



Predictive Technologies and Biomedical Semantics: A Study of Endocytic Trafficking

Radmila Juric¹^a, Elisabetta Ronchieri²^b, Gordana Blagojevic Zagorac³^c, Hana Mahmutefendic³
and Pero Lucin³

¹University of South Eastern Norway, Kongsberg, Norway

²CNAF/INFN, Bologna, Italy

³University of Rijeka, Faculty of Medicine, Rijeka, Croatia

Keywords: Training Data Set, Machine Learning, Endocytic Pathways.

Abstract: Predictive technologies with increased uptake of machine learning algorithms have changed the landscape of computational models across problem domains and research disciplines. With the abundance of data available for computations, we started looking at the efficiency of predictive inference as the answer to many problems we wish to address using computational power. However, the real picture of the effectiveness and suitability of predictive and learning technologies in particular is far from promising. This study addresses these concerns and illustrates them through biomedical experiments which evaluate Tf/TfR endosomal recycling as a part of cellular processes by which cells internalise substances from their environment. The outcome of the study is interesting. The observed data play an important role in answering biomedical research questions because it was feasible to perform ML classifications and feature selection using the semantic stored in the observed data set. However, the process of preparing the data set for ML classifications proved the opposite. Precise algorithmic predictions, which are ultimate goals when using learning technologies, are not the only criteria which measure the success of predictive inference. It is the semantic of the observed data set, which should become a training data set for ML, which becomes a weak link in the process. The recognised practices from data science do not secure any safety of preserving important semantics of the observed data set and experiments. They could be distorted and misinterpreted and might not contribute towards correct inference. The study can be seen as an illustration of hidden problems in using predictive technologies in biomedicine and is applicable to both: computer and biomedical scientists.


1 INTRODUCTION


This study explores the benefits and pitfalls of various types of data pre-processing, carried out under the umbrella of data science. The focus is on the role of data preparation for running machine learning (ML) algorithms and its role in assessing the quality of precision of ML classifiers, which has an impact on predictive inference. The case study is in the biomedical domain. It explores diverse endocytic routes of endosomal cargo molecules and recycling (Blagojevic et al., 2017), (Karleusa et al. 2018), (Mahmutefendic, et al., 2017). It would be beneficial to discover new knowledge on endocytic trafficking

by using predictive technologies and ML algorithms. Data science practices are needed for preparing biomedical data for at least ML classification, and running a selection of classifiers in order to find out if they biomedical data sets can be semantically labelled to fit supervised ML.

However popular the Data Science discipline is, we must not forget a few important concerns. The most obvious is the lack of the definition and an academic consensus on what exactly Data Science would mean and what it could do for biomedical science. Computations created under the umbrella of data science started diverging from main principles of computer science, built over the last 70 years.

^a <https://orcid.org/0000-0002-0441-0694>

^b <https://orcid.org/0000-0002-7341-6491>

^c <https://orcid.org/0000-0003-1249-3802>

Problems are numerous and could be collated into:

- Lacking of formalism in terms of defining which type of computational models data science generates and why;
- Shortcomings of practices of preparing huge data sets for training and testing, which very often legitimately distort the semantic of them (Danilchanka and Juric, 2020), (Juric et al., 2020);
- Looking at missing data values as places to be either eliminated or filled with “something”, without semantic justification for such changes;
- Assuming that noisy data in training data sets have no semantic and we wish to remove them;
- Using software tools for data science operations upon our data, without telling us exactly how they deal with the semantic of data. (Juric, 2018) (Ronchieri et al., 2019),
- Lacking an agreement, amongst computer scientists, on the role of predictive inference, after decades of successful applications of logic reasoning and inference (Newgard, 2015).

This study of endosomal trafficking detected all the problems itemised in the bullets above. In some of the problem domains, they might not be seen as serious concerns. However, in the field of biomedical science, when we strive for new discoveries, they should be taken seriously. Therefore this research, which initially focused on the problem of discovering more knowledge on endosomal trafficking through learning and predictive technologies, diverged into something different. We started questioning

- a) the process of preparing biomedical data sets from endocytic as required by data science,
- b) the lack of semantics in training data sets because of missing data values and
- c) the assumption that ML would help to find out what is missing in the puzzle of endosomal recycling.

The purpose of this paper is twofold.

We would like to draw attention of computer scientists and statisticians to the fact that the current climate of using various statistic inference upon learning data sets, without understanding the impact of the changes in the semantic of the data, is not advisable. Whenever we prepare data sets for running even a simple linear regression or classification, we have to understand what happens to the semantic of the data prepared for them.

We would also like to draw attention of biomedical scientists to the fact that the computational power is hidden in the data generated through biomedical experiments. The power is in the semantic of data observed in and recorded from these

experiments. We have to pay attention to computations before we start collecting data and formatting data sets.

In this study, in spite of our concerns, we have managed to create a set of quite reliable ML classifiers for endosomal trafficking. We are not convinced that the technology is a definitive answer to getting trustworthy insights into collected data and answering research questions. Therefore, the ultimate goal of this study is to initiate *a debate* across disciplines for assessing the future of predictive inference, and its reliability in biomedical science.

The paper is organised as follows. In section 2 we give the scenario of endosomal trafficking and define a hypothesis which would be of interest to biomedical scientists. In section 3 we analyse the semantic of the observed data set outline observations which might affect the definition of ML classifiers. In section 4 we illustrate the way data has been prepared for running ML classifiers and focus on the way the content of the data set changes in order to fit essential requirement of any ML classifier: definitions of features and semantics of data labelling. Section 5 we define ML classifiers and debate the problem of missing data values in the training data set. In section 6 we illustrate results of running a two set of classifiers, with and without missing data values in the training data set. Sections 7 debates the results and outlines what the future of ML algorithms in biomedical science research might hold.

2 ENDOCYTOSIS AND Tf/TfR ENDOCYTIC ROUTES

Endocytosis is an essential cellular process when cells internalize substances from their environment. There are two main types of endocytosis: clathrin dependent endocytosis (CDE) and clathrin independent endocytosis (CIE). The best model for studying CDE is the model of Transferrin/Transferrin receptor (Tf/TfR) endocytosis.

Transferrin (Tf) is a protein produced by liver that has high affinity for binding the iron and as such it has the most important role in regulating iron metabolism. Upon binding of Tf to its receptor (TfR) on the cell surface, the Tf/TfR complex is rapidly endocytosed by CDE. Following endocytosis Tf/TfR complex is transported into early endosomes (EE) where acid pH affects the release of iron. Tf/TfR complex can then either be recycled back to the cell surface (*fast recycling*) or directed to the jukstanuclear recycling compartment (JNRC). From

the JNRC, Tf/TfR complex is recycled back to the cell surface (*slow recycling*). In contrast to TfR, recycled Tf can not be detected on the cell surface because it is released to the medium.

The difference between endocytosis and internalization is important. Endocytosis means *entering into the cell* and internalization means *disappearance from the cell surface*. The latter is a net result of endocytosis and recycling (internalization = endocytosis - recycling). Consequently Tf can be used to track the endocytosis and TfR to track the internalization.

It is known that CDE is very fast and thus these biomedical experiments must have very short time intervals for collecting/observing data in each experiment. It is technically impossible to read results below 2 minutes of a time stamp, but the first 10 minutes of the experiment can be monitored carefully. TfR recycled from the EE can be early detected on the cell surface, even after the first 3 minutes from the beginning of the experiment. TfR, recycled from the JNRC is visible after 8-10 minutes from the beginning of the experiment.

It is important to note that there is a possibility of pre-EE (rapid) recycling, which occurs before EE (during the first 3 minutes). After 20 minutes TfR reaches its steady state (homeostasis) (Mahmutefendic, et al., 2018) and thus there is no need collect results of these experiments very frequently. Short time stamps are important at the beginning, but not after 20 minutes. Furthermore, considering that 2 min time stamp is expensive it is not necessary to perform them after 20 min. This means that during the first 20 minutes we can detect fast and slow recycling, kinetics of the endocytosis and after 20 minutes TfR reaches its homeostasis.

Each experiment uses fluorescently labeled antibodies for the detection of Tf and TfR. Antibodies can bind only Tf and TfR that are on the cell surface, regardless of their status. We do not know from the number of molecules on the cell surface if (a) they have not been endocytosed yet or (b) they have been endocytosed but they are recycled to the cell surface.

If antibody binds Tf or TfR, there is fluorescent signal, which is detected by flow cytometer. When cells are analyzed by flow cytometer, the number that represents *mean fluorescence intensity (MFI)* is read. MFI represents average fluorescent signal given by a single cell. The higher MFI value, the more molecules are present on the cell surface. Their presence is an indication of either slow endocytosis or fast recycling. The highest fluorescent signal is read at the beginning of each experiment (minute ZERO). After that, the signal slowly decreases due

to endocytosis. All the numbers given by flow cytometer are percentages calculates as MFI values. The example is given in Table 1.

Table 1: Calculating percentages from MFI Values.

	0	2	4	6	8
MFI VAL.	354	298	250	178	100
PERCENT.	100	84	70,6	50,3	28,2

The observed data set contains a set of numeric values, which show the presence of molecules on the cell surface in the relevant time stamp. Table 1 says that, after the beginning of the experiment (after minute 0), in which 354, i.e. 100% of molecules were present on the cell surface, in minute 8, there will be only 28% of them available on the cell surface. As mentioned earlier, this percentage does not indicate exactly how many of these molecules are present because of either slow endocytosis or fast recycling.

Considering results from the literature and earlier publications (Blagojevic et al., 2017), we would be interested in finding out if pre-EE does exists as a part of Tf/TfR endocytic routes, because it has not been proved in the literature yet. The question is could predictive and learning technologies help?

3 THE SEMANTIC OF THE OBSERVED DATA SET

The observed data set consisted of the rows of data, stored in a spreadsheet, where rows contained numeric values from each experiment (as shown in the second row of Table 1), and columns correspond to a time stamp at which the data was collected. However, we stumbled upon first problem immediately. Description of each experiment, in terms of the names of cells and molecules involved, data related either endocytosis or internalization, and data which describe conditions in which each experiment was conducted, are not part of the data set. This means that significant and semantically rich data form these experiments are of a descriptive nature, i.e. they are NOT numeric values, and as such, they were not entered into data set. The other problems are:

Labelling data set in classifications is extremely important and we had to ask if time stamps, i.e. the exact MINUTES in which the data is collected, would be suitable for defining features and deciding about labelling for ML classifiers?

Missing values in the data set is a result of not collecting data in all available minutes between 1 and

180. This might not be an obstacle, because NO DATA is semantic itself, according to the argument in the previous section. However, if a software automated tool is used for data preparation and processing, how can we be sure if the results of running ML algorithms upon data sets with more than 50% of missing data values, are good or bad. How could we know that if a software tool is in charge?

Removing columns or rows with missing data might be a double-edged sword: we may manually add semantic by removing “column headings” and converting them into the data values (Danilchanka and Juric, 2019). We may delete rows and columns with excessive amounts of missing information, or merge them following any justifiable prediction theory rules, but how do we know how much semantics we might lose or gain?

Noisy labels, Synonyms, Duplications, and many similar terms, are usually seen as obstacles in creating quality training data sets. However, they are very important semantically, i.e. why do we assume that there is no semantics in noisy labels? Is it safe to leave the interpretation of the meaning of noisy labels in the observed data to automated tools?

Semantic significance of a training data set may be confined to the data values stored in particular column(s) and therefore the column’s definition and its data values are essential in the precision of classifiers/predictions. Do we have to single out semantically significant columns of this training data set, or treat all columns equally when defining our own ML classifier suitable in this problem domain?

Classifier’s features are defined through the semantic stored in columns and the combination of columns of tabular formats, which could be chosen as *features*. However, how do we balance this? Which column will become a feature? What shall we do with the semantic of non-numerical data in terms of feature selection if we enter them into the training data set?

The above six observations appear after the manual analysis of the observed data set from Tf/TR endosomal experiments. This means that

- (a) the observed data set is not ready for any type of ML processing and
- (b) we need to find out which process we should use for preparing the data set for ML classifiers.

In order to answer as many questions as possible, regarding these observations, and address (a) and (b) above, we had to involve human intervention in restructuring the data set and focus on the following three principles:

- i. which semantic is relevant for preparing the data set further for learning technologies? Are MFI percentages sufficient?

- ii. would new columns in the data set, which store the semantic of conditions of each experiment, be suitable for ML classification?
- iii. how much could we trust our choice of features? Is our selection of features (time stamps) the adequate for data labelling/ defining classifiers?

An algorithm for data labelling is not difficult to create because of the explanations given in the previous section. Rich semantics are available within/around the experiments, even if it is not directly available in the observed data set. Human intervention in the process of restructuring the data set would take charge of that and it might enable the definition of the labelling mechanism.

4 PREPARING THE DATA SET FOR ML CLASSIFIERS

The observed data set was manually restructured according to the explanations from Section 3. No software tool could prepare the data set according to our analysis and add more semantics to it. The data set was restructured manually by

- Keeping the same time-stamps, which defined columns with numeric (MFI) values,
- Adding new columns with literal data values, which described additional semantics from the experiments (not MFI values).

Therefore numeric values were concatenated with strings: columns which explained conditions in which experiments were performed. Table 2 illustrates the additional semantic. There is a set of new columns, coloured as yellow boxes, and a sample of values stored in these columns are coloured as blue boxes.

Table 2: Additional columns in the observed data set.

Cell	Experiment	Condition MFI
Balb 3T3	INTERNALIZATION	around 20-25
Condition MCMV	Condition Temperature	Period
No	37oC	Long
Cell Status	Condition Interferon	
Normal	No	

Tables 3, 4 and 5 show excerpts from the training data set, i.e. only 11 interesting experiments. Each experiment occupies one row.

The top rows coloured in blue, green and amber in Tables 3, 4 and 5, show the exact minute in which data values are collected.

In Table 3, which shows the first 10 minutes of each experiment, the maximum number of reading of MFI values is 5, but in some cases not more than 2. Table 4 shows the same 11 experiments for minutes between 10 and 20. The time stamp changed and frequency of collecting numeric data values is decreasing: it is every 5 instead of 2 minutes. Table 5 shows the last 120 minutes of the same experiments. Readings were not conducted after 60 minutes.

Tables 3,4 and 5 show that we kept numerical values as they were observed in experiments. The only change in the observed data set was added semantics, which describes the experiments.

Adding more columns/semantic for describing experiment means adding strings and not numeric values: the types of cells chosen for experiments, their status, temperature, infection with MVN, presence of interferon and type of experiment (endocytosis or internalization).

It is important to note, that we could not declare the existence of any noisy labels for one obvious reason: all our numeric values were carefully collected / recorded and additional semantic was added manually. Therefore, there was no need for any other aspect of “cleaning” the data set, as a part of data preparation. However, there is only one serious concern in this particular case study: a significant amount of missing data values in the relatively semantically rich training data set.

5 DEFINING ML CLASSIFIERS

A significant number of missing data values in the data set requires looking at this problem differently. We are not confident that any of the existing data science practices of imputations would work here. The idea of imputations is both seductive and dangerous (Dempster and Rubin, 1977).

Firstly, we have to define a set of ML classifiers and we do not wish to replace missing data values, using any of the recommended data science practices. The reasons are obvious. These values are either not feasible to obtain (in the first 10 minutes) or irrelevant (after 20 minutes). Any other reasoning for imputing data into missing values would be inappropriate (Juric et al., 2020). How could we assume that the simulations of missing data values in the first 2-5 minutes of each experiment would be appropriate for answering the questions from the hypothesis?

Table 3: Excerpts from the observed numeric data (1).

0	1	2	3	4	5	6	7.5	8	10
100		24.8			18.9				11.5
100					48.7				37.4
100									
100					28.8				23.8
100	82.4		70.8			60.6		52.4	44.5
100	77.3		65.1			55.4		48.6	43.3
100	68.2		45.9			47.6		34.1	32.3
100	78.7				49.3				29.1
100					38.3				26
100					49.9				51.6
100	47.4		34.4						65

Table 4: Excerpts from the observed numeric data (2).

20	25	30	35	40	45	50	60
14.1		11.9					3.71
31.4	29.9	27.9	27.2	26			
	35			18.1			13.5
		13					5.28
29.8		14.3					7.8
38.1		22.1					
25.2		15					
20.4		18.1					
		7.4					4.53
41.9	43.6	42.2	39.41	38			
66.8							

Table 5: Excerpts from the observed numeric data (3).

35	40	45	50	60	75	90	105	120	180
				3.71					
27.2	26								
	18.1			13.5					
				5.28					
				7.8					
				4.53					
39.4	38								

However, simulating missing data values (after 20 minutes in each experiment) might be acceptable, but these data values are almost irrelevant because of homeostasis.

Secondly, we know that we have to use software tools for running ML algorithms, and therefore we have an opportunity to perform imputations for missing data values according to the options or functionalities tools offer. Treating missing values through software tools might give us at least an indication whether we are able to define ML classifier by doing something which is not semantically justifiable. We may be able to test these tools to find out if the imputation they performed would give good precisions for a set of ML classifier(s).

Consequently, it would be prudent to define and run ML classifiers twice: first time WITH missing data values and second time with imputations for replacing missing values with numbers generated by the chosen tool.

In this study we used RStudio (version 1.2.1335) for data visualisation and potential “cleaning” and the Weka ML framework (version 3.8.3) for defining and running ML classifiers. We decided to run ML experiments with relatively small data set in order to secure a variety of semantic for successful classification. Considering that we had 34 columns in our data set, then having 147 instances over 34 features, would not require the use of any other ML frameworks, such as scikit-learn/R combination. RStudio/Weka leaves us with a freedom to control the definition of classifiers and accommodate our involvement in these ML experiments.

Many of the classification and regression algorithms available in the Weka framework were tried and examined, but we show results given by Random Forest, REP Tree, JRIP, Bagging, Classification via Regression, Random Committee and Random Sub Space. They presented a kappa statistic greater than 0.81, which is the minimum value for which the agreement result can be considered good for further investigation.

The definition of the classifier includes a feature selection and an algorithm for labelling the data set, as a part of supervised ML techniques.

The feature selection was not a problem for one important reason. If we wished to run ML classifiers with a data set which has a significant amount of missing data values, than it would be unreasonable to exclude some of the columns in the definition of the classifier, due to nature of these experiments. Also, we added more semantic assuming that it is relevant for the definition and precision of classifiers. Practically all columns from our data set must be included in the feature selection. They are bulleted below. The features correspond to top rows of Tables 2,3 and 4 and yellow boxes in Table 1.

- minutes when reading was obtained 1, 2, 3, 4, 5, 6, 7.5, 8, 10, 12.5, 15, 20, 25, 30, 35, 40, 45, 50, 60, 75, 90, 105, 120, 180;
- cell name (String: list of different cells);
- type of experiment (String, Endocytosis, Internalisation);
- MFI level (String: low/normal/high/any range);
- Infection MCMV (String Yes / No)
- Cell temperature (String, number for Celsius)
- Interferon presence ((String Yes / No);
- Period of experiments (String: Short, Long);
- Cell status (String; normal).

Data Labelling has been done using a specific algorithm which correspond to the hypothesis: are we able to label each experiment into pre-EE, early-EE and Late-EE?

We should be able to find minutes in which we obtained either Min or Max MFI values of the Tf/TfR complex on the cell surface. This would directly imply that we should have an extra column in the data set, for each experiment, which specifies a minute in which these min and max values are obtained. The algorithm devised in this study might not be the best possible choice of reasoning needed for the labelling and it could change if the semantic described in Section 3 changes. What is important here is to see if we can classify the labels from the data set which were not aimed at being used for learning technologies and leave the algorithm from Figure 1 to serve only for illustration purpose.

```

For each cell and INTERNALISATION
If MAX Value is <=3 min label is pre-EE
If MAX Value is <=5 min label is EE
All others Max Values label is LE

For each cell and ENDOCYTOSIS
If MAX Value is <=3 min label is LE
If MAX Value is <=5 min label is EE
All other Max Values label is pre-EE
    
```

Figure 1: Potential algorithm for data labelling.

6 RUNNING ML EXPERIMENTS

Tables 6 and 7 are results of running a set of classifiers with and without missing data values. In case of imputing data to replace missing values, the tool replaced empty data values with the mean values obtained for the various instances: the mean was calculated by row. They show method, correct classified instances % (CCI%), incorrect classified

instances % (ICI%), Kappa statistic (KS), mean absolute error (MAE), root mean squared error (RMSE), relative absolute error % (RAE%), and root relative squared error % (RRSE%).

Table 6: Running ML Classifiers WITH Missing Data Values.

Method	CCI%	ICI%	KS	MAE	RMSE	RAE%	RRSE%
Random Forest	91.836	8.163	0.870	0.175	0.236	41.175	51.220
REP Tree	91.156	8.843	0.862	0.070	0.214	16.464	46.565
JRIP	98.639	1.360	0.978	0.070	0.101	16.560	21.925
Bagging	95.918	4.081	0.935	0.100	0.179	23.525	38.780
Class. Via Regression	96.598	3.401	0.946	0.138	0.193	32.569	42.020
Random Committee	91.156	8.843	0.860	0.126	0.2140	29.690	46.375
Random Sub Space	93.877	6.122	0.903	0.193	0.244	45.488	52.917
Average	94.169	5.830	0.908	0.125	0.197	29.353	42.829

There are two important outcomes from Tables 6/7. Firstly, the precision of ML classifiers WITH missing data values is better than without them. Missing data values, calculated by the chosen tool were replaced by the tool. Secondly, in spite of having a significant number of missing values, the precision of the algorithms which include them is quite good. We may say unexpectedly good.

Table 7: Running ML Classifiers WITH Imputations for Missing Data Values.

Method	CCI%	ICI%	KS	MAE	RMSE	RAE%	RRSE%
Random Forest	88.435	11.564	0.816	0.184	0.258	43.215	56.062
REP Tree	90.476	9.523	0.850	0.079	0.221	18.561	47.910
JRIP	95.238	4.761	0.924	0.048	0.169	11.432	36.686
Bagging	95.918	4.081	0.935	0.098	0.174	22.978	37.788
Class. Via Regression	94.557	5.442	0.914	0.154	0.221	36.180	47.973
Random Committee	89.115	10.884	0.829	0.139	0.256	32.702	55.627
Random Sub Space	93.197	6.802	0.892	0.208	0.258	48.919	56.081
Average	92.419	7.580	0.805	0.130	0.222	30.569	48.304

7 DISCUSSION AND CONCLUSIONS

This study could start debates on the future role of learning and predictive technologies across the disciplines of computer and biomedical sciences. The message from the study goes to both sides.

Computer scientists should be aware that current data science practices need attention because they may not guarantee the best possible semantic of

training data sets, which may affect ML results. In this study, we avoided all well accepted principles of removing rows/columns and noisy labels and address missing data values in order to prepare a data set. There was no justification for doing opposite. We achieved better precision of ML classifiers by keeping rows/columns with missing data values. We also entered semantic in the data set from biomedical experiments which was initially not considered. They were essential in creating labeling algorithms for supervised ML. Therefore human intervention was essential in preserving the semantic of the data set.

Biomedical scientist should be aware that biomedical experiments must be conducted with the type of data processing in mind. Moving from the statistical predictions towards learning technologies requires different ways of collecting data and possibly would need collaborations with computer scientists in order to evaluate which hypothesis could be feasible to (dis)approve. This takes us directly to the joint definition of the features and algorithms for labelling data set (as in Figure 1) with computer scientists.

This study also proved that we follow i.-iii. from section 3. MFI values could have been sufficient for defining classifiers, but the labelling of data set, which took into account the additional semantic form the experiments, outside the MFI values, was not difficult to define. Current combination of features with the labelling algorithms proved that it is feasible to run ML classification on this data set even with a significant amount of missing data values.

Our decision NOT to run imputations as described in (Acuna and Rodriguez, 2004), (Batista and Monard, 2003), might seem unusual but it is not isolated: there are examples where better precision in classification has been obtained by NOT removing missing data values (Danilchanka and Juric, 2020), (Danilchanka and Juric, 2018). Unfortunately, there are not so many publications which focus on this problem. Publications which are debating probabilistic uncertainty are old and rare (Newgard, 2015) (Craddock, 1986), (Rubin, 1976). Currently, Published papers are mostly concerned with achieving high precision of ML algorithms, without worrying that we might be scarifying the original semantic of training data sets to increase precision.

One of the most important outcome of this study is the fact that relatively good precision of our ML classifiers must encourage us to work further on creating semantically richer training data set and use them for example in unsupervised and deep learning. We can assume that this study would warrantee more research to be done in the field of predictive inference, as long as we can mange the semantic of the training data set.

There are two limitations.

Firstly, we were not able to answer one of the crucial questions from endosomal trafficking: “does pre-EE exists” for many reasons. In order to find out what is happening in the first 2-3 minutes in each experiment, we would need to use results from more experiments (we used only 147 experiments) and try some other ML algorithms. These are very expensive experiments, and thus we might re-think the way they are carried out. Simulating data for replacing values which can not be measured/obtained for the first two minutes, must be debated. Imputation used in the second set of ML experiments did not help to improve the precision. Also, Endocytosis and Internalizations semantically overlap and thus they should be addressed in future work, when defining the additional semantics of the training data set and revisiting the algorithm from Figure 1.

Secondly, we could have analyzed the results of the second set of ML classifiers, which had imputed mean values, calculated per each row. This would mean that we are trying to achieve precision in classification, but we will not know if we are improving the quality of the data set at the same time. Would this help us to find out if pre-EE existed?

Immediate future work should address our first set of limitations. The second set of limitations is a subject of more complicated debate: is predictive inference desirable in biomedical science if we could not guarantee that the semantic of the training data set will not be distorted. For this particular problem of endocytic trafficking, unfortunately the answer might be NO. However, this should not discourage us from searching for or finding more options where both predictive and logic inference cohabit (Basulto et al., 2017). In long term, this could lead towards discovering new insights in biomedical data

REFERENCES

- Acuna, E., Rodriguez, C., 2004. The treatment of missing values and its effect in the classifier accuracy. *Banks D., McMorris F. R., Arabie P., Gaul W. (eds) Classification, Clustering, and Data Mining Applications. Studies in Classification, Data Analysis, and Knowledge Organisation. Springer, 2004*
- Basulto, V. G. Jung, J.C., Schroeder, L., 2017. Probabilistic Description Logics for Subjective Uncertainty, *Journal of Artificial Intelligence Research 58 (2017) 1-66.*
- Batista, G., M., Monard, C., 2003. An Analysis of Four Missing Data Treatment Methods for Supervised Learning, *In Applied Artificial Intelligence, Vol 17, 2003, Issue 5-6 pp 519-533.*
- Blagojević Zagorac, G., Mahmutefendić, H., G., Maćešić, S., Karleuš, L. J., Lučin, P., 2017. Quantitative Analysis of Endocytic Recycling of Membrane Proteins by Monoclonal Antibody-Based Recycling Assays, *In Journal of Cellular Physiology 232(2017), 3; 463-476.*
- Craddock, A. J., Browse, R. A. 1986. Reasoning with Uncertain Knowledge, in *UAI'86, Second Conference on Uncertainty in Artificial Intelligence, pp 57-62*
- Newgard, C. D. Lewis, R. J., 2015. Missing Data: How to Best Account for What Is Not Known, *Clinical Review& Education, JAMAGuide to Statistics and Methods*
- Danilchanka; N., Juric, R., 2020. The Process of Creating a Training Data Set: Lessons Learned from Mechanical Engineering, *in SDPS 2018 Workshop of Accountability of AI Bologna, Italy.*
- Danilchanka; N., Juric, R., 2020. Reliability of Training Data Sets for ML Classifiers: a Lesson Learned from Mechanical Engineering, *in Proceedings of the 53rd HICSS conference, January 2020.*
- Dempster, A. P., Ruibn, D. P. 1997. Incomplete Data in Sample Surveys, *Theory and Bibliography, Vol 2 (ed, W.G. Madow, I. Olkin and D.B. Rubin), 3-10. New York Academic Press.*
- Juric, R., 2018. How BIASED Could AI Be? *In SDPS 2018 Workshop of Accountability of AI Bologna, Italy.*
- Juric, R., Ronchierri, E., Blagojević Zagorac, G., Mahmutefendić, H., Lučin, P. (20,20). Addressing the Semantic of Missing Data Values in Training Data Sets using MVL: A Study of Tf/TfR Endocytic Routes, *under review for the ISMVL 2020 Conference, Japan May 2020.*
- Karleušaa, L J., Mahmutefendić,H., Ilić Tomaš, M., Blagojević Zagorac, G., Lucin, P., 2018. Landmarks of endosomal remodelling in the early phase of cytomegalovirus infection, *in Virology 515 (2018) 108–122*
- Mahmutefendić, H., Blagojević Zagora, G., Grabušić, K., Karleuš, L. J., Maćešić, S, Momburg F., Lučin, P. (2017) Late endosomal recycling of open MHC-I conformers, *in Journal of cellular physiology, 2017 April, 232(4):872-887.*
- Mahmutefendić, H., Blagojević Zagora, G., Maćešić, S, Lučin, P. (2018) Rapid Endosomal Recycling, *Book Chapter, in Open Access Peer Review Chapter, IntechOpen, <https://www.intechopen.com/books/peripheral-membrane-proteins/rapid-endosomal-recycling>*
- Ronchieri, E., Juric, R., Canaparo, M., 2019. Sentiment Analysis for Software Code Assessment. *In proceedings of the 2019 IEEE NPSS Conference, Manchester, UK.*
- Rubin, D.B., 1976. “Inference and Missing Data” *Biometrika, Vol. 63, No. 3 (Dec., 1976), pp. 581-592.*