1, pp. 105-123, 2017.

[39] D. J. Newman and G. M. Cragg, "Plants as a source of anti-cancer agents," Journal of Ethnopharmacology, vol. 100, no. 1-2, pp. 72–79, Aug. 2005.

[40] J. Min, X. Li, K. Huang, H. Tang, X. Ding, C. Qi, X. Qin, and Z. Xu, "Phloretin induces apoptosis of non-small cell lung carcinoma A549 cells via JNK1/2 and p38 MAPK pathways," Oncology Reports, Feb. 2015.

[41] L. Ma, R. Wang, Y. Nan, W. Li, Q. Wang, and F. Jin, "Phloretin exhibits an anticancer effect and enhances the anti-cancer ability of cisplatin on non-small cell lung cancer cell lines by regulating expression of apoptotic pathways and matrix metalloproteinases," International Journal of Oncology, vol. 48, no. 2, pp. 843–853, Dec. 2015.

[42] J. B. Fordham, A. R. Naqvi, and S. Nares, "Leukocyte Production of Inflammatory Mediators Is Inhibited by the Antioxidants Phloretin, Silymarin, Hesperetin, and Resveratrol," Mediators of Inflammation, vol. 2014, pp. 1–11, 2014.

[43] "National Center for Biotechnology Information," National Center for Biotechnology Information. [Online]. Available: https://www.ncbi.nlm.nih.gov/.

[44] W. Yang, J. Soares, P. Greninger, E. Edelman, H. Lightfoot, S. Forbes, N. Bindal, D. Beare, J. Smith, I. Thompson, S. Ramaswamy, P. Futreal, D. Haber, M. Stratton, C. Benes, U. McDermott and M. Garnett, "Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells", *Nucleic acids research*, vol. 41, no. 1, pp. D955-D961, 2012.