INTERNATIONAL JOURNAL OF CIRCUITS, SYSTEMS AND SIGNAL PROCESSING
DOI: 10.46300/9106.2024.18.1 Volume 18, 2024

Stochastic machine learning models for mutation rate analysis of malignant cancer cells in patients with Acute Lymphoblastic Leukemia

Martsenyuk Vasyl[1,2], Abubakar Sadiq[2], Sverstiuk Andriy[2,3], Dimitrov Georgi[4], Gancarczyk Tomasz[1]
[1]Department of Computer Science and Automatics, University of Bielsko-Biala,
Bielsko-Biala 43-309,
Poland

[2]Department of Computer Science, Ternopil Ivan Puluj National Technical University,
Ternopil,
Ukraine

[3]Department of Medical Informatics, Ternopil Ivan Horbachevsky National Medical University,
Ternopil,
Ukraine

[4]University of Library Studies and Information Technologies,
Sofia,
Bulgaria

Received: March 21, 2023. Revised: February 24, 2024. Accepted: June 3, 2024. Published: July 5, 2024.

Abstract–Acute lymphoblastic leukemia, a pervasive form of the carcinogenic disease, is a lethal ailment subjecting numerous pediatric patients globally to terminal conditions. is a rapidly progressive condition, that exposes patients to conditions including Tumor Lysis Syndrome which often occurs early after the induction chemotherapy, contemporary research focuses primarily on the development of techniques for the early diagnosis of Acute Lymphoblastic Leukemia (ALL), leaving a gap within the literature. This study examines the application of machine learning techniques for the prognosis the mutation rate of cancer cells in pediatric patients with Acute Lymphoblastic Leukemia using clinical data from patients with ALL, who have undergone tests using Next Generation Sequencing (NGS) technology.

An overview of the clinical data utilized is provided in this study, with a comprehensive workflow encompassing, data analysis, dimensionality reduction, classification and regression tree algorithm (CART), and neural networks.

Results here demonstrate the efficiency with which these methods are able to target and decipher cancer cell proliferation in pediatric patients suffering from acute lymphoblastic leukemia. Valuable insights into relationships between key factors and conversion rates were also derived through data mining. However, tree classification and regression algorithms and neural networks used herein indicate the flexibility and the power of machine learning models in predicting the recurrence of cancer cells accurately. This study's results affirm previous findings thus giving clinical proof for mutational drivers among pediatric patients having Acute Lymphoblastic Leukemia. This adds value to results by providing an applicable utility in medical practice. Principally, this study denotes a substantial advancement in leveraging machine learning workflows for mutation rate analysis of cancer cells. By appraising clinical corroboration, emphasizing the explain ability and interpretability, and building upon these findings, future research can contribute to improving patient care and results in the field of Leukaemia.

Keywords–data exploration, machine learning modeling, decision trees, leukemia, neural network, principal component analysis.

I. Introduction

THE incidence of leukemia has doubled over the last decade, [1]. The incidence rate of Acute Lymphoblastic leukemia is about (3~5)/100,000, [2], and the age of onset is mostly before 15 years old, [3], [4]. The male-to-female ratio is approximately 1.2:1, [5]. In recent years, chemotherapy based on risk factor classification has colossally enhanced the prognosis of pediatric patients with ALL. In developed countries, the 5-year event-free survival (EFS) of pediatric ALL can reach more than 85%, and the overall survival (OS) can reach more than 90%, [6]. This inclination in conjunction

with recent trends in growing associated comorbidities, [7], medical expenses, and general mortality, signifies the rising global burden of Leukemia. Acute Lymphoblastic Lleukemia arises from the malignant transformation of lymphoid precursor cells, which are primarily found in the bone marrow and thymus. These precursor cells can differentiate into either B-lymphocytes or T-lymphocytes. The two major subtypes of Acute Lymphoblastic Leukaemia are B-cell Acute Lymphoblastic Leukemia and T-cell Acute Lymphoblastic Leukaemia, each with distinct genetic and clinical features, [8].

Acute Lymphoblastic Leukemia is characterized by a variety of genetic abnormalities that contribute to its heterogeneity. Common chromosomal aberrations include the Philadelphia chromosome (resulting from a translocation between chromosomes 9 and 22), which is more prevalent in adult Acute Lymphoblastic Leukemia cases, [9]. Other genetic alterations involve abnormalities in genes such as IKZF1, CDKN2A, ETV6, and TP53, which play crucial roles in cell cycle regulation and tumor suppression.

Tumour mutation burden (TMB), is a way to identify and quantify the number of non-synonymous, somatic mutations in cancer cells that occur per mega-base of genetic regions of interest, [10]. TMB is a predictor for patient stratification in response to immunotherapy. For example, melanoma studies have correlated high TMB to response to anti-CTLA-4 checkpoint inhibitors, [11], [12], bolstering T-cell response to target tumor cells. High TML is also associated with high efficacy of anti-PD-1 in non-small cell lung cancer (NSCLC), [13].

The analysis of independent factors that contribute to the TMB in children with Acute Lymphoblastic Leukemia at the time of initial diagnosis and the construction of a stochastic model are integral for the prescription of optimal treatment and monitoring of terminal patients to improve patients, life span, and survival rate and quality of life.

Contemporary research employs machine learning in spheres of disease diagnosis, predictive analysis, and individualization of treatment courses, this includes Acute Lymphoblastic leukemia detection and classification using an ensemble of classifiers and pre-trained convolutional neural networks, [14]. Preceding studies have proposed machine learning and deep learning techniques for the prediction of tumor mutation burden, [15]. Nonetheless, machine learning-based TMB prediction models for pediatric ALL have not been analyzed or reported in modern research.

Relying on clinical data from St. Judes Children Hospital, this study resolves to develop several Stochastic machine learning models for Tumour Mutation Burden analysis in pediatric patients with Acute Lymphoblastic Leukaemia, by analyzing the clinical features of actual pediatric patients diagnosed with ALL to provide valuable insights for decision-making regarding the assessment of Tumor Mutation Burden.

II. Related Works

The application of artificial intelligence techniques for Tumor mutation Burden analysis, based on clinical data on Acute Lymphoblastic Leukaemia has gained significant recognition in recent years. State-of-the-art analysis in this field involves the use of advanced machine learning and deep learning algorithms to analyze clinical datasets and extract valuable insights regarding cancer cell mutation.

Top-notch clinical data is essential for precise mutation rate analysis. Modern studies often involve collecting comprehensive electronic health records, [16], or claims data that capture detailed patient information, including disposes of, treatments, medications, and outcomes, [17]. Adequate data preprocessing techniques are applied to handle missing values, and outliers, and ensure data quality, [18].

Efficacious feature engineering is crucial for capturing significant information from clinical data. Modern stratagems focus on designing informative features that represent comorbidity patterns, [19]. This may include encoding various variables, creating derived schemas, and encoding temporal information, [20].

Mutation analysis employs different workflows in machine learning for risk assessment. The contemporary techniques involve ensemble methods like random forests, gradient boosting, and deep learning models such as neural networks, [21]. The ability of these algorithms to capture intricate interrelations and patterns within clinical data ensures a trustworthy risk prognosis. In the context of mutation analysis, it is imperative that the decisions made by machine learning models be easily understandable, [22].

Top-notch research works toward developing approaches to make the predictions explicable and easy to paraphrase by methods including feature importance analysis, attention mechanisms, and rule extraction. This is done with the aim of enhancing transparency and trust in risk assessment for mutation rate analysis. In this course they present solid reasons behind every action taken by ML during the process of making a prediction based on gathered information. They hope their findings will be useful for decision-makers who have little technical knowledge about machine learning but are involved in tasks related to this field at work or elsewhere where such solutions can support their daily work which requires interpretation of results delivered by these systems, [23].

Assessing the risk factors in the body is also a very important part of the quantitative analysis because the presence of other health problems in the patient with acute leukemia affects their overall well-being and response to the prescribed treatment in association with patients without complications. Critical risk assessment often benefits from integrating multiple data sources, including clinical notes, medical imaging, genetic information, and patient report results, [24]. Pioneering approach to techniques such as natural language processing and image analysis to extract useful information from these different data sources, making risk assessment more comprehensive, [25].

Thorough validation and evaluation of machine learning models are important to ensure their dependability and generalisability. Modern studies often conduct immense validation on extensive-scale, diverse datasets to assess the

performance of risk assessment models, [26]. Furthermore, efforts are made to transaxle these artificial intelligence e techniques into clinical practice, taking into consideration factors like usability, clinical relevance, and regulatory compliance, [27].

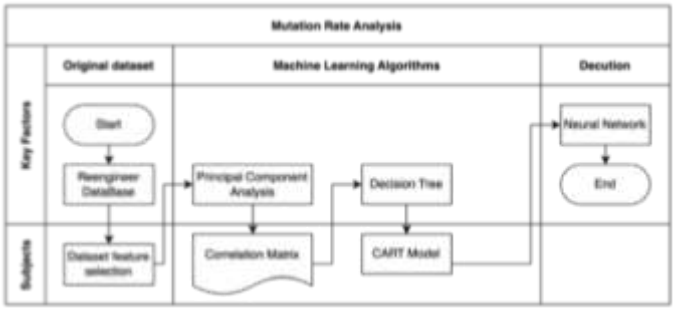

Fig. 1 workflow for the Tumor Mutation Burden analysis based on machine learning

Predominantly, the Modern analysis of applying machine learning workflows for mutation rate analysis and risk assessment based on clinical data of Acute Lymphoblastic Leukemia involves leveraging advanced machine learning algorithms, incorporating explainability and interpretability, integrating multi-modal data sources, and ensuring rigorous validation for clinical adoption, [28]. These developments aim to ameliorate the precision, reliability, and applicability of machine learning-based, mutation rate analysis, and risk assessment tools in the medical sphere.

III. Materials and Methods

Pediatric patients from St. Judes Children's Research Hospital, [29] with Acute Lymphoblastic Leukaemia were identified for the purpose of this research. The prospective cohort study included 73 pediatric patients with Acute Lymphoblastic Leukaemia who have undergone testing using Next Generation Sequencing Technology, falling within the age demographic 3 ± 26. The original database includes 26 attributes and diagnostic deductions based on the results of non-invasive clinical studies conducted in several medical institutions, including St. Judes Children's research hospital, Memphis, Tennessee, USA. Patients' names and Social Security numbers have been removed from the database and replaced with fictions values in accordance with ethical standards. For further analysis, an aggregated dataset with 9 main factors which contains 73 unique records, in which the average age of patients is 13 years, is used to build a correlation matrix and execute principal component analysis. In the input sample, data on the gender distribution indicate 54.3% boys and 44.3% girls.

Acute Lymphoblastic Leukaemia was confirmed according to medical records and detailed observations. Subjects were classed according to the standardized vocabulary and taxonomy for cancer that is developed and maintained by the National Cancer Institute, OncoTree code.

The genetic abnormalities associated with Acute Lymphoblastic Leukaemia in subjects analyzed using cytogenetics. This analysis involved studying the chromosomes of leukemia cells to identify specific chromosomal changes or abnormalities that may contribute to the development and progression of the disease. The patients were also subject to Next-Generation-Sequencing tests, enabling Genome sequencing, [30], identification of driver mutations, [31], and minimal residual disease monitoring, [32], aiding in diagnosis, prognosis, and the development of targeted therapies.

A workflow for the Tumour Mutation Burden analysis based on prior assessment, [33], an indexation of the potential risk factors leading to increase mutation rate in patients with Acute Lymphoblastic Leukaemia (Table I), and a correlation matrix, [34] to assess the feasibility of using all factors for further model building are subsequently provided.

The correlation matrix method as shown in Fig 2, is applied to the dataset investigating the mutation rate analysis of malignant cancer cells in patients with Acute Lymphoblastic Leukaemia, encompassing factors such as Age, Age Class, Institute Source, Platform, Protocol, Gender, Mutation Count, Number of samples per patient, and Tumor Mutation Burden. This analytical approach elucidates the associations between these crucial variables, quantifying the degree of interdependence. The resultant matrix of correlation coefficients discerns patterns of co-variation, identifying potential relationships that may be indicative of underlying biological mechanisms governing mutation rates in Acute Lymphoblastic leukemia. Positive correlations indicate concordant variation and may reflect common genetic determinants, whereas negative correlations indicate inverse correlations and indicate subtle interactions within the genomic landscape of malignant cells. This differentiated study using correlation matrices lays the foundation for a comprehensive understanding of the complex mutational dynamics in acute lymphoblastic leukemia, with implications for targeted therapy strategies and prognostic assessment.

The succeeding course of action is to analyze and index 4 potential risk factors for the increased mutation rate of malignant cancer cells in pediatric patients with Acute Lymphoblastic leukemia.

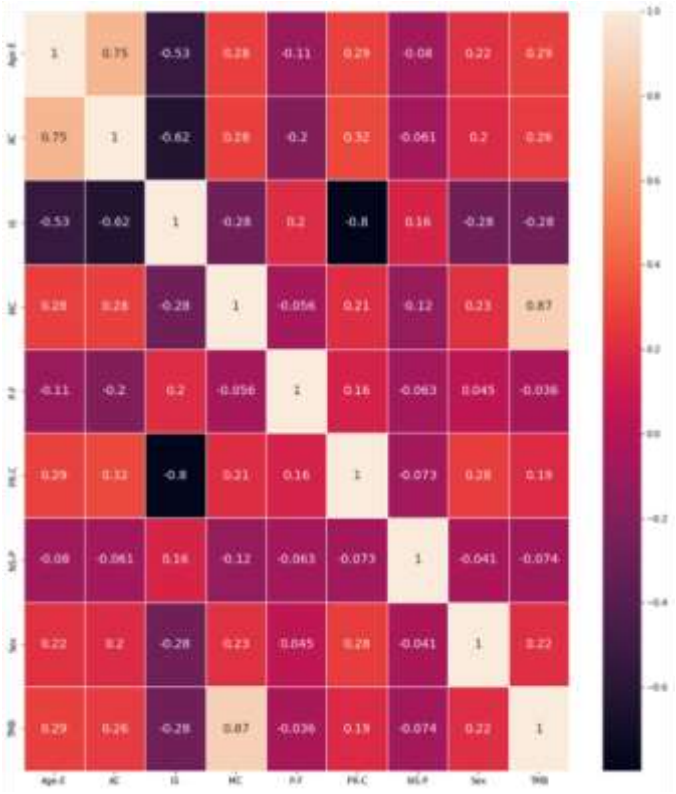

Fig. 2 Correlation Matrix of key factors

Table I. Notation of key factors for classification and regression models for mutation rate analysis

Name of factors	Label of the investigated factors in the input data set	Factor ranges and denotation of their viable variants	Numerical values of factor ranges
Age	Age-E	<10	1
		10 - 15	2
		16 - 20	3
		> 20	4
Gender	Sex	Female	1
		Male	2
		NA	0
Mutation Count	MC	0 - 5	1
		6 - 10	2
		11 - 15	3
		16 - 20	4
		21 - 25	5
		25 - 33	6
Tumor Mutation Burden	TMB	0.00 - 0.30	1
		0.31 - 0.60	2
		0.61 - 0.90	3

IV. Machine Learning Workflows

An interchangeable approach is utilized for the analysis of the mutation rate of malignant cancer cells in pediatric patients with Acute Lymphoblastic Leukemia. For this purpose, a machine learning workflow as displayed in Fig. 1, is proposed in this paper, distinct units are delineated in the respective Algorithms 1-3.

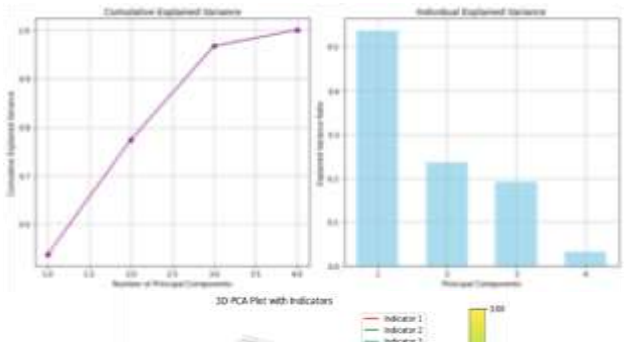

Fig. 3 Cumulative and Individual Explained Variance

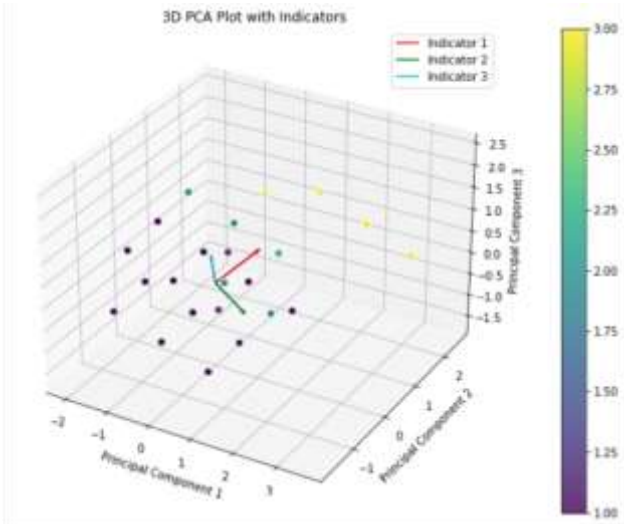

Fig. 4 3D Principal Component Analysis Visualisation

The proposed workflow based on machine learning algorithms for the clinical data of the subjects can be encapsulated as follows.

The initial step of data compilation involves gathering and preprocessing the clinical data. It encompasses tasks including data, collection, refactoring, and handling missing values or outliers. During this stage, the data is properly formatted and transformed into a suitable format for machine learning algorithms.

On the condition that the input dataset has a large number of features, dimensionality reduction methods are applied to reduce the intricacy of the data. Techniques such as Principal Component Analysis (PCA) with the modified feature selection, [35], can help identify the paramount features that are conspicuous to the target variable.

Prior to the construction of the machine learning models, data exploration is vital to comprehend the relationships between different features. This is achieved through depictions, statistical evaluation, and correlation matrices. Correlation analysis facilitates the identification of how distinguished features are related to one another and their potential effect on the target variable. Employing a correlation matrix, the coefficients between the variables in the input data set are shown in a symmetrical order.

Decision trees are renowned machine learning algorithms used for classification and regression tasks, in the scope of this paper, the applicability of decision trees is limited to classification problems. A decision tree is a visual representation of a decision-making process that resembles a tree-like structure. Decision trees can be constructed using the CART algorithm, [36]. The decision tree facilitates the creation of classification rules that can be used to predict a categorical value for the mutation rate of a novel subject based on their Indications.

This paper employs Principal Component Analysis to scrutinize mutation patterns within pediatric patients afflicted with Acute Lymphoblastic Leukemia. Following the standardization of mutation features such as Mutation Count and Tumor Mutation Burden, the cumulative explained variance and individual explained variance plots elucidate the retained variability across principal components.

The cumulative explained variance ratio reveals that the first principal component explains approximately 49.53% of the variance, with the first two components together explaining around 75.92% and all three components explaining 100% of the variance, as indicated in Fig. 3. The top features contributing to each principal component, including 'MC' (a feature), 'Age-E', and 'Sex', provide insights into the dataset's structure and underlying patterns. Specifically, PC1 is heavily influenced by 'MC', 'Age-E', and 'Sex', indicating their significant contributions to the captured variance. This analysis aids in understanding the relationships between the original features and the principal components extracted through PCA, offering valuable insights into the dataset's characteristics and potential associations with Acute Lymphoblastic Leukemia in pediatric patients.

Determining the optimal number of components, which are well-defined with 95% total detail, enables a complete and concise visualizationof the changing nature of the dataset.

The analysis of the following primary factors provides a complex representation of childhood acute lymphoblastic leukemia, in which each patient is represented as a point in the distribution center based on the mutation profile. them. The inclusion of relevant clinical variables, including age, sex, number of mutations, and tumor burden, facilitates subtle interpretation. Complementary cycle techniques work to emphasizespecific genes or subgroups that can be emphasized, defining complex patterns in mutational patterns that may have clinical relevance for prognosis and treatment strategies in pediatric oncology.

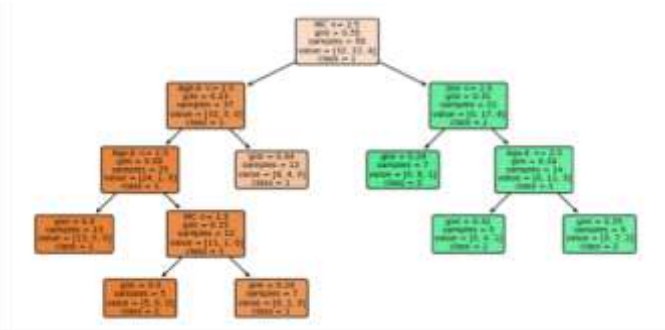

Fig. 5 Classification decision tree

The Principal Component Analysis, as depicted in Fig. 4, uncovers the complex mutational mechanisms of childhood acute leukemia, using advanced statistical methods to break down complex genomic information into meaningful insights. By combining therapies, monitoring helps identify subtle genetic mechanisms, promoting a deeper understanding of the disease spectrum and paving the way for targeted and personalized therapies in childhood leukemia.

Accuracy: 0.87
Decision Tree Rules:
|--- MC <= 2.50
| |--- Age-E <= 2.50
| | |--- Age-E <= 1.50
| | | |--- class: 1
| | |--- Age-E > 1.50
| | | |--- MC <= 1.50
| | | | |--- class: 1
| | | |--- MC > 1.50
| | | | |--- class: 1
| |--- Age-E > 2.50
| | |--- class: 1
|--- MC > 2.50
| |--- Sex <= 1.50
| | |--- class: 2
| |--- Sex > 1.50
| | |--- Age-E <= 2.50

Fig. 6 Precise elucidation of the decision tree with. transition conditions between vertices

Determined from the Kaiser-Meyer-Olkin factor adequacy assessment method, [37], an adequacy assessment of the input dataset is conducted, and a KMO value of 0.60 is obtained from this analysis, which is sufficient for further construction of the decision tree model. A distinct Kaiser-Meyer-Olkin value for all the key factors in the dataset is obtained through supplemental analysis, Age-E - 0.84, MC - 0.55, Sex - 0.82, TMB - 0.55.

Amid the 4 investigated key factors and the constructed correlation matrix, 3 factors were designated to prognosticate the last key factor that indicated the mutation rate of malignant cancer cells in pediatric patients with Acute Lymphoblastic Leukemia. The key factors that most affect the mutation rate of malignant cancer cells in patients with Acute Lymphoblastic Leukemia, which were used to build the classification decision tree model: Age-E, Sex, and Mutation Count.

Explicit information and acceptable values of the analyzed key factors for the mutation rate analysis of malignant cancer cells in pediatric patients with Acute Lymphoblastic Leukemia, [38], used to build the model are provided in Table I. The classification decision tree obtained is based on the results of the multifactorial analysis of analyzing the mutation rate of malignant cancer cells. The CART algorithm (Classification and Regression Trees) was used to build the classification decision tree model provided above, as illustrated in Fig. 5. A comprehensive representation of the decision tree is shown in Fig. 6. The transition conditions for each vertex are also shown.

An accuracy rate of 87% was reached by a decision tree model tailored to pin-point mutations as well as count tumors among a huge number of childhood leukemia. This accuracy level is high because it encapsulates patterns of genetic changes present among cancers across different age groups leading significantly to the comprehension of how genes work.

The decision tree structure is more complex in that it reveals a more complex set of conditions, each of which is closely related to distinct characteristics in the molecular profile of childhood acute leukemia. These factors in complex ways help us understand mutation rate, E-age, and gender contribute to explaining a systematic perspective of complex interactions between genes and clinical variables.

The cross-validation portion of the analysis further shows the model's reliability, showing an accuracy of 85% and a standard deviation of 6%. The validated statistical performance attests to the consistent predictive power and reliability of the model across different sets of data. Such stability makes the model better suited to real-world situations, where it can be navigated without the complex problems inherent in the mutational analysis of childhood lymphoblastic leukemia cells.

The decision tree design model is emerging as a sophisticated and powerful tool in the medical field, providing not only a high level of accuracy but also a complex understanding of the factors that make the fruit Hereditary life affects the rate of tumor recurrence in pediatric patients dealing with complications of serious problems;.lymphoblastic leukaemia.

The application of classification and regression tree models, guided by the Gini index, [39], was made for the classification and prediction of mutation rates in benign cancer cells in patients diagnosed with acute lymphoblastic leukemia. The inclusion of key factors such as age, gender, number of mutations, and tumor mutation burden led to the construction of a decision tree aimed at dividing the dataset into different groups. The Gini index served as an impurity metric, enabling algorithmic decision-making and classification criteria to optimize the purity of emerging nodes. The resulting decision tree, reflecting the complex relationships within the dataset, provides a flexible framework for describing how specific factors interact in

the treatment of acute leukemia patients. lymphoblastic depends on the number of their changes. This methodological approach aligns with a rigorous exploration of the complex genomic landscape underlying Acute Lymphoblastic Leukemia, providing a valuable tool for both prognostic assessments and potential insights into the underlying molecular mechanisms governing mutation patterns in this specific oncological context.

As a way of evaluating the quality of the decision tree model, it is obligatory to obtain the classification accuracy of the model using Python, [40], adhering to the required measures the accuracy of the model is calculated, as a result of which a value of 0.93 is obtained, which corresponds to 93% accuracy of prognosis based on the input parameters. The confusion matrix, as presented in Fig. 7, provides a comprehensive overview of the model's true positive, true negative, false positive, and false negative classifications, enabling a nuanced understanding of its predictive capabilities. Also we present here the normalized confusion matrix. It allows us to evidence the performance of the classifier in a more comprehensive manner. The research based on decision tree model is fitted with the predicting results of clinical study. Moreover splitting conditions of the decision tree are consistent with clinical experience of Acute Lymphoblastic Leukemia.

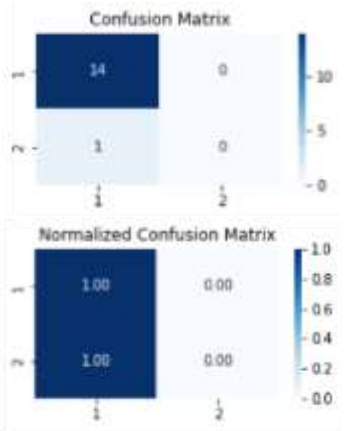

Fig. 7 Confusion matrices

Also the decision tree models were trained on dataset for diagnosing Tumour Mutational Burden (TMB) for pediatric patients with pediatric acute lymphoblastic leukemia. The models have appeared to be an effective tool for early diagnostics of high TMB cases since they have shown the appropriate values of bias, recall, and F1 score metrics. There may be difficulties in classifying specimens that have low TMB levels but this model's efficiency makes up for them by far. It accurately diagnoses most of the cases where children suffer from a form of cancer, thus serving as the backbone for future developments in this area. In the future, the progress of this model continues to provide an opportunity to improve its ability to detect cases with different levels of TMB, thus improving the correct diagnosis and tailored treatment for pediatric patients. all. Macroeconomics and equity methods provide a comprehensive explanation of the model. overall performance of, considering both balance and class inequality in the situation of children's acute lymphoblastic leukemia mutation analysis. This granular analysis is consistent with the overall research topic, increasing our understanding of the complexity of the patterns and enhancements that can predict the number of abnormal cancer cells in a cohort of childhood acute lymphoblastic leukaemia. A comparison of the Gini index decision-making process for quantitative analysis in acute lymphoblastic leukemia patients reveals a descriptive score of 1.0. Specificity, which indicates the negative rate, means the exceptional ability of the model to correctly identify cases where the prediction is correct and there is no mutation. Using a descriptive score of 1.0, the model demonstrates flawless performance in discrimination cases without replication, thus reducing the risk of false positives. In addition to specificity, sensitivity (accuracy rate) evaluates the model's ability to correctly identify situations with variable rates. A thorough understanding of the details and sensitivities, in terms of key factors such as age, age class, source, platform, protocol, gender, number of mutations, number of samples per patient, and tumor mutation burden, explains the model, as shown in Fig. 8. By being able to discriminate it shows different powers that it possesses compared with others hence increasing the capacity to detect right from wrong instances while dealing with acute lymphoblastic leukemia a disease with many pitfalls. In this intricate field of health policy making use of decision tree modeling has emerged as an intricate approach aimed at recognizing intricate patterns in acute lymphoblastic leukemia-specific dataset, [41].

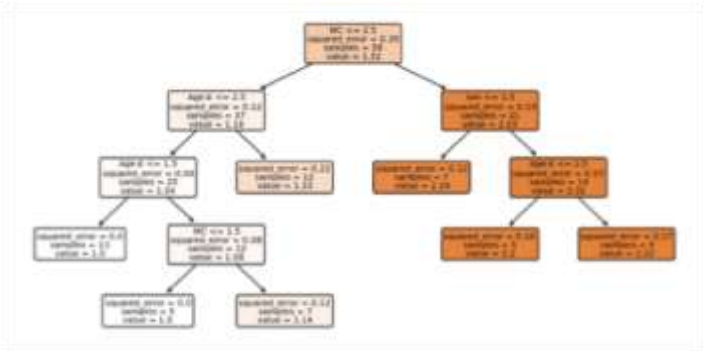

Fig. 8 Regression Decision tree

The model presented an RMSE of 0.13, showing how well it can represent fine-line relationships among clinical variables, as depicted in Fig. 9. Moreover, in supporting consistency; consistency regarding the model, cross-validation RMSE which varies from one part to another, was quoted as 0.14 validating this factor, [42].

This medical record analyses in depth numerous clinical variations using a tree regression model, that outperforms other algorithms due to its special capability of handling intricate interactions thus ensuring maximum adaptability and responsiveness. The predictive power of the model

improves over time such that progressively higher accuracy renders an evolving image of acute lymphoblastic leukemia.

$$Cross-ValidationMSE = \frac{1}{k}\sum_{j=1}^{k}\left(\frac{1}{n_j}\sum_{i=1}^{n_j}(Y_{ij}-y_{ij})^2\right) \quad (1)$$

Where, k is the number of folds, n_j represents the number of data points in the *j*-th fold, and Y_{ij} and y_{ij} denote the predicted and actual values for the *i*-th data point in the *j*-th fold, respectively.

The research approach, in this case, focuses on childhood acute leukemia whereby the neural networks' integration has increased the area of study together with decision trees and regression models. The neuronal network, whose complexity is depicted by its structure, covers the intricacies applicable to the dataset such as a number of people and different variables, [43]. In other words, it provides a way to describe the rate of change in childhood leukemia in a comprehensive manner and the neural network that acts as a computational center.

```
Mean Squared Error: 0.13
Decision Tree Rules:
|--- MC <= 2.50
|   |--- Age-E <= 2.50
|   |   |--- Age-E <= 1.50
|   |   |   |--- value: [1.00]
|   |   |--- Age-E >  1.50
|   |   |   |--- MC <= 1.50
|   |   |   |   |--- value: [1.00]
|   |   |   |--- MC >  1.50
|   |   |   |   |--- value: [1.14]
|   |--- Age-E >  2.50
|   |   |   |--- value: [1.33]
|--- MC >  2.50
|   |--- Sex <= 1.50
|   |   |--- value: [2.14]
|   |--- Sex >  1.50
|   |   |--- Age-E <= 2.50
|   |   |   |--- value: [2.20]
|   |   |--- Age-E >  2.50
|   |   |   |--- value: [2.22]
|   |   |--- Sex >  1.50
|   |   |   |--- value: [2.00]
Cross-Validation Mean
```

Fig. 9 Regression decision tree rules

At its core, this improvement is based on metrics such as mean squared error, mean absolute error, and R-squared which determine how accurate the model is in capturing complex relationships within the dataset, [44]. This integration of quantitative and =qualitative analysis includes a juxtaposition of actual versus predictive values for tumor mutation burden with the trajectory of loss of function enriches our research toolkit. Such systematic approaches with neural networks surrounding these models help us understand more about the underlying mutational landscape thus directing ongoing research towards a global comprehension regarding the dynamics of leukemia.

A neural network model, [45], trained to evaluate tumor burden rates in childhood acute lymphoblastic leukemia showed commendable performance in a variety of analyses. With an accuracy of 60%, the model shows a significant ability to predict the weight very well, achieving an accuracy of 0.60 for class 1, as outlined in Fig. 10. This accuracy shows that the prediction model for a high number of brain tumors is correct in 60% of cases. time. In addition, despite the negative R-squared score of -0.5698, suggesting that the model does not simplify the process, its proximity to zero indicates that the model prediction is not bad than the average forecast of the main, [46].

Furthermore, the model shows promising results with small root mean square error and mean absolute error values, reported at 0.6000 and 0.4667, respectively (see Table II). These measurements show that the model's predictions match the actual tumor size, highlighting its potential for clinical research. Taken together, these findings demonstrate the power of the model as a useful tool for health professionals in evaluating brain tumors in children with acute lymphoblastic leukemia. Using additional information and the integration of other data, the model shows promise for improving diagnostic findings to guide other treatment decisions in clinical practice.

V. Conclusion

With the realms of this research, the key factors associated with the mutation rate of malignant cancer cells in pediatric patients with Acute Lymphoblastic Leukemia are analyzed. The study focused on the application of machine learning workflows for mutation rate analysis, using clinical data of pediatric patients from St. Jude Children's Research Hospital with Acute Lymphoblastic Leukemia. An overview of the clinical data used is provided to elucidate the key factors present with the clinical data of the subjects. The workflow, including correlation matrix, data exploration, and efficacious artificial intelligence modeling using classification and decision tree regression algorithm, confusion and normalized confusion matrix, Principal Component Analysis with cumulative and individual variances explained, and other machine learning workflows are comprehensively discussed and diagrams are provided to show the implementation results of theses modeling techniques.

The results of this study evince the efficacy of these machine-learning techniques in analyzing the mutation rate of malignant cancer cells in pediatric patients with Acute Lymphoblastic Leukemia. Through correlation analysis and data exploration, vital information is obtained regarding the relationships between various factors and the tumor mutation rate in the clinical data of subjects.

Additionally, the application of principal component analysis, decision trees, and classification rules demonstrates the versatility and potential of machine learning models in precisely prognosticating tumor mutation rates in patients with Acute Lymphoblastic Leukemia.

Advancing, it is paramount important to expatiate the sphere of comprehensive and interpretable Artificial Intelligence models for mutation rate analysis of malignant cancer cells. Supplementing the understanding and clarity of these models will facilitate their adoption in clinical practices. Progressive studies contributing to this research

field should focus on exploring techniques to interpret and explain the decisions made by machine learning models, in order to guarantee that healthcare professionals can comprehend and rely on the results.

In conclusion, this study is a remarkable progressive step in implementing machine learning workflows for tumor mutation rate analysis. Taking into consideration the clinical evidence, and augmenting the results of this study, future research can ameliorate patient care and outcomes in the sphere of pediatric Acute Lymphoblastic Leukemia.

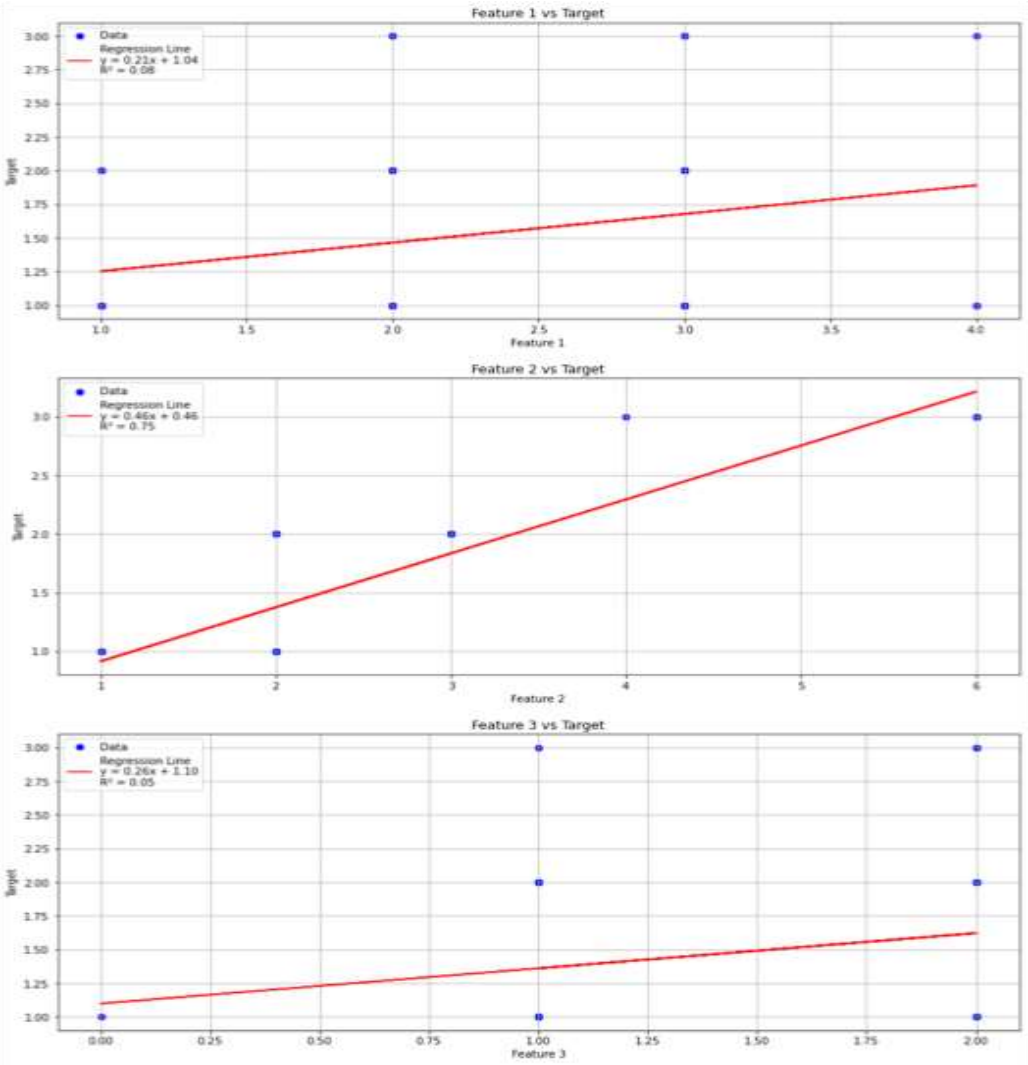

Fig. 10 Neural network model accuracy over Epochs

TABLE II. Neural Network Classification Report

	precision	recall	f1-score	support
1.0	0.63	1.00	0.75	9
2.0	0.00	0.00	0.00	5
3.0	0.00	0.00	0.00	1
accuracy			0.60	15
macro avg	0.20	0.33	0.25	15
weighted avg	0.36	600	0.45	15

ACKNOWLEDGEMENT

This work was supported in part by the Erasmus + Program for Education of the European Union through the Key Action 2 Grant (the Future Is in Applied Artificial Intelligence) under Grant 2022-1-PL01- KA220-HED000088359 (work package 5: "Piloting," activity A5.6 "Project deliverables").

References

[1] R. Machii and K. Saika, "Age-specific incidence rate of leukemia in the world,"Japanese Journal of Clinical Oncology, vol. 52, no. 1, pp. 101–102, Jan. 2022, doi: 10.1093/jjco/hyab199.

[2] National Cancer Institute. *SEER cancer statistics review 1975-2017*. [EB/OL]. (2020). Available at: https://seer.cancer.gov/csr/1975_2017/browse_csr.php?sectionSEL=28&pageSEL=sect_28_table.08 [Accessed: 09.06.2024].

[3] Hunger, S. P., Lu, X., Devidas, M., Camitta, B. M., Gaynon, P. S., Winick, N. J., Reaman, G. H., & Carroll, W. L. (2012) "Improved Survival for Children and Adolescents With Acute Lymphoblastic Leukemia Between 1990 and 2005: A Report From the Children's Oncology Group," *Journal of Clinical Oncology*, 30(14), pp. 1663–1669, https://doi.org/10.1200/JCO.2011.37.8018.

[4] Bhojwani, D., Yang, J. J., & Pui, C.-H. (2015) "Biology of Childhood Acute Lymphoblastic Leukemia," *Pediatric Clinics*, 62(1), pp. 47-60. Available at: https://doi.org/10.1016/j.pcl.2014.09.004.

[5] Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2021. *CA Cancer J Clin*. (2021) 71:7–33, doi: 10.3322/caac.21654.

[6] Xiao, Y., Xiao, L., Zhang, Y., Xu, X., Guan, X., Guo, Y., Shen, Y., Lei, X., Dou, Y., & Yu, J. (2024) "Prediction of tumor lysis syndrome in childhood acute lymphoblastic leukemia based on machine learning models: a retrospective study," *Frontiers in Oncology*, Vol. 14, 2024, https://doi.org/10.3389/fonc.2024.1337295.

[7] ERotbain, E. C., Niemann, C. U., Rostgaard, K., da Cunha-Bang, C., Hjalgrim, H., & Frederiksen, H. (2021) "Mapping comorbidity in chronic lymphocytic leukemia: impact of individual comorbidities on treatment, mortality, and causes of death," *Leukemia*, Published: 18 February 2021.

[8] J. M. Chessells, R. M. Hardisty, N. T. Rapson, and M. F. Greaves, "Acute Lymphoblastic Leukæmia in Children: Classification and Prognosis," The Lancet, vol. 310, no. 8052, pp. 1307–1309, Dec. 31, 1977, doi: 10.1016/S0140-6736(77)90361-0.

[9] F. Malard and M. Mohty, "Acute lymphoblastic leukaemia," The Lancet, vol. 395, no. 10230, pp. 1146–1162, Apr. 04, 2020, doi: 10.1016/S0140-6736(19)33018-1.

[10] "What is tumor mutation load (TML)?," Illumina Sequencing Learning Center, Illumina, [Online]. https://www.illumina.com/science/sequencing-method-explorer/kits-and-arrays/tumor-mutation-load.html. [Accessed: 03.17. 2024].

[11] Snyder, A., Makarov, V., Merghoub, T., Yuan, J., Zaretsky, J. M., Desrichard, A., Walsh, L. A., et al. (2014) "Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma," *New England Journal of Medicine*, vol. 371, pp. 2189-2199. DOI: 10.1056/NEJMoa1406498.

[12] VVan Allen, E. M., Miao, D., Schilling, B., Shukla, S. A., Blank, C., Zimmer, L., Sucker, A., Hillen, U., Geukes Foppen, M. H., Goldinger, S. M., Utikal, J., Hassel, J. C., Weide, B., Kaehler, K. C., Loquai, C., Mohr, P., Gutzmer, R., Dummer, R., Gabriel, S., Wu, C. J., Schadendorf, D., & Garraway, L. A. (2015) "Genomic correlates of response to CTLA-4 blockade in metastatic melanoma," *Science*, vol. 350, no. 6257, pp. 207-211. DOI: 10.1126/science.aad0095.

[13] Zhang, X., Lopes, I. M., Ni, J.-Q., Yuan, Y., Huang, C.-H., Smith, D. R., Chaubey, I., & Wu, S. (2021) "Long-term performance of three mesophilic anaerobic digesters to convert animal and agro-industrial wastes into organic fertilizer," *Journal of Cleaner Production*, vol. 307, 20 July 2021, p. 127271. DOI: 10.1016/j.jclepro.2021.127271.

[14] Bhute, A. ., Bhute, H. ., Pande, S., Dhumane, A., Chiwhane, S. and Wankhade, S. (2023) "Acute Lymphoblastic Leukemia Detection and Classification Using an Ensemble of Classifiers and Pre-Trained Convolutional Neural Networks", *International Journal of Intelligent Systems and Applications in Engineering*, 12(1), pp. 571–580, https://ijisae.org/index.php/IJISAE/article/view/3955.

[15] Jain, M.S., Massoud, T.F. (2020). "Predicting tumor mutational burden from histopathological images using multiscale deep learning." *Nature Machine Intelligence*, 2, 356–362. https://doi.org/10.1038/s42256-020-0190-5.

[16] AAsfaw, A., Ascha, M., Yerram, P., Reiss, S., Brake, S., & Wadé, N. B. (2022) "SA27 Comparison of Comorbidity Indices between Electronic Health Records (EHR) Derived Database and Claims Data among Patients with Metastatic Breast Cancer," *Value in Health*, vol. 25, issue 12, supplement, p. S488, December 2022. DOI: 10.1016/j.jval.2022.09.2421.

[17] Hong, N., Wen, A., Stone, D. J., Tsuji, S., Kingsbury, P. R., Rasmussen, L. V., Pacheco, J. A., Adekkanattu, P., Wang, F., Luo, Y., Pathak, J., Liu, H., & Jiang, G. (2019) "Developing a FHIR-based EHR phenotyping framework: A case study for identification of patients with obesity and multiple comorbidities from discharge summaries," *Journal of Biomedical Informatics*, vol. 99, November 2019, p. 103310. DOI: 10.1016/j.jbi.2019.103310.

[18] Albahra, S., Gorbett, T., Robertson, S., D'Aleo, G., Suseel Kumar, S. V., Ockunzzi, S., Lallo, D., Hu, B., & Rashidi, H. H. (2023) "Artificial intelligence and machine learning overview in pathology & laboratory medicine: A general review of data preprocessing and basic supervised concepts," *Seminars in Diagnostic*

Pathology, vol. 40, issue 2, March 2023, pp. 71-87.DOI: 10.1053/j.semdp.2023.02.002.

[19] BBen-Assuli, O., Heart, T., Klempfner, R., & Padman, R. (2023) "Human-machine collaboration for feature selection and integration to improve congestive heart failure risk prediction," *Decision Support Systems*, vol. 172, September 2023, p. 113982. DOI: 10.1016/j.dss.2023.113982.

[20] Navazi F, Yuan Y and Archer N. An examination of the hybrid meta-heuristic machine learning algorithms for early diagnosis of type ii diabetes using big data feature selection. Healthcare Analytics 2023; https://doi.org/10.1016/j.health.2023.100227.

[21] Uddin, S., Wang, S., Lu, H., Khan, A., Hajati, F., & Khushi, M. (2022)"Comorbidity and multimorbidity prediction of major chronic diseases using machine learning and network analytics," *Expert Systems with Applications*, vol. 205, 1 November 2022, p. 117761. DOI: 10.1016/j.eswa.2022.117761.

[22] Nikolaou, V., Massaro, S., Garn, W., Fakhimi, M., Stergioulas, L., & Price, D. (2021) "The cardiovascular phenotype of Chronic Obstructive Pulmonary Disease (COPD): Applying machine learning to the prediction of cardiovascular comorbidities," *Respiratory Medicine*, vol. 186, September 2021, p. 106528. DOI: 10.1016/j.rmed.2021.106528.

[23] AAlsaleh, M. M., Allery, F., Choi, J. W., Hama, T., McQuillin, A., Wu, H., & Thygesen, J. H. (2023) "Prediction of disease comorbidity using explainable artificial intelligence and machine learning techniques: A systematic review," *International Journal of Medical Informatics*, vol. 175, July 2023, p. 105088. DOI: 10.1016/j.ijmedinf.2023.105088.

[24] Van Vleck TT, Farrell D and Chan L. Natural language processing in nephrology. Advances in Chronic Kidney Disease 2022; 29(5): 465–471. doi: https://doi.org/10.1053/j.ackd.2022.07.001.

[25] Shouse, G., Kaempf, A., Gordon, M. J., Artz, A., Yashar, D., Sigmund, A. M., Smilnak, G., Bair, S. M., Mian, A., Fitzgerald, L. A., Bajwa, A., Jaglowski, S., Bailey, N., Shadman, M., Patel, K., Stephens, D. M., Kamdar, M., Hill, B. T., Gauthier, J., Karmali, R., Nastoupil, L. J., Kittai, A. S., & Danilov, A. V. (2023) "A validated composite comorbidity index predicts outcomes of CAR T-cell therapy in patients with diffuse large B-cell lymphoma," *Blood Advances*, vol. 7, no. 14, pp. 3516–3529. DOI: 10.1182/bloodadvances.2022009309.

[26] Ghabril, M., Gu, J., Yoder, L., Corbito, L., Ringel, A., Beyer, C. D., Vuppalanchi, R., Barnhart, H., Hayashi, P. H., & Chalasani, N. (2019) "Development and Validation of a Model Consisting of Comorbidity Burden to Calculate Risk of Death Within 6 Months for Patients With Suspected Drug-Induced Liver Injury," *Gastroenterology*, vol. 157, issue 5, November 2019, pp. 1245-1252.e3. DOI: 10.1053/j.gastro.2019.07.006.

[27] Vitzthum, L., Noticewala, S. S., Hines, P., Nguyen, C., Shen, H., & Mell, L. K. (2017) "A Web-Based Tool to Compare Comorbidity Models and Geriatric Risk-Assessment in Head and Neck Cancer Patients," *International Journal of Radiation Oncology, Biology, Physics*, vol. 99, issue 2, supplement, p. E379, October 01, 2017. DOI: 10.1016/j.ijrobp.2017.06.1508.

[28] Ayyappan, V., Chang, A., Zhang, C., Paidi, S. K., Bordett, R., Liang, T., Barman, I., & Pandey, R. (2020) "Identification and Staging of B-Cell Acute Lymphoblastic Leukemia Using Quantitative Phase Imaging and Machine Learning," *ACS Sens.*, vol. 5, issue 10, pp. 3281–3289, October 14, 2020. DOI: 10.1021/acssensors.0c01811.

[29] Rivera, G. K., Pinkel, D., Simone, J. V., Hancock, M. L., & Crist, W. M. (1993) "Treatment of Acute Lymphoblastic Leukemia -- 30 Years' Experience at St. Jude Children's Research Hospital," *New England Journal of Medicine*, vol. 329, no. 18, pp. 1289-1295, October 28, 1993. DOI: 10.1056/NEJM199310283291801.

[30] S. T. Park and J. Kim, "Trends in Next-Generation Sequencing and a New Era for Whole Genome Sequencing," *Int Neurourol J.*, vol. 20, no. Suppl 2, pp. S76–S83, Nov. 2016, doi: 10.5213/inj.1632742.371.

[31] J. F. Cutigi, A. F. Evangelista, and A. Simao, "Approaches for the identification of driver mutations in cancer: A tutorial from a computational perspective," *Journal of Bioinformatics and Computational Biology*, vol. 18, no. 03, pp. 2050016, 2020, doi: 10.1142/S021972002050016X.

[32] F. Ravandi (Section Editor)," Minimal Residual Disease Monitoring in Adult ALL to Determine Therapy," *Acute Lymphocytic Leukemias*, vol. 10, pp. 86–95, May 01, 2015.

[33] Desai, S., Rashmi, S., Rane, A., Dharavath, B., Sawant, A., & Dutt, A. (2021) "An integrated approach to determine the abundance, mutation rate and phylogeny of the SARS-CoV-2 genome," *Briefings in Bioinformatics*, vol. 22, issue 2, March 2021, pp. 1065–1075. DOI: 10.1093/bib/bbaa437. [Accessed: 09.06.2024].

[34] Asgari, P., Miri, M. M., & Asgari, F. (2022) "The comparison of selected machine learning techniques and correlation matrix in ICU mortality risk prediction," *Informatics in Medicine Unlocked*, vol. 31, 2022, p. 100995. DOI: 10.1016/j.imu.2022.100995.

[35] M. Mollaee and M. H. Moattar, "A novel feature extraction approach based on ensemble feature selection and modified discriminant independent component analysis for microarray data classification," *Biocybernetics and Biomedical Engineering*, vol. 36, no. 3, pp. 521-529, 2016, doi: 10.1016/j.bbe.2016.05.001.

[36] R. J. Lewis, "An Introduction to Classification and Regression Tree (CART) Analysis," presented at the 2000 Annual Meeting of the Society for Academic Emergency Medicine, San Francisco, CA.

[37] N. Shrestha, "Factor Analysis as a Tool for Survey Analysis," *American Journal of Applied Mathematics and Statistics*, vol. 9, no. 1, pp. 4-11, 2021. doi: 10.12691/ajams-9-1-2.

[38] Ding, L.-W., Sun, Q.-Y., Tan, K.-T., Chien, W., Thippeswamy, A. M., Yeoh, A. E. J., Kawamata, N.,

Nagata, Y., Xiao, J.-F., Loh, X.-Y., Lin, D.-C., Garg, M., Jiang, Y.-Y., Xu, L., Lim, S.-L., Liu, L.-Z., Madan, V., Sanada, M., Fernández, L. T., Preethi, H., Lill, M., Kantarjian, H. M., Kornblau, S. M., Miyano, S., Liang, D.-C., Ogawa, S., Shih, L.-Y., Yang, H., & Koeffler, H. P. (2017) "Mutational Landscape of Pediatric Acute Lymphoblastic Leukemia," *Cancer Research*, vol. 77, issue 2, pp. 390–400, January 16, 2017. DOI: 10.1158/0008-5472.CAN-16-1303.

[39] S. Tangirala, "Evaluating the Impact of GINI Index and Information Gain on Classification using Decision Tree Classifier Algorithm," *International Journal of Advanced Computer Science and Applications*, 2020, doi: 10.14569/ijacsa.2020.0110277.

[40] N. Kumari, A. K. Bhatt, R. K. Dwivedi, and R. Belwal, "Accuracy Testing of Data Classification using TensorFlow, a Python Framework in ANN Designing," presented at the IEEE International Conference on Systems, Man, and Cybernetics (SMC), Moradabad, India, Nov. 23-24, 2018, doi: 10.1109/SYSMART.2018.8746945.

[41] S. S. Rathore and S. Kumar, "A Decision Tree Regression based Approach for the Number of Software Faults Prediction," *ACM SIGSOFT Software Engineering Notes*, vol. 41, no. 1, pp. 1–6, Feb. 22, 2016, doi: 10.1145/2853073.2853083.

[42] A. Botchkarev, "Performance Metrics (Error Measures) in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology," arXiv:1809.03006 [stat.ME], 2018. [Online]. Available: https://doi.org/10.48550/arXiv.1809.03006. Journal reference: Interdisciplinary Journal of Information, Knowledge, and Management, 2019, vol. 14, pp. 45-79, doi: 10.28945/4184 [Accessed: 09.06.2024].

[43] V. Plevris, G. Solorzano, N. P. Bakas, and M. E. A. Ben Seghier, "Investigation of performance metrics in regression analysis and machine learning-based prediction models," Conference Object, Nov. 24, 2022, https://doi.org/10.23967/eccomas.2022.155.

[44] M. W. Browne, "Cross-Validation Methods," *Journal of Mathematical Psychology*, vol. 44, no. 1, pp. 108-132, Mar. 2000, doi: 10.1006/jmps.1999.1279.

[45] M. Islam, G. Chen, and S. Jin, "An Overview of Neural Network," *American Journal of Neural Networks and Applications*, vol. 5, no. 1, pp. 7-11, Jun. 29, 2019, doi: 10.11648/j.ajnna.20190501.12.

[46] J. Karch, "Improving on Adjusted R-Squared," *Collabra: Psychology*, vol. 6, no. 1, p. 45, Sep. 29, 2020, doi: 10.1525/collabra.343.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

- Martsenyuk Vasyl contributed to conceptualization, formal analysis, funding acquisition, methodology, project administration, supervision, validation, writing – original draft, writing – review & editing.
- Abubakar Sadiq contributed to conceptualization, methodology, software development, validation, formal analysis, investigation, data curation, visualization, writing - original draft, writing – review & editing.
- Sverstiuk Andriy contributed to conceptualization, methodology, validation, project administration, supervision, writing - original draft, writing – review & editing.
- Dimitrov Georgi contributed to conceptualization, formal analysis, methodology, supervision, validation, writing – original draft, writing – review & editing.
- Gancarczyk Tomasz contributed to conceptualization, formal analysis, methodology, supervision, validation, writing – original draft, writing – review & editing.

All authors have read and agreed to the published version of the manuscript.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

This work was supported in part by the Erasmus + Program for Education of the European Union through the Key Action 2 Grant (the Future is in Applied Artificial Intelligence) under Grant 2022-1-PL01- KA220-HED000088359 (work package 5: “Piloting,” activity A5.6 “Project deliverables”)

Conflict of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.