

## PREDICTING IMMUNOGENICITY IN MURINE HOSTS WITH USE OF RANDOM FOREST CLASSIFIER

Anna Marciniak<sup>1,2</sup>, Martyna Tarczewska<sup>2</sup>, Sylwester Kloska<sup>1</sup>

<sup>1</sup> Faculty of Medicine, Nicolaus Copernicus University in Toruń,  
Ludwik Rydygier Collegium Medicum  
ul. Jagiellońska 13-15, 85-067 Bydgoszcz Poland  
e-mail: {503015,503013}@stud.umk.edu.pl

<sup>2</sup> Faculty of Telecommunications, Computer Science and Electrical Engineering,  
UTP University of Science and Technology,  
Al. prof. S. Kaliskiego 7, 85-796 Bydgoszcz, Poland  
e-mail: {annmar00,martar003}@utp.edu.pl

*Summary:* Biomedical data are difficult to interpret due to their large amount. One of the solutions to cope with this problem is to use machine learning. Machine learning can be used to capture previously unnoticed dependencies. The authors performed random forest classifier with entropy and Gini index criteria on immunogenicity data. Input data consisted of 3 columns: epitope (8-11 amino acids long peptide), major histocompatibility complex (MHC) and immune response. Presented model can predict the immune response based on epitope-MHC complex. Achieved results had accuracy of 84% for entropy and 83% for Gini index. The results are not fully satisfying but are a fair start for more complexed experiments and could be used as an indicator for further research.

**Keyword:** Random Forest Classifier, Immunogenicity, Machine Learning, Entropy, Gini index

### 1. INTRODUCTION

Biological data is a term that has multiple meanings. It can be used to describe proteomics and genomics data, as well as experimental biology data and patient clinical/diseases data. The screening methods of various studies and experiments provide enormous amounts of new data that must be analyzed. There are techniques that provide data on the level of gene expression or the genomic sequence of various organisms. One of the significant complications during analysis this type of data is that the format of the data and the way it is stored differ. This is the result of various devices are used to obtain them, and various formats and naming methods are used for recording. Moreover, further difficulties in dealing with data are their enormous amount and complex relationships between the results of different studies, providing many details about the subject of the research [1]. Biomedical data are difficult to interpret due to their large amount, so the use of machine learning can facilitate the analysis process as well as capture previously unnoticed dependencies. Therefore, lead to progress in the treatment of certain diseases.

Random forest is a technique used in modeling forecasts and behavioral analyzes and is based on decision trees. It contains many decision trees that represent a separate case of classification of data entered a random forest. The random forest technique considers cases individually, considering the majority vote as the chosen forecast (Fig. 1) [6, 11].

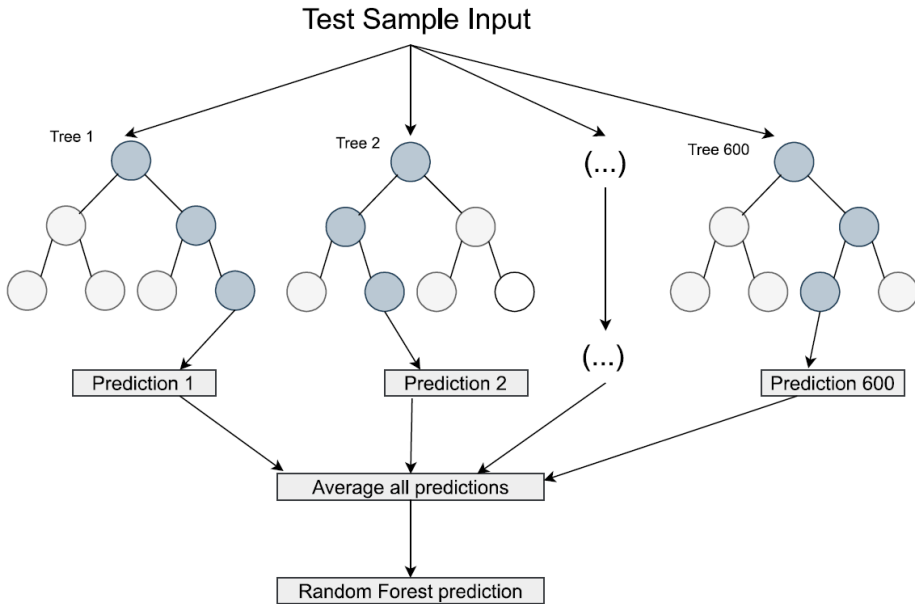


Fig. 1. Visualization of Random Forest prediction algorithm (Source: own study)

Decision trees are a popular method for various machine learning tasks [7, 12, 13]. Tree learning is a method that is the closest to fulfil the requirements of standard data mining procedure. Due to the way this process takes place, decision trees are resistant to all kinds of transformations and do not consider unnecessary data that could negatively affect the quality of results. However, the downside of decision trees is that they do not get as accurate results as other machine learning methods [5, 8].

‘Trees’ that are very deep tend to learn highly irregular patterns: they over-adapt to training sets, i.e., have a low load but a very large variance. Random forests are means of averaging many deep decision trees trained on different parts of the same training set to reduce variance. This is done at the expense of a slight increase in load and some loss of interpretability, but in general significantly increases performance in the final model [10].

Bootstrapping is a sampling technique in which one can randomly sample, replacing data from a dataset. During bootstrapping one can use only about 2/3 of the data. About 1/3 of the data is not used in the model and can be conveniently used as a test kit. The final predicted value is the average value of all decision trees. One decision tree has a large variance (it tends to overlap), so by connecting many weak individuals with strong individuals, we average the variance. That is the majority of votes.

Random forest streamlines trees by introducing a division into a random subset of features (Fig. 2). This means that with each tree subdivision, the model includes only a small subset of the features, not all the features of the model. That is, from the set of

available features  $n$ , the subset of  $m$  features ( $m$  is the square root of  $n$ ) is selected randomly. It is important that the variance can be averaged.

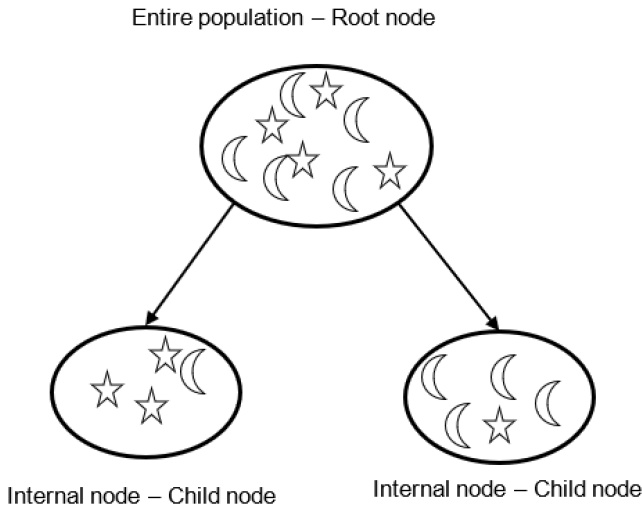


Fig. 2. Visualization of dividing dataset to subsets based on features (Source: own study)

The choice of Random Forest Classifier was dictated by the following advantages:

- predictive performance can compete with the best supervised learning algorithms;
- they provide reliable estimation of function validity;
- they offer efficient estimates of test error without the cost of retraining the model associated with cross-validation.

The model we propose is based on data from the examination of the immune response to anticancer drugs performed on the mouse model. The conducted experiment aims to predict the immune system response in order to be able to create a targeted therapy and improve the existing methods of cancer treatment.

## 2. METHODS

Input data were prepared as one .tsv file. Data fragment is shown in Fig. 3. Data consist of 3 columns:

1. Peptide – amino acid sequence. There are 20 amino acids in nature, each of them has different physico-chemical properties. In this case, they are so-called epitopes on the cell surface.
2. Major Histocompatibility Complex (MHC) – tissue compatibility system; it is a group of proteins responsible for presenting an antigen to other components of the immune system. In mice 6 MHC proteins can be distinguished.
3. Immune response – 1 if the drug responds positively; 0 if there is no answer.

The data we used was made available by Ardigen for the purposes of BioHack 2019 and are free to use and download.

	Peptide	MHC	Immunogenicity
0	AAALSPMEI	H2-Db	0.0
1	AAASVVGAPV	H2-Db	0.0
2	AAEEFAFL	H2-Kb	0.0
3	AAFNLPIEL	H2-Kb	0.0
4	AAFTFTKI	H2-Kb	1.0
5	AAFTFTKV	H2-Kb	1.0
6	AAFTNLLAM	H2-Db	0.0
7	AAGINVGPI	H2-Db	1.0
8	AAHEFGHAL	H2-Db	1.0
9	AAIESLREM	H2-Kb	0.0
10	AAIFSYLAAL	H2-Kb	0.0
11	AAIFSYLAALI	H2-Kb	0.0
12	AAINNRICV	H2-Db	0.0
13	AAIPNRTFA	H2-Db	0.0
14	AAIRGNDVI	H2-Db	0.0

Fig. 3. Part of data frame. Index is added as a first column to original 3 column data (Source: <http://biohack.com.pl/biohack-ii-exemplary-tasks/>)

### 2.1. Data preparation

Data preparation was divided into three stages as shown below (Fig. 4). One Hot Encoding have been used because the input data is text. The table has also been extended with additional physicochemical data of amino acids present in the peptide structure.

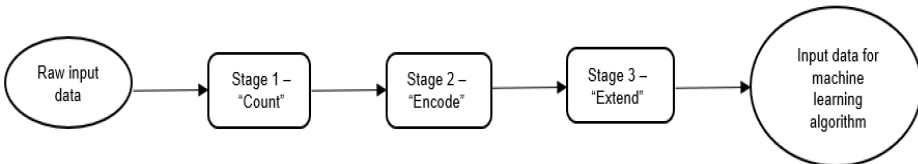


Fig. 4. Data preparation diagram (Source: own study)

#### 2.1.1. Stage 1 – „Count”

At this stage, the first column (peptide) was processed. The data in this column consists of a string of amino acids (length 8-11) – a string of the following characters: 'A', 'C', 'D', 'E', 'F', 'G', 'H', 'I', 'K', 'L', 'M', 'N', 'P', 'Q', 'R', 'S', 'T', 'V', 'W', 'Y'. Each of these letters stand for a specific amino acid. The algorithm counts how many times a given character appears in each string and adds a column with the appropriate name.

After this stage, the data contain additional columns named: 'A', 'C', 'D', 'E', 'F', 'G', 'H', 'I', 'K', 'L', 'M', 'N', 'P', 'Q', 'R', 'S', 'T', 'V', 'W', 'Y', which contains the number of occurrences of a given character in each string.

The Fig. 5 shows only some of the new columns, for the first row we can see that 'A' – occurs 3 times, 'C' does not occur. The algorithm finishes counting given row on 'Y' column.

	Peptide	MHC	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
0	AAALSPMEI	H2-Db	3	0	0	1	0	0	0	1	0	1	1	0	1	0	0	1	0	0	0	0
1	AAASVVGAPV	H2-Db	4	0	0	0	0	1	0	0	0	0	0	0	1	0	0	1	0	3	0	0
2	AAEEFAFL	H2-Kb	3	0	0	2	2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
3	AAFNLPIIEL	H2-Kb	2	0	0	1	1	0	0	2	0	2	0	1	1	0	0	0	0	0	0	0
4	AAFTFTKI	H2-Kb	2	0	0	0	2	0	0	1	1	0	0	0	0	0	0	0	2	0	0	0

Fig. 5. Part of data after stage 1 (Source: own study)

### 2.1.2. Stage 2 – “Encode”

At this stage One Hot Encoding of MHC column occurs. In the model there are 6 MHC types that can be distinguished: H2-Dd, H2-Db, H2-Kd, H2-Kb, H2-Kk and H2-Ld. After this stage, the following columns are added to the table from stage 1:

- 1 – means that the given MHC type appears in this line.
- 0 – means that the given MHC type does not appear in this line.

Each row can have only one MHC type assigned. The so-called real hot encoding was used because first unnecessary column needs to be removed. Sample row is shown in Fig. 6.

	H2-Dd	H2-Kb	H2-Kd	H2-Kk	H2-Ld
0	0	0	0	0	0
1	0	0	0	0	0
2	0	1	0	0	0
3	0	1	0	0	0
4	0	1	0	0	0
...	...	...	...	...	...
7351	0	1	0	0	0
7352	0	0	1	0	0
7353	0	0	1	0	0
7354	0	1	0	0	0
7355	0	0	1	0	0

7356 rows × 5 columns

Fig. 6. Visualization of encoding MHC in data (Source: own study)

### 2.1.3. Stage 3 - "Extend"

At this stage, the table is expanded with vectors containing the amino acid properties that are in the given string. If any amino acid is not present in the sequence, 0 is entered in place of its properties.

The amino acid properties are shown in Table 1.

Table 1. Amino acid properties. HP is abbreviation for hydrophaty;  $K_d$  is a dissociation constant; pKa is acid dissociation constant [4]

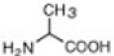
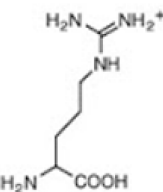
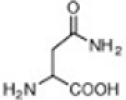
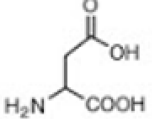
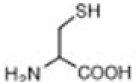
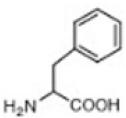
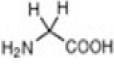
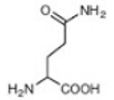
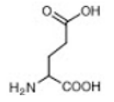
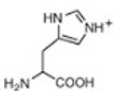
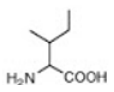
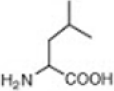
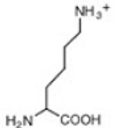
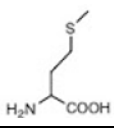
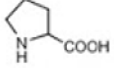
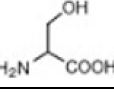
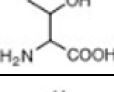
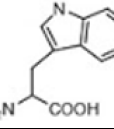
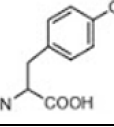
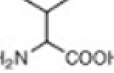
Chemical structure	Amino acid	HP	$K_d$	Mass	HP index	Charge	pKa NH <sub>2</sub>	pKa COOH
	Alanine	1	-7	89.094	1.8	0	9.87	2.35
	Arginine	-1	-4	174.203	-4.5	1	9.09	2.18
	Asparagine	-1	-5	132.119	-3.5	0	8.8	2.02
	Aspartic Acid	-1	-2	133.104	-3.5	-1	9.6	1.88
	Cysteine	0	-5	121.154	2.5	0	10.78	1.71
	Phenylalanine	1	-4	165.192	2.8	0	9.24	2.58
	Glycine	1	-4	75.067	-0.4	0	9.6	2.34
	Glutamine	-1	-5	146.146	-3.5	0	9.13	2.17
	Glutamic Acid	-1	-5	147.131	-3.5	-1	9.67	2.19

table 1. cont.

	Histidine	0	-3	155.156	-3.5	0	8.97	1.78
	Isoleucine	1	-4	131.175	4.5	0	9.76	2.32
	Leucine	1	-5	131.175	3.8	0	9.6	2.36
	Lysine	1	-3	146.189	-3.9	1	10.28	8.9
	Methionine	0	-4	149.208	1.9	0	9.21	2.28
	Proline	1	-6	115.132	-1.6	0	10.6	1.99
	Serine	1	-5	105.093	-0.8	0	9.15	2.21
	Threonine	1	-6	119.119	-0.7	0	9.12	2.15
	Tryptophan	1	-3	204.228	-0.9	0	9.39	2.38
	Tyrosine	1	-4	181.191	-1.3	0	9.11	2.2
	Valine	1	-5	117.148	4.2	0	9.72	2.29

After this stage, the table prepared from our data consists of 167 columns and 7356 rows.

## 2.2. Machine learning model

### 2.2.1. Input data

After the stages 1-3 the data are prepared to be used in a machine learning model. To do so the data was divided into test and training set. Test data accounts for 5% of all data.

### 2.2.2. Model description

As a predictor model random forest was used. As criteria both entropy and Gini index were used. As an evaluation of a model, confusion matrix was created, and ROC curve and AUC were tested. A grid search (process of scanning the data to configure optimal parameters for a given model) from `sklearn.model_selection` library was performed. Afterwards the values determined by grid search were put into a model as shown below.

- `classifier = RandomForestClassifier(n_estimators = 416, criterion = 'gini', random_state = 0)`
- `classifier = RandomForestClassifier(n_estimators = 416, criterion = 'entropy', random_state = 0)`

## 3. RESULTS

### 3.1. Grid search results

The results of the grid search for both criteria are:

- for entropy criterion:  
Best Accuracy: 81.19%,  
Best Parameters: `{'max_depth': 17, 'max_features': 'sqrt', 'n_estimators': 416}`,
- for gini criterion:  
Best Accuracy: 81.21%,  
Best Parameters: `{'max_depth': 17, 'max_features': 'sqrt', 'n_estimators': 416}`.

### 3.2. Random Forest

The results obtained for the entropy (Fig. 8) and Gini index (Fig. 9) criteria are presented below as the confusion matrix. The matrix was analyzed according to the pattern shown on Fig. 7. True positives (TP) are true values correctly classified. False positives (FP) are negative values incorrectly classified as true. False negatives (FN) are true values incorrectly classified as negative. True negatives (TN) are negative values correctly classified. Type I error is the rejection of a true null hypothesis, while a type II error is the non-rejection of a false null hypothesis.



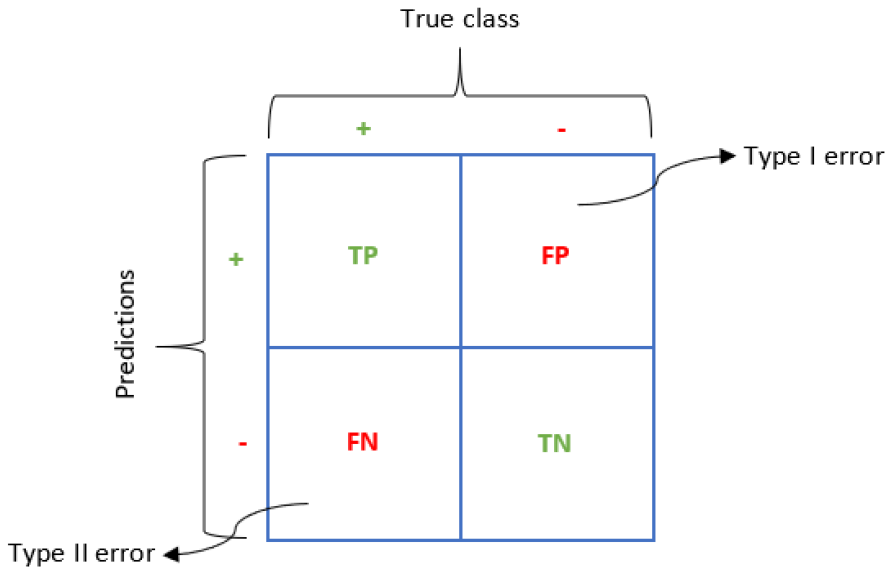


Fig. 7. Confusion matrix pattern (Source: own study based on [https://en.wikipedia.org/wiki/Confusion\\_matrix](https://en.wikipedia.org/wiki/Confusion_matrix))

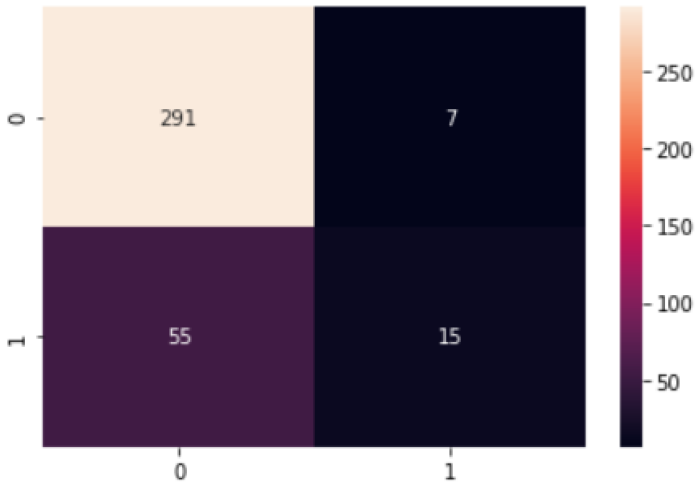


Fig. 8. Confusion matrix for entropy criterion (Source: own study)

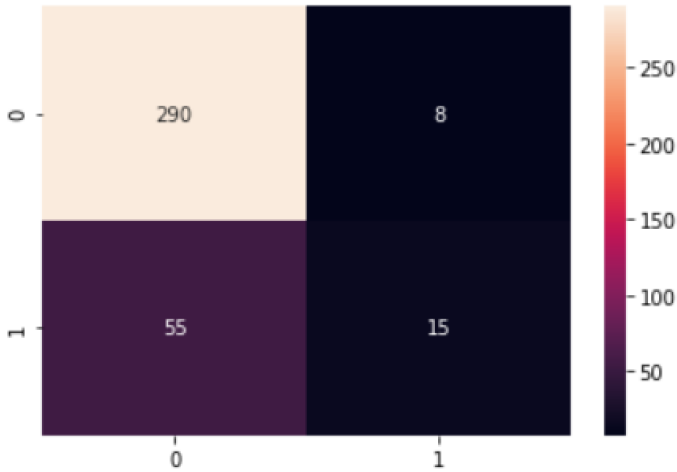


Fig. 9. Confusion matrix for Gini index criterion (Source: own study)

From confusion matrix one can calculate various scores that describe given data. The accuracy can be calculated according to formula (1). For proposed model accuracy is: 84% for entropy criterion, 83% for Gini index criterion.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Next rate one can calculate is ‘Precision’, also called measure of “Correct Positives”. Precision answers the question: “When model predicted true, how often was it right?”. In this case model predicted that 291 combinations (TP and FP) give answer “0”, of which 288 (TP) was predicted right. Generally, the dataset had 345 combinations that give “0” answer. Precision is calculated as shown in formula (2). For this dataset precision is: 99% for both, entropy, and Gini index criterion.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Another rate one can calculate is ‘Recall’, also called “sensitivity of True Positive rate”. Recall answers the question: “when the class was actually true, how often did classifier get it right?”. Recall can be calculated as shown in formula (3). For this dataset we had 345 negative immune response, but the classifier detected only 288 of them, so recall rate is 83% for entropy criterion and 82% for Gini index criterion.

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

F1 score is an overall measure of model’s accuracy and combines two already calculated scores – recall and precision. F1 score is harmonic mean of precision and recall. This score differs from accuracy and is often used for unbalanced datasets, where true negative rate is high accuracy could be misleading. F1 score can be calculated as shown in formula (4). Even though tested dataset is slightly unbalanced, true negative rate is not significant and F1 score is 91% (entropy) and 90% (Gini index). All rates for both criterions are shown in table below (Table 2).

$$F1\ score = \frac{2(PRECISION*RECALL)}{(PRECISION+RECALL)} \quad (4)$$

Table 2. Comparison of results for both criteria (Source: own study)

	Accuracy	Precision	Recall	F1 score
Entropy	<b>83.15%</b>	<b>97.65%</b>	<b>84.10%</b>	<b>90.37%</b>
Gini Criterion	82.88%	97.32%	84.06%	90.20%

### 3.3. Roc curve

ROC Curve is a metric used to assess the model ability to distinguish between binary (0 or 1) classes. It is created by plotting true positive rate against false positive rate at various threshold settings. As the model performance improves it becomes skewed towards upper left corner. ROC Curve for this experiment is shown below. For both criterion AUC is similar and both ROC Curves are above random predictor line (diagonal line) (Fig. 10).

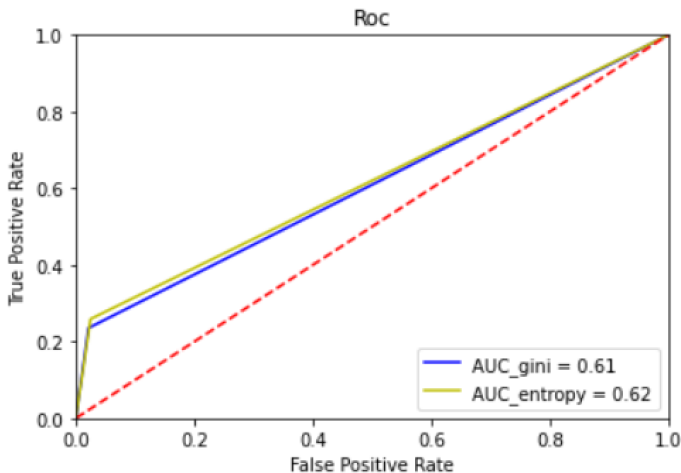


Fig. 10. ROC Curve for Random Forest Classifier for both criteria (Source: own study)

## 4. DISCUSSION

There are few methods that allow prediction of binding affinity between peptide (as an epitope) and MHC, so they form p-MHC complex. They all use slightly different approaches to obtain the best possible results. Developing new strategies to fight against various infections and illnesses, even cancer is high priority for scientists around the world. This fact brings hope for the patients that are currently waiting for treatment. Currently it seems like no single method can be used to predict the p-MHC binding formation with 100% probability.

The results obtained in the model are not fully satisfactory, although they may suggest the direction of further research, development, and improvement of the method. Undoubtedly, these results are the basis for the claim that the use of ML in biological and clinical issues is justified.

Devette et al. also worked on tumor response prediction in mice. They combined laboratory experiments as well as an MHC prediction tool: NetH2pan [4]. It is based on an artificial neural network (ANN). Based on the chemical-biological structure of the particles, this tool is able to conclude with a certain degree of probability the usefulness of therapy. However, as the authors themselves point out, further research and attempts to improve the platform are still necessary in order to identify the peptides unique to cancer cells as accurately as possible, which will allow for effective treatment.

Cheng et al. pay attention to genomic databases, both from healthy people and from people suffering from cancer, are constantly evolving and provide more and more data [2]. It is impossible to analyze each data set to spot differences that can have a key impact on disease development. Thanks to the development of machine learning algorithms and their use in medicine, it is increasingly possible to find important information about the sequence of nucleotides or amino acids in a peptide and their impact on changes taking place in cells. This information can then be used to devise effective therapies, including personalized immunotherapy.

Nussinov et al. indicate a large group of bioinformatics tools used in precision oncology that are already available. They state, however, that these tools are useful but insufficient – thanks to their use, it is more and more often possible to determine the cause of the disease (mutations responsible for its development), but this information still does not allow for a sufficient increase in the effectiveness of the therapy [9]. ML methods are extremely useful as they allow scientists to see patterns that are invisible to the naked eye. Unfortunately, biology and medicine are governed by laws that often cannot be described only with raw data. The current state of knowledge allows the use of ML algorithms to improve the theories and diagnostics, but often, despite the high percentage of probability, the solution proposed in this way may turn out to be ineffective.

The analyzed issue is extremely complex, so it is worth continuing research on this topic. Additionally, successes in this field may contribute to the progress in the treatment of cancer. Further development paths include, for example, testing other prediction models and focusing on other properties of amino acids and proteins included in the analyzed complexes.

## BIBLIOGRAPHY

- [1] Achan P., Warriar A.G., Chitturi B., 2011. Biological Data Handling Methods. In: Proceedings of the International Conference on Bioinformatics & Computational Biology (BIOCOMP).
- [2] Cheng F., Liang H., Butte A.J., Eng C., Nussinov R., 2019. Personal mutanomes meet modern oncology drug discovery and precision health. *Pharmacol Rev.*
- [3] Creighton T.E., 1993. Proteins: structures and molecular properties. Macmillan;
- [4] DeVette C.I., Andreatta M., Bardet W., Cate S.J., Jurtz V.I., Jackson K.W., et al., 2018. NetH2pan: A computational tool to guide MHC peptide prediction on murine tumors. *Cancer Immunol Res.*
- [5] Hosseinzadeh A., Edalatpanah S., 2017. Classification Techniques in Data Mining: Classical and Fuzzy Classifiers. In: Emerging Research on Applied Fuzzy Sets and Intuitionistic Fuzzy Matrices.
- [6] Kamiński B., Jakubczyk M., Szufel P., 2018. A framework for sensitivity analysis of decision trees. *Cent Eur J Oper Res.*

- [7] Kotsiantis S.B., 2013. Decision trees: a recent overview. *Artif Intell Rev.*
- [8] Luna J.M., Gennatas E.D., Ungar L.H., Eaton E., Diffenderfer E.S., Jensen S.T., et al., 2019. Building more accurate decision trees with the additive tree. *Proc Natl Acad Sci.*
- [9] Nussinov R., Jang H., Tsai C-J., Cheng F., 2019. Precision medicine and driver mutations: Computational methods, functional assays and conformational principles for interpreting cancer drivers. *PLoS Comput Biol.*
- [10] Painsky A., Rosset S., 2016. Cross-validated variable selection in tree-based methods improves predictive performance. *IEEE Trans Pattern Anal Mach Intell.*
- [11] Patel H.H., Prajapati P., 2018. Study and analysis of decision tree based classification algorithms. *Int J Comput Sci Eng.*
- [12] Topirceanu A., Grosseck G., 2017. Decision tree learning used for the classification of student archetypes in online courses. *Procedia Comput Sci.*
- [13] Zeng X., Yuan S., Li Y., Zou Q., 2014. Decision tree classification model for popularity forecast of Chinese colleges. *J Appl Math.*

## PRZEWIDYWANIE IMMUNOGENNOŚCI U MYSZY PRZY UŻYCIU KLASYFIKATORA *RANDOM FOREST*

### Streszczenie

Dane biomedyczne są trudne do interpretacji ze względu na ich dużą ilość. Jednym z rozwiązań radzenia sobie z tym problemem jest wykorzystanie uczenia maszynowego. Techniki te umożliwiają wychwycenie wcześniej niezauważonych zależności. W artykule przedstawiono wykorzystanie klasyfikatora *Random Forest* z kryterium entropii i indeksem Gini na danych dotyczących immunogenności. Dane wejściowe składają się z 3 kolumn: epitop (peptyd o długości 8-11 aminokwasów), główny kompleks zgodności tkankowej (MHC) i odpowiedź immunologiczna. Zaprezentowany model przewiduje odpowiedź immunologiczną na podstawie kompleksu epitop-MHC. Uzyskane wyniki osiągnęły dokładność na poziomie 84% (entropia) i 83% (indeks Gini). Wyniki nie są w pełni satysfakcjonujące, ale stanowią dobry początek dla bardziej złożonych eksperymentów i wyznacznik do dalszych badań.

Słowa kluczowe: Random Forest Classifier, immunogenność, uczenie maszynowe, entropia, Gini index