AUTOMATED CLASSIFICATION OF GENETIC MUTATIONS IN CANCER USING MACHINE LEARNING

Saroj Doranga[1], Rajeev Nepal[2], Pratigya Timsina[3]
[1]Department of Electronics and Computer Engineering, Thapathali Campus, Institute of Engineering, Kathmandu, Nepal
Email: 9846656375saroj@gmail.com
[2]Department of Community Medicine, Universal College of Medical Sciences and Teaching Hospital (UCMS-TH)
Email: nepalrajeev11@gmail.com
[3]MBS Scholar, Prithvi Narayan Campus, Tribhuvan University, Pokhara, Nepal
Correspondence should be addressed to Saroj Doranga,
Email: pratigyatimsina324@gamil.com

ABSTRACT

Efforts to decipher the genomic data of cancer and its implications for treatment face challenges. Robust preclinical models reflecting human cancer's genomic diversity, along with comprehensive genetic and pharmacological annotations, can greatly aid in this endeavor. Large collections of cancer cell lines effectively capture the genomic diversity and provide valuable insights into the response to anti-cancer drugs. In this study, we demonstrate significant agreement and biological consistency between drug sensitivity measurements and their corresponding genomic predictors from two publicly available pharmacogenomics databases: The Cancer Cell Line Encyclopedia and the Genomics of Drug Sensitivity in Cancer. Despite ongoing efforts to identify cancer-related metabolic changes that may reveal vulnerabilities to targeted drugs, systematic evaluations of metabolism in relation to functional genomics features and associated dependencies are still uncommon. To gain further insights into the metabolic diversity of cancer, we analyzed 225 metabolites in 928 cell lines representing over 20 cancer types using liquid chromatography-mass spectrometry (LC-MS) in the Cancer Cell Line Encyclopedia (CCLE). The analysis revealed missing data for various features, with certain percentages exceeding 40%, leading to the removal of 12 features according to standard procedures. Further analysis revealed 25 unique chromosomes and 4 unique Variant_Types in the dataset. Model performance assessment showed an accuracy score of 96% using a logistic regression model.

Keywords: Cancer Cell Line Encyclopedia, Chromatography-mass spectrometry, Genomic, preclinical models, drug sensitivity

Introduction

Millions of individuals worldwide succumb to cancer each year, making it a leading cause of death. Among men, lung cancer is the primary contributor to

cancer-related fatalities, while breast cancer holds that position among women. Timely identification and treatment are crucial in order to minimize mortality rates. While there has been significant discourse about the transformative potential of precision medicine, particularly genetic testing, in revolutionizing cancer treatment, the actual implementation has been limited due to the extensive manual labor involved. The Cancer Cell Line Encyclopedia (CCLE) Project, a collaborative effort between the Broad Institute and the Novartis Institutes for Biomedical Research, aims to conduct comprehensive genetic characterizations of around 1000 cancer cell lines on a large scale. In this project, we analyze the Cancer Cell Line Encyclopedia (CCLE) dataset to gain insights into the genetic mutations of cancer cell lines. The CCLE dataset provides comprehensive information on gene expression, chromosomes, and sequencing data for various cancer cell lines. Our main objectives are to preprocess the dataset by handling missing values and outliers, apply appropriate feature encoding techniques, and build a logistic regression model for classification. We evaluate the performance of the model using metrics such as accuracy, confusion matrix, and classification report. By leveraging the CCLE dataset and employing analytical techniques, we aim to improve our understanding of cancer genetics, identify mutation patterns, and contribute to the broader field of cancer research.

This project has undergone three distinct phases, beginning with Phase I in January 2008. The primary objectives of this collaboration include detailed genetic and pharmacological profiling of a diverse collection of human cancer models, development of integrated computational analyses linking specific genetic characteristics, gene expression patterns, and cell lineage information to distinct vulnerabilities to pharmaceutical agents, as well as translation of cell line integrative genomics into patient stratification strategies. The participating teams directly obtained 1000 cell lines from various publicly accessible repositories, such as ATCC (American Type Culture Collection), DSMZ (Deutsche Sammlung von Mikroorganismen und Zellkulturen), and the KCLB (Korean Cell Line Bank), ensuring that the genomic data generated closely align with the original cell line derivatives.

Phase II of the CCLE project built upon the initial characterizations by incorporating advanced Next-Generation sequencing techniques to further enhance and refine the profiling of expressed mRNA through RNA-seq. This phase also involved deeper exploration of genetic alterations through exome sequencing, complementing the efforts of the Sanger Center by covering

previously unanalyzed cell lines. Additionally, the miRNA content of all cell lines was characterized, metabolite abundance across the CCLE was quantified for 225 metabolites using mass reaction monitoring (MRM) mass spectrometry, bulk Histone H3 tail modifications were quantified through mass spectrometry-based reverse phase protein array analysis, and collaboration with Michael Davis and Gordon Mills at MD Anderson facilitated further investigations.

Phase III of the Cancer Cell Line Encyclopedia project focuses on attaining a higher level of completeness in the characterization of cell lines at the nucleic acid level. Given that the majority of therapeutic interventions function by disrupting or altering protein activity, and with growing interest in antibody-drug conjugates, antibody-mediated cellular cytotoxicity (ADCC), and CAR-T cells targeting surface proteins, efforts are directed towards defining the CCLE proteome through mass spectrometry techniques.

The aim of our project is to create a Machine Learning algorithm that can classify genetic variations based on the provided knowledge base. This classification will involve assigning cancer genetic mutations to one of nine predefined classes of cancer. Our project is guided by two hypotheses:

a. Null Hypothesis (H0): In this hypothesis, researchers assume that the occurrence of cancer is primarily attributed to specific gene mutations rather than a wide range of genetic variations.

b. Alternative Hypothesis (H1): Conversely, the alternative hypothesis suggests that we can identify numerous genetic variations and mutations associated with cancer. These identified variations and mutations can be classified into nine distinct classes, representing different types of cancer.

By developing the Machine Learning algorithm and leveraging the knowledge base, we aim to investigate and validate either the null hypothesis or the alternative hypothesis.

Literature Review

A brain tumor is an abnormal cell collection with four degrees (Iqbal et al., 2019). Also, in other study author proposed a deep learning model using LSTM and ConvNet for brain tumor delineation used 3DCNN for segmenting brain regions (Ramzan et al., 2020). Other researcher developed a CNN ResNet for Gliomas classification (Shen et al., 2017).

Lung cancer is a leading cause of death worldwide (Jiang et al, 2017). He introduced a 2D CNN system for pulmonary nodule detection. Asuntha & Srinivasan (2020) employed deep learning techniques for lung nodule identification.

Skin cancer is characterized by uncontrolled cell growth in the skin. Premaladha& Ravichandran, (2016) proposed an intelligent system for melanoma classification. Bareiro Paniagua et al., (2016) focused on dermoscopy image analysis, and Khan et al., (2019) utilized common features for image analysis.

Acute lymphocytic leukemia (ALL) affects blood and bone marrow. Sharma & Kumar (2019) used Artificial Bee Colony and BPNN for Leukemia diagnosis. Zhang et al., (2020) assessed leukocyte impact using image analysis.

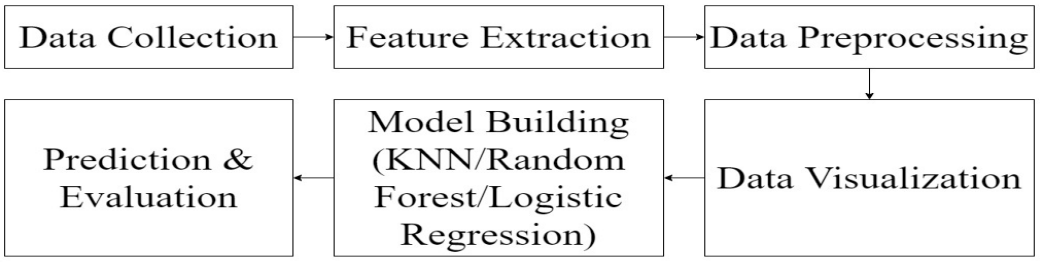
Breast cancer is prevalent in women worldwide. They suggested a hybrid selection model for identifying biased genes (Vijayarajeswari et al., 2019). Abdel-Zaher & Eldeib, (2016) proposed a CNN-based method for breast carcinoma detection. Etemadi et al., (2016) developed a classification model for breast cancer.

Liver tumor detection utilized machine learning techniques (Chang et al., 2017). Frid-Adar et al., (2018) applied deep learning methods for medical image generation. Romero et al., (2019) proposed an end-to-end solution for differentiating liver metastases and healthy cysts in CT images.

Methodology

To carry out this this, we employed "Automated Classification of Genetic Mutations in Cancer using Machine Learning" method that involved several key steps. Firstly, relevant genetic mutation data was collected from reliable sources. Then, meaningful features were extracted from the collected data. The data was preprocessed to clean and transformed it into a suitable format, followed by visualizations to gain insights. Different machine learning algorithms such as KNN, Random Forest, or Logistic Regression were utilized to build predictive models. These models were trained using preprocessed data and evaluated by comparing their predictions with actual labels. This methodology enabled the automated classification of genetic mutations in cancer using machine learning techniques and block diagram was shown in figure below.

Fig 1: Block diagram of overall process



Dataset. The data utilized in this study was sourced from DepMap (Dependency Map), a database portal focused on cancer genetic and pharmacological dependencies. The dataset comprised the CCLE_Mutation file, which contained the most recent and updated cancer mutation data. CCLE (Cancer Cell Line Encyclopedia) provided comprehensive biological information about various cancer cell lines, including a vast collection of human cancer cell lines with compiled data on gene expression, chromosomes, and sequencing.

The CCLE_Mutation datafile consisted of 27 features and 10,000 instances. These features encompassed important information such as Hugo_Symbol, Entrez_Gene_Id, NCBI_Build, Chromosome, Start_position, End_position, Strand, Variant_Type, Reference_Allele, Alternate_Allele, dbSNP_RS, dbSNP_Val_Status, Annotation_Transcript, DepMap_ID, isDeleterious, isTCGAhotspot, TCGAhsCnt, isCOSMIChotspot, COSMIChsCnt, ExAC_AF, Variant_annotation, HC_AC, RD_AC, RNAseq_AC, SangerWES_AC, WGS_AC, and Variant_Classification.

Researchers gathered this data to discern the genetic mutations present in cancer cell lines. With fourteen distinct variant classes identified within the dataset, this posed a multiclass classification problem.

Feature handling. The dataset consisted s of features that fell into three different data types: object, integer, and boolean. Each of these data types was processed and adjusted to ensure compatibility with the machine learning model. This involved appropriate transformations and encoding techniques to effectively incorporate these features into the model's training and prediction processes. By handling the diverse data types appropriately, the information from the machine learning algorithm believing to contribute to accurate and meaningful predictions.

Data pre-processing. In order to prepare the data for analysis, the following steps were devised:

1) **Calculation of missing values:** The missing no library was utilized to determine the presence of missing values in the dataset. Features with more than 40% missing values were removed, while the remaining missing values are imputed using appropriate techniques.
2) **Imputation of missing values:** Categorical variables with a low percentage of missing values were imputed using the mode (most frequent items), while numerical variables with a low percentage of missing values are imputed using the median.
3) **Treatment of outliers:** The describe (???) function was employed to identify outliers by generating a summary of the dataset, including minimum and maximum values, as well as quartiles. Boxplots from the seaborn library were used for visualizing the outliers.
4) **Handling categorical and numerical data:** Based on the chosen machine learning model, suitable featurization techniques were applied. For this multiclass classification problem, one-hot encoding was used for categorical features, while numerical data was processed using StandardScaler. The dataset did not require any column discretization.
5) **Imbalance analysis:** To ensure unbiased results, an analysis of data imbalance was performed on the target column, which was the Variant_Classification. The dataset contained a total of fourteen unique variants with imbalanced distribution. The variant classes were listed, indicating the presence of class imbalance in this multi-class classification problem.
6) **Univariate analysis:** Each feature was individually analyzed to gain insights into its characteristics, including the total number of subdivisions and percentage distribution. Bar charts and pie charts were utilized for visual representation.

Data visualization. Data visualization was carried out using the matplotlib and seaborn libraries to depict the distribution of the dataset and evaluate performance metrics.

Model building. Since the dataset consisted of categorical data, classification algorithms are employed. The data was prepared using various preprocessing and feature encoding techniques such as standardization and one-hot encoding. The dataset was then divided into two subsets: a train set and a test set. Logistic

regression was one of the classification algorithms used to build the model, and its performance was evaluated using different metrics.

Logistic regression. Logistic regression was a supervised learning algorithm that is based on the concept of odds ratio. It is commonly used for predicting the probability of a target variable. In our case, since the dataset is categorical, we employed different encoding techniques such as one-hot encoding or response coding (Berger et al., 2018). Additionally, since we had an imbalanced dataset, we utilized logistic regression with a class balancing strategy. This approach aimed to improve recall and precision for classes that had fewer data points.

Prediction & evaluation. Once the logistic regression model was built, we proceeded to predict the target variable and evaluated the model's performance using various metrics. Accuracy, confusion matrix, and classification report were some of the metrics employed to assess the model's effectiveness. These metrics provided insights into the overall accuracy, as well as the classification performance for different classes in the dataset.

Results

The results of this study include various visualizations and analyses. A histogram illustrates the distribution of variations between the index of a variation and the number of occurrences, providing insights into the frequency of different variations. A cumulative distribution plot demonstrates the distribution of variation classes, giving an overview of the proportion of each class. We also present the distribution of chromosomes, indicating whether the data is evenly distributed across different chromosomes. The distribution of variant types in the CCLE mutation data is visualized, revealing the prevalence of different types of genetic variants. Additionally, we provide the percentage of missing values in the CCLA mutation data, which indicates the completeness of the dataset. During the data preprocessing phase, we ensure that the features have no outliers, which enhances the reliability of the subsequent analyses. Lastly, we present interpretations of the features and their correlations, providing insights into the relationships and patterns within the data. These findings are visually represented in the figures as illustrated below.

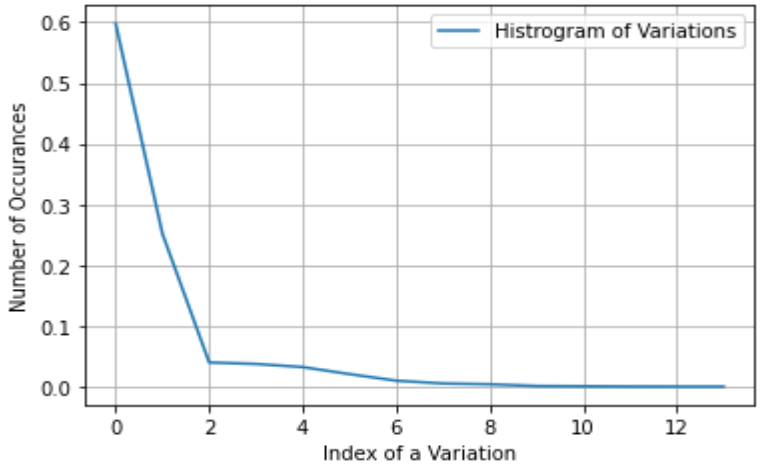


Fig 2: *Histogram of variations between index of a variation number of occurrences*

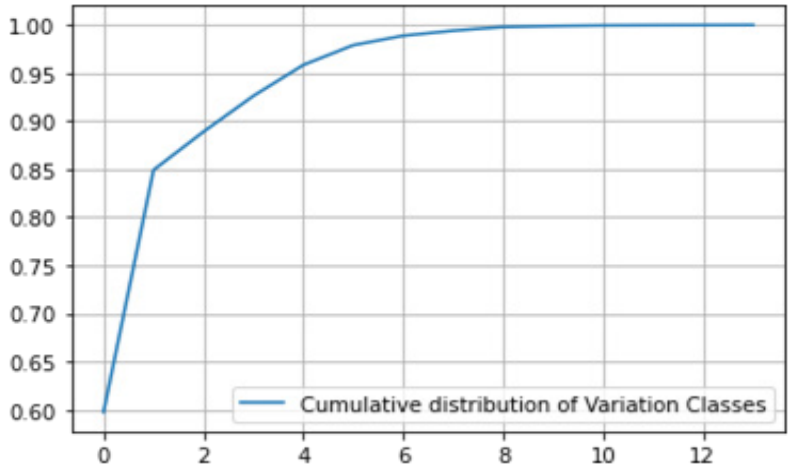


Fig 3: *Cumulative distribution of variation classes*

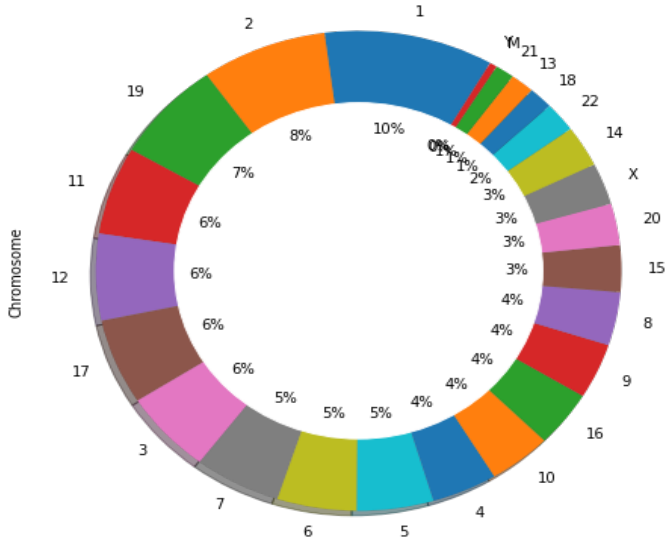


Fig 4: *Distribution of chromosome having balanced distribution of data*

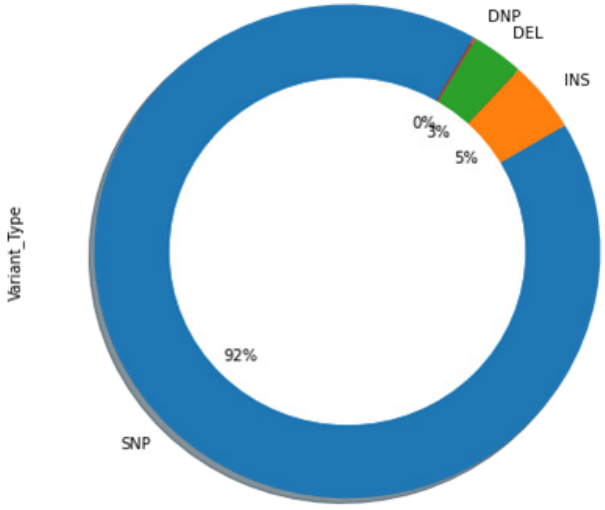


Fig 5: *Distribution of variant types of CCLE mutation data*

The analysis of the CCLE dataset revealed interesting findings regarding missing values, model performance, and data characteristics. Among the features examined, it was discovered that several variables had a considerable number of missing values. Specifically, the features dbSNP_RS, dbSNP_Val_Status, Annotation_Transcript, TCGAhsCnt, ExAC_AF, HC_AC, RD_AC, RNAseq_AC, SangerWES_AC, and WGS_AC exhibited missing values at rates of 81.07%, 96.73%, 0.05%, 70.02%, 73.05%, 91.22%, 99.70%, 47.76%, 62.58%, and 59.76% respectively.

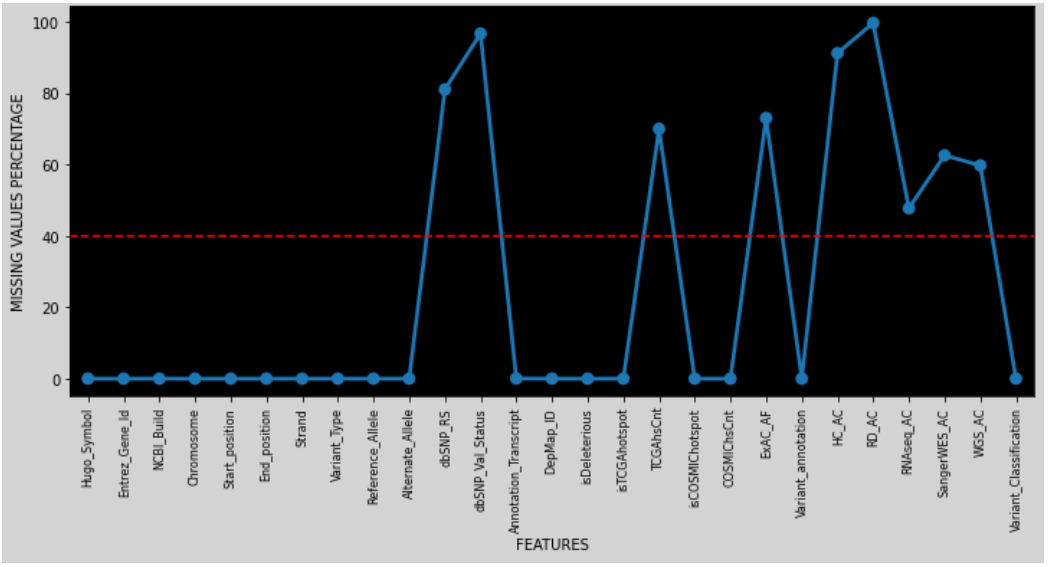


Fig 6: *Percentage of missing values in CCLA mutation data*

Having onto model evaluation, a logistic regression model demonstrated promising performance on the test data, achieving an impressive accuracy score of 96%. Additionally, the logistic model's accuracy remained high at 97% when applied to the testing test dataset.

In order to handle missing values effectively, a threshold of 40% was established, denoted by a red line. Any features surpassing this threshold were deemed to have an excessive number of missing values and were subsequently removed from the CCLE_Mutation data. As a result, a total of 12 features were identified for deletion.

Further analysis of the dataset unveiled interesting insights into its composition. There were 25 unique chromosomes present, providing diverse genetic information. Additionally, the dataset contained four unique Variant_Types, with the SNP type representing the majority, accounting for approximately 92% of the records. Notably, the feature NCBI_Build exhibited no outliers and consisted of a single value (37), while the features Start_position and End_position also displayed no outliers.

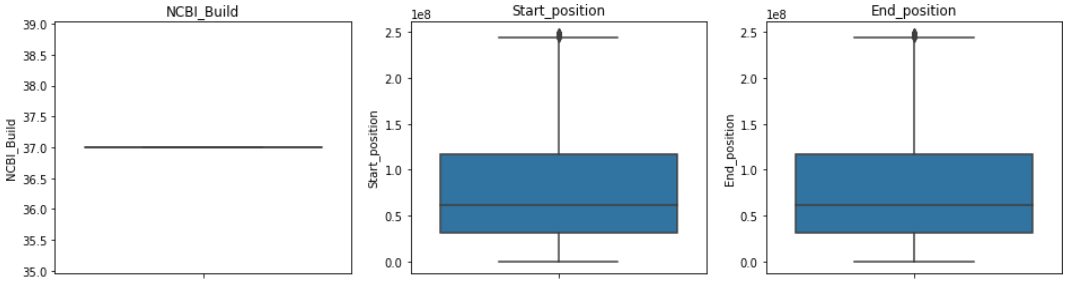


Fig 7: *Features Has no Outliers During Pre processing*

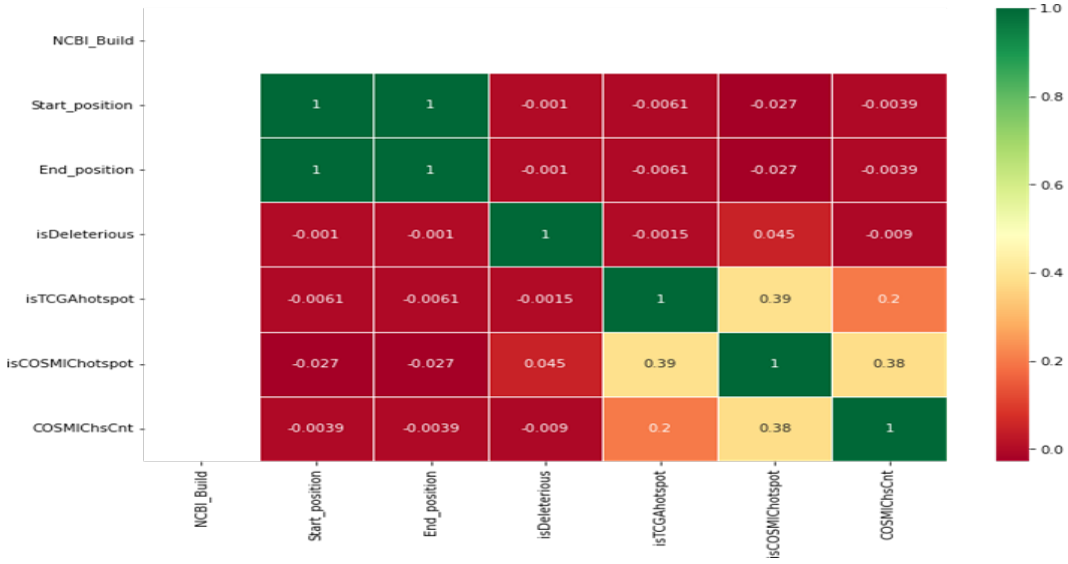


Fig 8: *Interpretation of features and their correlations*

These findings shed light on the characteristics of the CCLE dataset, emphasizing the importance of handling missing values, understanding the distribution of variant types, and identifying outliers. By considering these factors and leveraging the logistic regression model's high accuracy, this project contributes to the field of cancer research and facilitates a deeper understanding of genetic mutations in cancer cell lines.

Discussion

This research incorporated various theoretical frameworks and concepts to provide a comprehensive understanding. The application of machine learning theories was crucial in our study, as it enabled us to develop predictive models by training algorithms using labeled data. We utilized specific machine learning algorithms, including K-Nearest Neighbors (KNN), Random Forest, and Logistic

Regression, which have well-established theoretical foundations and assumptions. We also delved into the theoretical underpinnings of the relationship between genetic mutations and cancer. By discussing how genetic mutations contribute to cancer development and progression, we established the importance of accurately classifying and identifying these mutations. This theoretical knowledge played a vital role in the interpretation and analysis of our results, highlighting the potential clinical significance of our automated classification approach.

Furthermore, we explored the concept of feature extraction and its relevance in our study. By discussing the selection and extraction of meaningful features from the genetic mutation data, we demonstrated the importance of capturing relevant information for our machine learning models. This involved drawing upon established biological theories and knowledge regarding genetic markers and factors associated with cancer. Additionally, we incorporated theoretical concepts related to data preprocessing and visualization. We emphasized the significance of data cleaning, normalization, and scaling techniques to ensure the reliability and consistency of our analysis. The utilization of data visualization techniques, such as histograms, cumulative distribution plots, and correlation analysis, enabled us to gain insights into the data distribution and identify potential patterns and relationships. By integrating these theoretical foundations throughout our discussion and analysis, we presented a comprehensive overview of our research approach and its theoretical underpinnings. This helped to establish the validity and relevance of our findings, paving the way for further advancements in automated classification of genetic mutations in cancer using machine learning techniques.

Initially, the collected dataset, consisting of 10000 data points and 27 features, undergoes preprocessing to handle missing values and remove outliers. Fortunately, the dataset does not have many missing values. Cancer cell lines serve as crucial models for studying cancer biology, validating cancer targets, and assessing drug efficacy. Previously, investigations were limited to a few commonly used cell lines or, at most, the NCI60 panel comprising 60 cell lines. For instance, when EGFR mutations were discovered in lung cancer, EGFR inhibitors were developed using a single cell line, A549, as the sensitive model for epidermal growth factor receptor (EGFR) inhibitors. However, this approach starkly contrasts with the number of patients treated in the initial phase III trials of EGFR inhibitors. Consequently, the sensitivity of cancers with activating EGFR mutations went unnoticed initially due to the lack of large-scale, well-defined cancer cell line models. The missing table in Fig. 9 represents the gaps in knowledge.

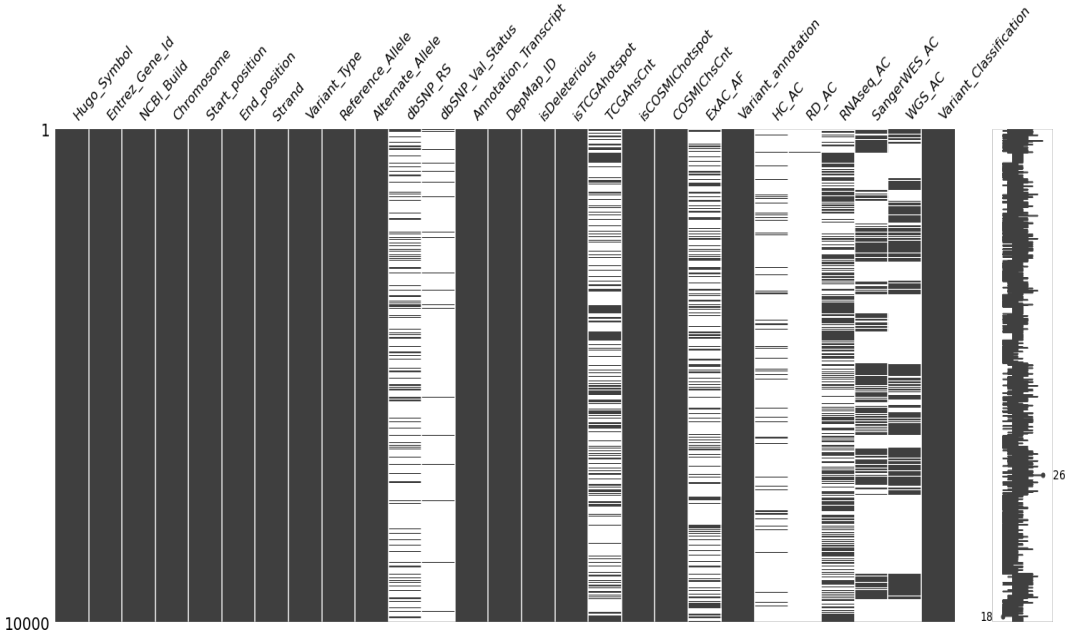


Fig 9: *Matrix showing missing values, indicated by white lines during preprocessing*

Since models do not inherently understand categorical variables and require numerical features, categorical features are converted into numeric features using one-hot encoding. This conversion provides more detailed information by splitting a column containing a categorical feature into multiple columns. The boolean features isDeleterious, isTCGAhotspot, and isCOSMIChotspot are transformed into numerical values, where true is represented by 1 and false by 0. The number of additional columns created during this process depends on the number of distinct categories in the categorical feature. The newly created columns store only ones and zeros based on the presence of each category in the original categorical feature column.

Upon analyzing the preprocessed data, it is observed that Start_position exhibits a high correlation with End_position but not with other features. The isDeleterious feature shows very minimal correlation with the isCOSMIChotspot feature. On the other hand, the isTCGAhotspot feature demonstrates a strong correlation with the isCOSMIChotspot feature and a weaker correlation with the COSMIChsCnt feature. There are a total of twenty-five unique chromosome data points, and the distribution of each chromosome shows a nearly balanced distribution of data.

Conclusion

The preprocessing phase of the project involved handling missing values and removing outliers from the collected dataset, which consisted of 10,000 data points and 27 features. Fortunately, the dataset had a limited number of missing values, alleviating potential challenges during the preprocessing stage.

Cancer cell lines play a crucial role in cancer research, enabling the study of cancer biology, validation of cancer targets, and evaluation of drug efficacy. Traditionally, investigations were restricted to a small set of commonly used cell lines or, at most, the NCI60 panel comprising 60 cell lines. For instance, the discovery of EGFR mutations in lung cancer led to the development of EGFR inhibitors using a single cell line, A549, as a model for sensitivity to these inhibitors. However, this approach starkly contrasted with the scale of patients enrolled in the initial phase III trials of EGFR inhibitors. As a result, the sensitivity of cancers with activating EGFR mutations remained unnoticed due to the lack of comprehensive and well-defined cancer cell line models. The missing gaps in knowledge are depicted in Figure 3, emphasizing the need for more extensive and robust cancer cell line models.

Since models do not inherently comprehend categorical variables and rely on numerical features, categorical features in the dataset were transformed using one-hot encoding. This conversion allowed for a more detailed representation of information by splitting a column containing a categorical feature into multiple columns. Additionally, boolean features such as isDeleterious, isTCGAhotspot, and isCOSMIChotspot were converted into numerical values, with true represented by 1 and false by 0. The number of additional columns created during this process depended on the number of distinct categories present in the original categorical feature column.

Upon analyzing the preprocessed data, notable correlations among the features were observed. Specifically, Start_position exhibited a high correlation with End_position while displaying minimal correlation with other features. The isDeleterious feature demonstrated very limited correlation with the isCOSMIChotspot feature. Conversely, the isTCGAhotspot feature showed a strong correlation with the isCOSMIChotspot feature and a weaker correlation with the COSMIChsCnt feature. Furthermore, the dataset encompassed a total of 25 unique chromosome data points, and the distribution of each chromosome demonstrated a nearly balanced distribution of data.

These findings highlight the essential steps taken during data preprocessing and provide insights into the relationships and characteristics of the preprocessed dataset. By converting categorical features, identifying correlations, and analyzing chromosome distribution, this project contributes to the understanding and utilization of cancer cell line models for cancer research purposes.

References

Abdel-Zaher, A. M., & Eldeib, A. M. (2016). Breast cancer classification using deep belief networks. *Expert Systems with Applications*, 46, 139-144.

Asuntha, A., & Srinivasan, A. (2020). Deep learning for lung Cancer detection and classification. *Multimedia Tools and Applications*, 79, 7731-7762.

Bareiro Paniagua, L. R., Leguizamón Correa, D. N., Pinto-Roa, D. P., Vázquez Noguera, J. L., & Salgueiro Toledo, L. A. (2016). Computerized Medical Diagnosis of Melanocytic Lesions based on the ABCD approach. *CLEI Electronic Journal*, *19*(2), 6-6.

Berger, A. C., Korkut, A., Kanchi, R. S., Hegde, A. M., Lenoir, W., Liu, W., & Moore, R. A. (2018). A comprehensive pan-cancer molecular study of gynecologic and breast cancers. *Cancer Cell*, 33(4), 690-705.

Chang, C. C., Chen, H. H., Chang, Y. C., Yang, M. Y., Lo, C. M., Ko, W. C., & Chang, R. F. (2017). Computer-aided diagnosis of liver tumors on computed tomography images. *Computer Methods and Programs in Biomedicine*, 145, 45-51.

Etemadi, R., Alkhateeb, A., Rezaeian, I., & Rueda, L. (2016, December). Identification of discriminative genes for predicting breast cancer subtypes. *In 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 1184-1188).

Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., & Greenspan, H. (2018). GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*, 321, 321-331.

Iqbal, S., Ghani Khan, M. U., Saba, T., Mehmood, Z., Javaid, N., Rehman, A., & Abbasi, R. (2019). Deep learning model integrating features and novel classifiers fusion for brain tumor segmentation. *Microscopy Research and Technique*, 82(8), 1302-1315.

Jiang, H., Ma, H., Qian, W., Gao, M., & Li, Y. (2017). An automatic detection system of lung nodule based on multigroup patch-based deep learning network. *IEEE Journal of Biomedical and Health Informatics*, *22*(4), 1227-1237.

Khan, S. A., Nazir, M., Khan, M. A., Saba, T., Javed, K., Rehman, A., & Awais, M. (2019). Lungs nodule detection framework from computed tomography images using support vector machine. *Microscopy Research and Technique*, *82*(8), 1256-1266.

Premaladha, J., & Ravichandran, K. S. (2016). Novel approaches for diagnosing melanoma skin lesions through supervised and deep learning algorithms. *Journal Of Medical Systems*, *40*, 1-12.

Ramzan, F., Khan, M. U. G., Iqbal, S., Saba, T., & Rehman, A. (2020). Volumetric segmentation of brain regions from MRI scans using 3D convolutional neural networks. *IEEE Access*, *8*, 103697-103709.

Romero, F. P., Diler, A., Bisson-Gregoire, G., Turcotte, S., Lapointe, R., Vandenbroucke-Menu, F.,& Kadoury, S. (2019, April). End-to-end discriminative deep network for liver lesion classification. *In 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)* (pp. 1243-1246).

Sharma, R., & Kumar, R. (2019). A novel approach for the classification of leukemia using artificial bee colony optimization technique and back-propagation neural networks. *In Proceedings of 2nd International Conference on Communication, Computing and Networking: Chandigarh, India* (pp. 685-694).

Shen, W., Zhou, M., Yang, F., Yu, D., Dong, D., Yang, C., & Tian, J. (2017). Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification. *Pattern Recognition*, *61*, 663-673.

Vijayarajeswari, R., Parthasarathy, P., Vivekanandan, S., & Basha, A. A. (2019). Classification of mammogram for early detection of breast cancer using SVM classifier and Hough transform. *Measurement*, *146*, 800-805.

Zhang, C., Wu, S., Lu, Z., Shen, Y., Wang, J., Huang, P., & Li, D. (2020). Hybrid adversarial-discriminative network for leukocyte classification in leukemia. *Medical Physics*, *47*(8), 3732-3744.