BioMedInformatics

MDPI

Article

Predictions of Programmed Cell Death Ligand 1 Blockade Therapy Success in Patients with Non-Small-Cell Lung Cancer

Taksh Gupta *, Tamara Qawasmeh and Serena McCalla

The Lawrenceville School, Lawrenceville, NJ 08648, USA; drmccalla@iresearchinstitute.com (S.M.)
* Correspondence: taksh.tg@gmail.com; Tel.: +1-609-495-4004

Abstract: Lung cancer is responsible for the most cancer deaths worldwide, with non-small-cell lung cancer (NSCLC) making up 80% of cases. Some genetic factors leading to NSCLC development include genetic mutations and Programmed Cell Death Ligand 1 (PD-L1) expression. PD-L1 proteins are targeted in an NSCLC treatment called PD-L1 blockade therapy (immune therapy). However, this treatment is effective in a low percentage of patients. This study aimed to create machine learning models to use features, like the number of mutations and the number of PD-L1 proteins in cancer cells, along with others, to predict whether a patient will receive clinical benefits from immune therapy. This was carried out by downloading and merging datasets from cbioportal.org to create a sample size for the model. Features that were highly correlated with clinical benefits were identified. Three machine learning models (Gaussian naïve Bayes, decision tree, and logistic regression) were created using these features to predict clinical benefits in patients, and each model's accuracy was evaluated. All three models had accuracy rates between 55 and 85%, with two of the models averaging an accuracy rate of around 75%. Doctors can use these models to more accurately predict whether immune therapy treatment is likely to work in a patient before prescribing it to them.

Keywords: non-small-cell lung cancer; PD-L1 blockade therapy; immune therapy; immune checkpoint inhibitors; PD-L1 proteins; Gaussian naïve Bayes; decision tree; logistic regression; machine learning

Citation: Gupta, T.; Qawasmeh, T.; McCalla, S. Predictions of Programmed Cell Death Ligand 1 Blockade Therapy Success in Patients with Non-Small-Cell Lung Cancer. *BioMedInformatics* **2023**, *3*, 1060–1070. https://doi.org/10.3390/biomedinformatics3040063

Academic Editors: José Machado and Alexandre G. De Brevern

Received: 1 July 2023
Revised: 20 August 2023
Accepted: 23 October 2023
Published: 7 November 2023

1. Introduction

Spread of Non-Small-Cell Lung Cancer

Non-small-cell lung cancer (NSCLC) is responsible for the most cancer deaths worldwide. Unlike small-cell lung cancer, NSCLC commonly occurs in smokers and non-smokers and makes up 80% of lung cancer cases [1]. For non-smokers, there are both environmental and genetic risk factors for developing NSCLC [2]. Some genetic factors that can lead to NSCLC include mutations and the level of PD-L1 protein expression in cancer cells [3,4]. The PD-L1 protein is a protein present in cells that acts as brakes for immune system cells, such as T cells. The PD-L1 protein in cancer cells binds to the PD-1 protein on T cells [5]. When this binding occurs, the T cell is not activated and knows not to attack the cell with the PD-L1 protein. This results in cells with higher PD-L1 levels being less likely to be attacked by T cells [5–7] (Figure 1). This is exploited by cancer cells, some of which have been observed to have very high levels of PD-L1, which causes them not to be attacked by T cells [5–7]. This mechanism allows the cancer to escape the body's natural response to cancerous cells and grow and spread at a more rapid rate through the patient's body.

The PD-L1 protein on cancer cells is targeted by a type of NSCLC treatment, called PD-L1 blockade therapy. Immune therapy drugs, known as immune checkpoint inhibitors (ICIs), bind to PD-L1 proteins. This binding prevents the PD-L1 proteins in cancer cells from being able to attach to PD-1 proteins on T cells [8–10]. As a result, T cells are left free to launch immune attacks on the cancer cells, being able to kill them before the cancer cells can grow and multiply (Figure 2). ICI-based treatment has shown to be effective with long-lasting results, but it also has various limitations. These can include immunoresistance

and adverse effects for patients for whom the treatment is unsuccessful. The largest of these limitations is that ICI treatment has only proven to be effective in about 20–40% of patients [9]. This makes it important to understand whether ICI treatment will be beneficial for a patient before prescribing it to them. However, it is currently difficult to predict if a patient will benefit from ICI treatment, making it less cost-effective overall [11]. Hellmann et al. established two features that are associated with the success of ICI treatment, also known as a clinical benefit. These two features are the number of mutations in cancer cells and PD-L1 expression, which were both shown to be positively correlated with a clinical benefit [12].

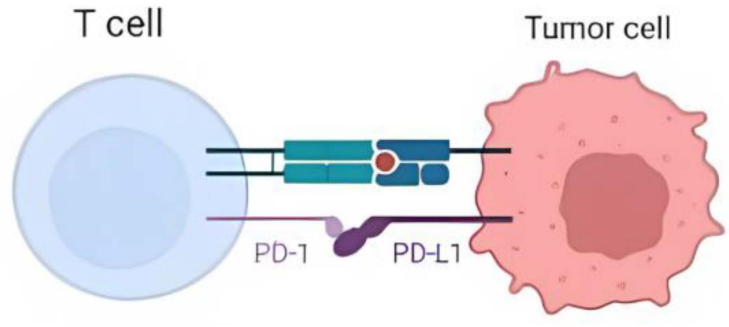

Figure 1. Diagram of PD-L1 and PD-1 binding; this PD-L1 and PD-1 binding process prevents the T-cell from killing the tumor cell.

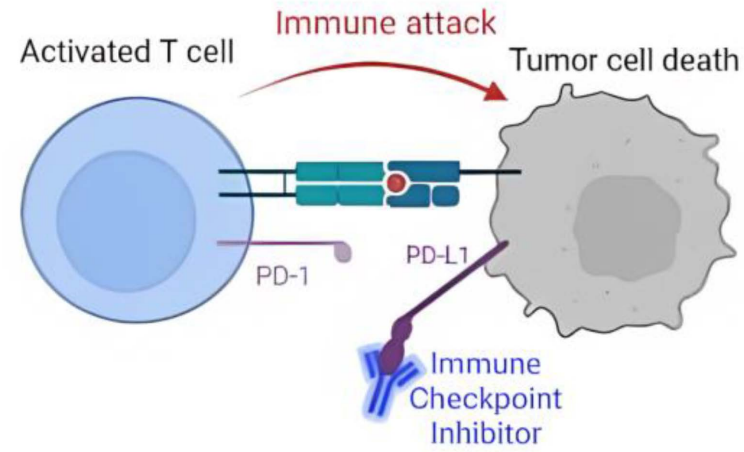

Figure 2. Diagram of ICI preventing PD-L1 and PD-1 binding; the ICI binding to the PD-L1 protein allows the T-cell to attack the tumor cell.

The first objective of this study was to determine if the correlations outlined above are present in datasets other than the one presented in the study by Hellmann et al. This would strengthen the validity of these correlations. The next goal was to understand which other demographic, environmental, or genetic features would be the best to use to predict clinical benefits in patients. Supervised machine learning models would then be created using these features to predict a patient's clinical benefit from ICI treatment. This would allow clinical benefits to be predicted before treatment begins, improving the cost-effectiveness of the treatment. Finally, the performance of this model will be evaluated by looking at both its accuracy and how well it fits the data. These models have the potential to help doctors decide which patients may benefit from PD-L1 blockade therapy treatment, improving treatment efficacy for patients with NSCLC.

2. Materials and Methods

2.1. Materials

There were two data sources used for this project. The primary dataset (MSK) is from a study conducted with 75 NSCLC patients [13]. It was downloaded as a.tsv file from https://www.cbioportal.org/study/clinicalData?id=nsclc_mskcc_2018 (accessed on 25 October 2023).

The secondary dataset (MSKCC) is from a study conducted with 240 NSCLC patients [14]. It was downloaded as a.tsv file from https://www.cbioportal.org/study/clinicalData?id=nsclc_pd1_msk_2018 (accessed on 25 October 2023).

These datasets were chosen for a variety of reasons. Firstly, the studies that produced the datasets were conducted by much of the same group of researchers at Memorial Sloan Kettering Cancer Institute. As a result, the methodology for collecting these datasets was similar in both studies, as Hellmann et al. state when citing the Rizvi et al. study [12]. These similar methodologies mean that the data are collected in a similar way, so comparisons between the studies can be made with more accuracy. Additionally, both studies looked specifically at what factors lead to clinical benefits, meaning that there were a multitude of these factors in the datasets that could be used to train the models.

2.2. Methodology

The MSK and MSKCC datasets were downloaded from the cBioPortal website. The datasets were merged into one combined dataset to provide more data to train and test the model. Once this was complete, the merged dataset was split into 80% training data and 20% test data. These steps provided enough data to train the model while also allowing a large enough test data size to reliably test the models' accuracy. To decide which features the model would use to determine clinical benefits, a Pearson correlation coefficient test was run on the clinical benefit variable to see which features had the highest correlation to this variable. The features with the six highest correlation coefficients were used to train the model. Six features were chosen because this provided enough features for the model to accurately predict clinical benefits without overcomplicating the model and risking overfitting to the training data, a common issue with machine learning programs. Leave-one-out cross-validation (LOOCV) was used to measure the model fitness in relation to the data. LOOCV works in a dataset with "n" entries by training the model with "n-1" entries and testing it on the final entry. It then repeats this process "n" times to use each entry as the test entry. This process is used to compute a score based on how accurate and well fitted the data and the model are. This score is found by performing LOOCV and then calculating the mean absolute error (MAE) of the model. Three different supervised machine learning models were created to fit the training data. The first was a Gaussian naïve Bayes (GNB) model. The GNB model works by assuming the data for each feature follows a normal distribution and classifying each new entry by looking at the likelihood that each feature fits into this normal distribution [15]. The second model created was a decision tree model. The decision tree model works by creating conditions based on specific features in the data and making a tree-like representation of these conditions. When a new entry is given to the model, it goes down a specific path on the tree based on if each condition is true or false for that entry. After enough conditions, the model classifies the entry based on the path it took through the 'branches' of the 'tree' [16]. The final model was a logistic regression model. Logistic regression models classify data into binary outcomes by fitting a logistic curve to the dataset's features [17]. The three models were trained using a part of the merged dataset allocated for training, and then tested on the remaining data. The accuracy levels of each model (in classifying the patient on the MSK, MSKCC, and merged datasets) were recorded. While these three models have limitations in scope and efficiency, they would be sufficient given the sample sizes in the studies that produced the datasets. Additionally, these models perform better for classification tasks than continuous value tasks, which fits the objectives of this study. These three models were also chosen due to their accessibility and the limited preprocessing needed. Pandas, NumPy, Sklearn, Seaborn, and Matplotlib

were the main packages used. Pandas was used for data manipulation. Numpy was used for numerical computation. Sklearn was used to create the machine learning models. Seaborn and Matplotlib were used to create the visualizations. All programs were run in Python version 3.7.13 using a Google Colabatory notebook.

3. Results and Discussions

3.1. Background Information for Datasets

The first tests provide context for some of the demographic and genetic features of the merged dataset. The merged dataset contains patients of a wide range of ages (38% of patients were 60–70 years old, 28% were 50–60 years old, and 20% were 70–80 years old). The other 15% of patients were either above 80 years old or below 50 years old (Figure 3). It is important to note the ages of the participants in the studies because a younger or older sample may include patients with different features. For example, a younger demographic of patients may consist of fewer smokers than this dataset has [18]. The following visualization is a pie chart to show what percentage of patients received clinical benefits from the ICI treatment. Approximately 40% of the patients in the merged dataset received a clinical benefit from the treatment (Figure 4). This is ideal for creating a supervised learning model because it provides a large amount of labeled data for each outcome (clinical benefit and no clinical benefit). Next, a histogram showing the mutation count in each dataset was created. The mutation counts in patients in both the MSK and MSKCC datasets were clustered on the lower side of the range of mutation counts (Figure 5). The MSK dataset had patients with mutation counts over 600, while the MSKCC dataset did not have any patients with a mutation count above 100 (Figure 6). This difference in mutation counts is important to note because the MSKCC dataset had a larger sample size compared to the MSK dataset. Therefore, the mutation count will be skewed toward the MSKCC dataset's lower average count. This trend of mutations being clustered on the lower end of the range was present when a histogram of the mutation counts in the merged dataset was also created (Figure 7).

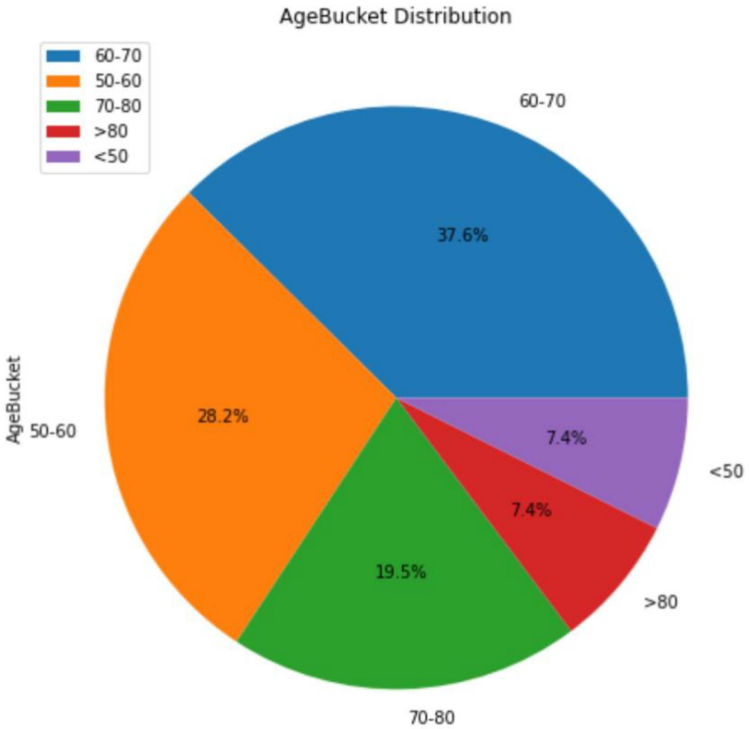

Figure 3. Pie chart of patient ages in the merged dataset; a large age range was represented in the merged dataset with most patients being between 50 and 70 years old.

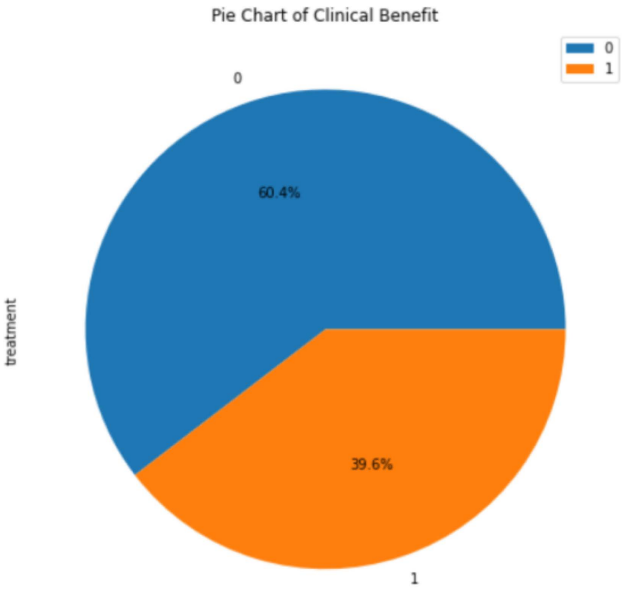

Figure 4. Pie chart showing clinical benefits for the merged dataset; about 40% of the patients received a clinical benefit from the treatment.

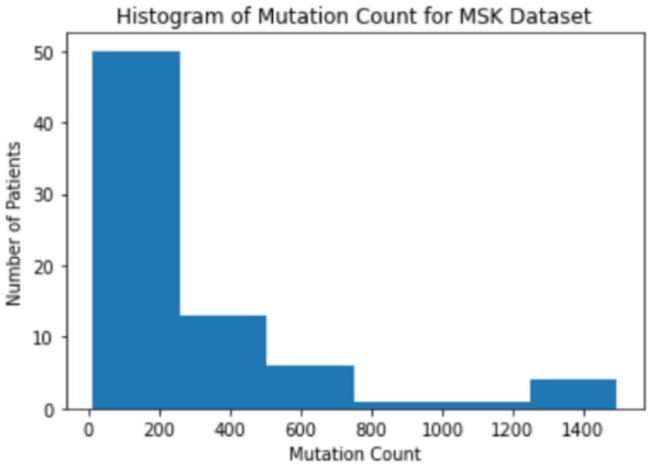

Figure 5. Histogram showing mutation counts (MSK); this database includes outlier patients with mutation counts well over 600.

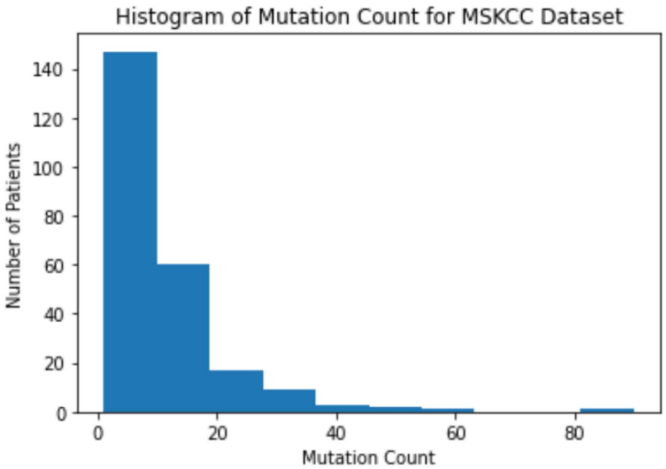

Figure 6. Histogram of mutation count (MSKCC); this dataset does not include any patients with a mutation count above 100.

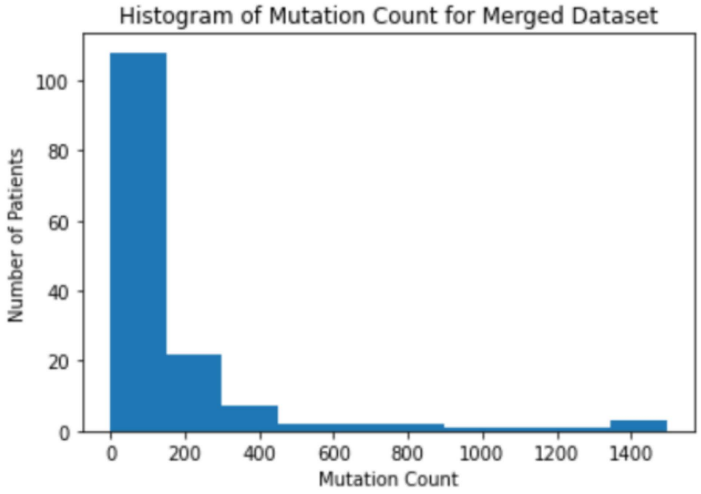

Figure 7. Histogram of mutation count (merged); most mutations were clustered in the 0–200 range.

3.2. The Correlation between Mutation Counts and Clinical Benefits

Once this background information was established, the mutation count variable was examined to see if they it significantly correlated with clinical benefits. It had already been previously shown in a study that produced an MSK dataset that there was a significant correlation between mutation counts and clinical benefits, so this dataset was not tested. In the MSKCC dataset, it was shown that patients that received a clinical benefit from the treatment had a significantly higher mutation count on average than those who did not (Figure 8). This same finding was echoed in the merged dataset (Figure 9). This supports the findings of Hellmann et al. in the study that produced the MSK dataset. However, seeing this trend in a larger dataset helped to confirm this correlation. This also means that mutation count is an important feature that can be used to train the models since it strongly correlates to clinical benefits in patients.

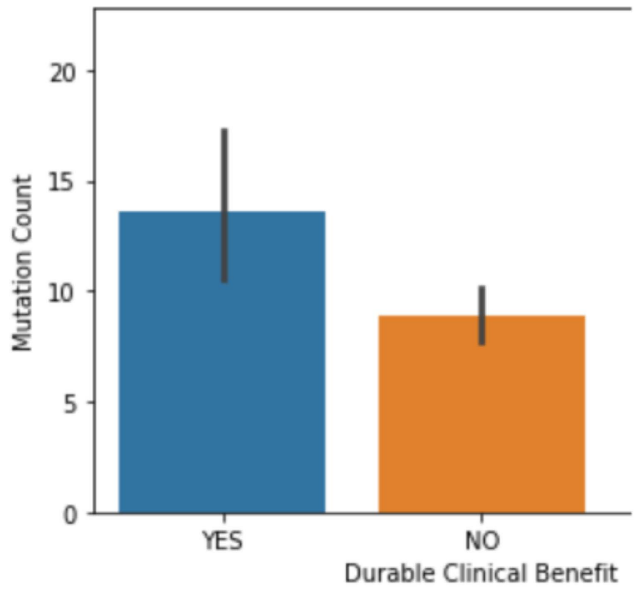

Figure 8. Bar chart comparing clinical benefits and mutation counts in the MSKCC dataset; patients with a clinical benefit had significantly higher mutation counts on average in the MSKCC dataset. The error bars do not overlap, showing that this relationship is statistically significant (p-value < 0.001).

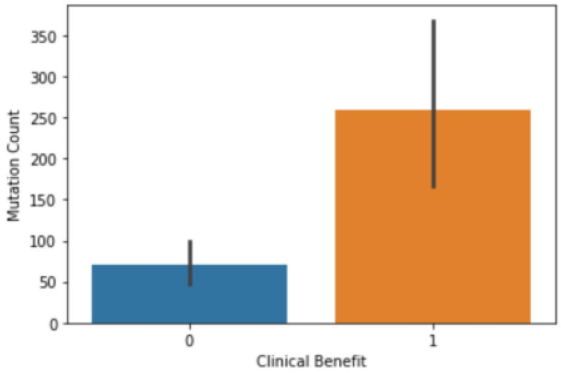

Figure 9. Bar chart comparing clinical benefits and mutation count (merged); patients that received a clinical benefit (1) had significantly higher mutation counts on average in the merged dataset. The error bars do not overlap, showing that this relationship is statistically significant (p-value < 0.001).

3.3. Correlation Coefficient Test

A correlation coefficient test was run for all the features in the merged dataset to see which features had the highest scores. Of the eight features tested, six features with high correlation coefficients were found. These features were nonsynonymous mutation burden, predicted neoantigen burden, mutation count, tumor mutation burden, PD-L1, and smoking history (Table 1). This is an interesting result because the four features with the highest correlation were all found to be related to the number of mutations present in cancer cells. This indicates that mutation counts in cancer cells are highly correlated with the success of ICI treatment, a finding that is supported by additional research in this field [12]. The other two features selected were PD-L1 expression and the smoking history of the patient. It should be expected that these features are highly correlated with treatment success, given the important role of the PD-L1 protein in cancer cells discussed earlier and smoking's causal effect on NSCLC. In fact, it is interesting that PD-L1 expression is not more highly correlated with clinical benefits, as the PD-L1 proteins on cancer cells are the target of the treatment. Once these six features were identified, they were used to train each of the models. It is important to note that the MSKCC dataset did not include the nonsynonymous mutation burden or predicted neoantigen burden features, so the MSKCC rows had null values in these columns in the merged dataset. While there are limitations to using this method for selecting features, it provides a straightforward way to train the models while also preserving one of the study's objectives to identify specific features that can be used to predict clinical benefits. These features can be used independently from the machine learning models for specific patients for whom the data for only one or two features are available. This allows doctors to achieve some of this predictive ability without the need for as much data from each patient.

Table 1. Correlation coefficient for machine learning features; the six highest correlation coefficients were used for the model, with the four highest being related to mutation counts in cancer cells.

Feature	Correlation Coefficient
Nonsynonymous Mutation Burden	0.3730
Predicted Neoantigen Burder	0.3392
Mutation Count	0.3261
Tumor Mutation Burden	0.2655
PD-L1	0.2362
Smoking History	0.1445

3.4. Leave-One-Out Cross-Validation

Once each model was trained, a LOOCV score was used to evaluate how well each model fits the data. This score was evaluated by performing LOOCV and then calculating the MAE for each model on each dataset. All three models had LOOCV scores close to zero, meaning that they were well fitted to the data (Table 2). The decision tree model was the least well fitted of the three models on average, but the difference between the LOOCV scores for each model was marginal. It is also worth noting that the models were better fitted to the MSKCC dataset than the MSK dataset. This could be because the MSKCC dataset had fewer features or a larger sample size than the MSK dataset.

Table 2. LOOCV scores (mean absolute error) of different models when different datasets were used; all three models were well fitted to the datasets with low LOOCV scores, with the decision tree model being the least well fitted of the three.

Models/Datasets	MSK	MSKCC	Merged
GNB	0.4107	0.2500	0.3025
Decision Tree	0.5000	0.3529	0.4958
Logistic Regression	0.3928	0.2352	0.3109

3.5. Machine Learning Models

Once the fit of the models was established, each model was trained and tested. Each model used 80% of the data in the dataset for training and 20% for testing. Each model was tested on the MSK, MSKCC, and the merged datasets. The models' accuracy for all three datasets were consistently above 50% (Table 3). The GNB and logistic regression models consistently were accurate at a rate of 70–80%, while the decision tree model had a larger disparity in its accuracy across the datasets. The three models were, on average, slightly more accurate on the MSK dataset than the MSKCC or merged datasets. This could be because the MSKCC dataset complicated the model in a few different ways. First, the MSKCC dataset held many null values for the PD-L1, nonsynonymous mutation burden, and predicted neoantigen burden columns (three of the primary key features). Additionally, this dataset had lower correlation coefficients for each feature than the MSK or merged dataset. However, the MSKCC dataset's larger sample size may be why the GNB model performed best on this dataset. Regardless, the models were consistently accurate above 50% for all three datasets.

Table 3. The accuracy of machine learning models when different datasets were used; all three models showed accuracy levels above 50% for all three datasets. The GNB and logistic regression models proved to be the most accurate on average.

Models/Datasets	MSK	MSKCC	Merged
GNB	71.43%	77.78%	73.33%
Decision Tree	85.71%	55.56%	73.33%
Logistic Regression	78.57%	77.78%	70.00%

4. Conclusions

4.1. Conclusions

The merged dataset contained data from patients of a wide range of ages, although approximately 85% of the patients were between the ages of 50 and 80. In this merged dataset, 39.6% of patients received clinical benefits from the treatment, which was ideal for the creation of machine learning models. Next, histograms were created to show mutation counts for the MSK, MSKCC, and merged datasets, which showed that most patients had mutation counts below two hundred. The relationship between mutation counts and clinical benefits was then analyzed. In both the MSKCC and merged datasets, it was

found that patients who received clinical benefits from the PD-L1 blockade therapy had significantly more mutations than those who did not. This indicates that mutation counts and clinical benefits are correlated in the MSKCC and merged datasets. This validates the finding in Hellmann et al.'s study (which produced the MSK dataset) that these two variables are correlated. A Pearson correlation coefficient test was run to find features that were highly correlated with clinical benefits in the merged dataset. Six total key features (nonsynonymous mutation burden, predicted neoantigen burden, mutation count, tumor mutation burden, PD-L1, and smoking history) were found to be highly correlated with clinical benefits in the merged dataset, with correlation coefficients ranging from 0.37 to 0.15. These six key features were then used to create three supervised machine learning models. The models use these six features to predict whether treatment results in clinical benefits for a patient. These models were a Gaussian naïve Bayes model, a decision tree, and a logistic regression model, each of which specialize in classification tasks. The merged dataset was split into 80% training data and 20% testing data. This gave the models enough data to learn without overfitting the training data. Once the models were trained, LOOCV was used to assess the models' accuracy and overall fit to the data. Each of the models had low LOOCV scores on each of the datasets, which shows that the models were well fitted to the data. These models were then applied to the MSK, MSKCC, and merged datasets to predict patient clinical benefits. All three models were found to accurately predict clinical benefits greater than 50% of the time, averaging an accuracy rate of 73.72% across the three models and the three datasets. The accuracy rates of the models ranged from 55% to 85%, both of which were achieved using the decision tree model. The GNB and logistic regression models were more consistent, with accuracies ranging between 70 and 80%. Although they were not perfect, all three of these models were accurate at a higher rate than random chance, showing that machine learning models can have practical applications in predicting clinical benefits for patients with NSCLC using PD-L1 blockade therapy treatment.

4.2. Future Investigations

In the future, even more accurate models can be created using larger datasets with additional patient data. The sample sizes for the datasets used in this study were not very large, so it stands to reason that a larger sample size could provide even more accurate models. Additionally, datasets with more features and higher correlations with clinical benefits could be used to provide the models with more variables to classify the patients. Another way of increasing the accuracy of the models could be to use more complex classification machine learning models. This study used GNB, decision tree, and logistic regression models due to their lack of necessary preprocessing time and the ease of implementation. However, there are more complex machine learning models that could be used to provide more accurate results. In the future, it is also important to understand why certain features are correlated with clinical benefits. It is currently not clear why a higher mutation count or higher PD-L1 expression leads to better results in treatment [12]. Understanding why these correlations exist could further advance ICI treatment and its effectiveness in a larger pool of patients. It could also help find more features that correlate with clinical benefits, creating more accurate models in the future.

4.3. Applications

Immune therapy treatments such as ICIs have shown great potential in helping patients with NSCLC. However, ICIs cost patients and insurers millions of dollars each time the treatment is used [19]. As a result, knowing when a patient is likely to benefit from ICIs is extremely important, especially since only 20–40% of patients receive a clinical benefit from this treatment. Previously, understanding whether a patient would benefit from ICI treatment was a large challenge which slowed down the implementation of this treatment. Predicting which patients will receive a clinical benefit would dramatically increase the cost-effectiveness of this treatment and could prevent any potential adverse effects that may come with ICI treatment without clinical benefits [20–24]. Using these models, doctors

can better understand in advance whether a patient is likely to benefit from PD-L1 blockade therapy to maximize the effectiveness of these powerful treatments.

Author Contributions: Conceptualization, T.G.; Methodology, T.G. and T.Q.; Software, T.G.; Formal Analysis, T.G.; Investigation, T.G.; Resources, T.G.; Data Curation, T.G. and T.Q.; Writing—original draft preparation, T.G.; Writing-review and editing, T.G. and T.Q.; Visualization, T.G.; Supervision, T.Q. and S.M.; Project Administration, S.M.; Funding, Not applicable. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Datasets used for this study can be found at https://www.cbioportal.org/study/clinicalData?id=nsclc_pd1_msk_2018 and: https://www.cbioportal.org/study/clinicalData?id=nsclc_mskcc_2018.

Acknowledgments: I would like to acknowledge Tamara Qawasmeh for her mentorship throughout the process of conducting the research and writing this paper. It would not have been possible without her guidance and expertise.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Torre, L.A.; Siegel, R.L.; Jemal, A. Lung cancer statistics. *Adv. Exp. Med. Biol.* **2016**, *893*, 1–19.
2. Gene Mutations in Non-Small-Cell Lung Cancer. Available online: https://www.webmd.com/lung-cancer/story/nsclc-gene-mutations (accessed on 29 July 2022).
3. Cheng, P.-C.; Cheng, Y.-C. Correlation between familial cancer history and epidermal growth factor receptor mutations in Taiwanese never smokers with non-small cell lung cancer: A case-control study. *J. Thorac. Dis.* **2015**, *7*, 281–287. [PubMed]
4. Ettinger, D.S.; Akerley, W.; Bepler, G.; Blum, M.G.; Chang, A.; Cheney, R.T.; Yang, S.C. Non-small-cell lung cancer. *Nat. Rev. Dis. Primers* **2015**, *1*, 15009.
5. NCI Dictionary of Cancer Terms. National Cancer Institute, 2 February 2011. Available online: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/pd-l1 (accessed on 29 July 2022).
6. PDL1 (Immunotherapy) Tests. Available online: https://medlineplus.gov/lab-tests/pdl1-immunotherapy-tests/ (accessed on 29 July 2022).
7. Gavrielatou, N.; Shafi, S.; Gaule, P.; Rimm, D.L. PD-L1 expression scoring: Non-interchangeable, non-interpretable, neither, or both. *J. Natl. Cancer Inst.* **2021**, *113*, 1613–1614. [CrossRef]
8. Immune Checkpoint Inhibitors. National Cancer Institute, 24 September 2019. Available online: https://www.cancer.gov/about-cancer/treatment/types/immunotherapy/checkpoint-inhibitors (accessed on 29 July 2022).
9. Doroshow, D.B.; Bhalla, S.; Beasley, M.B.; Sholl, L.M.; Kerr, K.M.; Gnjatic, S.; Hirsch, F.R. PD-L1 as a biomarker of response to immune-checkpoint inhibitors. *Nat. Rev. Clin. Oncol.* **2021**, *18*, 345–362. [CrossRef] [PubMed]
10. Akinleye, A.; Rasool, Z. Immune checkpoint inhibitors of PD-L1 as cancer therapeutics. *J. Hematol. Oncol.* **2019**, *12*, 92. [CrossRef] [PubMed]
11. Verma, V.; Sprave, T.; Haque, W.; Simone, C.B.; Chang, J.Y.; Welsh, J.W.; Thomas, C.R. A systematic review of the cost and cost-effectiveness studies of immune checkpoint inhibitors. *J. Immunother. Cancer* **2018**, *6*, 128. [CrossRef]
12. Hellmann, M.D.; Nathanson, T.; Rizvi, H.; Creelan, B.C.; Sanchez-Vega, F.; Ahuja, A.; Wolchok, J.D. Genomic features of response to combination immunotherapy in patients with advanced non-small-cell lung cancer. *Cancer Cell* **2018**, *33*, 843–852.e4. [CrossRef] [PubMed]
13. CBioPortal for Cancer Genomics. 2018. Available online: https://www.cbioportal.org/study/clinicalData?id=nsclc_pd1_msk_2018 (accessed on 22 October 2023).
14. CBioPortal for Cancer Genomics. 2018. Available online: https://www.cbioportal.org/study/clinicalData?id=nsclc_mskcc_2018 (accessed on 22 October 2023).
15. 1.9. Naive Bayes. Scikit. (n.d.-b). Available online: https://scikit-learn.org/stable/modules/naive_bayes.html (accessed on 22 October 2023).
16. 1.10. Decision Trees. Scikit. (n.d.-c). Available online: https://scikit-learn.org/stable/modules/tree.html (accessed on 22 October 2023).
17. 1.1. Linear Models. Scikit. (n.d.-a). Available online: https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression (accessed on 22 October 2023).

18. CDC. Current Cigarette Smoking among Adults in the United States, Centers for Disease Control and Prevention. 16 March 2022. Available online: https://www.cdc.gov/tobacco/data_statistics/fact_sheets/adult_data/cig_smoking/index.htm (accessed on 29 July 2022).
19. Bailey, C.; MIIA Health Trust Manager. Gene Therapies Offer Breakthrough Results but Extraordinary Costs. Massachusetts Municipal Association (MMA), 18 March 2020. Available online: https://www.mma.org/gene-therapies-offer-breakthrough-results-but-extraordinary-costs/ (accessed on 29 July 2022).
20. Kim, H.; Liew, D.; Goodall, S. Cost-effectiveness and financial risks associated with immune checkpoint inhibitor therapy. *Br. J. Clin. Pharmacol.* **2020**, *86*, 1703–1710. [CrossRef] [PubMed]
21. Ding, H.; Xin, W.; Tong, Y.; Sun, J.; Xu, G.; Ye, Z.; Rao, Y. Cost effectiveness of immune checkpoint inhibitors for treatment of non-small cell lung cancer: A systematic review. *PLoS ONE* **2020**, *15*, e0238536. [CrossRef] [PubMed]
22. Iivanainen, S.; Koivunen, J.P. Possibilities of improving the clinical value of immune checkpoint inhibitor therapies in cancer care by optimizing patient selection. *Int. J. Mol. Sci.* **2020**, *21*, 556. [CrossRef] [PubMed]
23. Dijkstra, K.K.; Voabil, P.; Schumacher, T.N.; Voest, E.E. Genomics- and transcriptomics-based patient selection for cancer treatment with immune checkpoint inhibitors: A review. *JAMA Oncol.* **2016**, *2*, 1490–1495. [CrossRef] [PubMed]
24. Ding, H.; Xin, W.; Tong, Y.; Sun, J.; Xu, G.; Ye, Z.; Rao, Y. Adverse effects of immune-checkpoint inhibitors: Epidemiology, management and surveillance. *Nat. Rev. Clin. Oncol.* **2019**, *16*, 563–580.