Reliability of Training Data Sets for ML Classifiers: a Lesson Learned from Mechanical Engineering

Natallia Danilchanka
University of South East Norway
Kongsberg, Norway
213762@usn.no

Radmila Juric
University of South East Norway
Kongsberg, Norway
rju@usn.no

Abstract

The popularity of learning and predictive technologies, across many problem domains, is unprecedented and it is often underpinned with the fact that we efficiently compute with vast amounts of data and data types, and thus should be able to resolve problems, which we could not in the past. This view is particularly common among scientists who believe that the excessive amount of data, we generate in real life, is ideal for performing predictions and training algorithms. However, the truth might be quite different. The paper illustrates the process of preparing a training data set for an ML classifier, which should predict certain conditions in mechanical engineering. It was not the case that it was difficult to define and choose classifiers, in order to secure safe predictions. It was our inability to create a safe, reliable and trustworthy training data set, from scientifically proven experiments, which created the problem. This places serious doubts on the way we use learning and predictive technologies today. It remains debatable what the next step should be. However, if in ML algorithms, and classifiers in particular, the semantic which is built-in data sets, influences classifier's definition, it would be very difficult to evaluate and rely on them, before we understand data semantics fully. In other words, we still do not know how the semantic, sometimes hidden in a data set, can adversely affect algorithms trained by them.

1. Introduction

The proliferation of applications of Machine Learning (ML), across numerous problem domains, has surprised many computer scientists, but also opened the door to exploring learning and predictive technologies for addressing a variety of problems, which use an excessive amount of data. In this jungle of availability and popularity of ML algorithms, which run almost instantaneously, as soon we be obtain data, by anyone who downloaded software suites, which can automatically create and run almost any ML algorithm, we started worrying. Research focuses often on computational models and algorithms, because they process data, deliver functionalities and therefore they are supposed to be scrutinized. In ML researchers measure their success in evaluating results of predictions through the precision of algorithms they define. They often move from one algorithm to another, to find the best possible solution for a given problem. We all know theoretically which ML algorithms are suitable for which problem(s), but in most cases we are able to run one algorithm after another, in a sequence, in a very short period of time using the same data set. Sometimes, it is not even important if we run probabilistic algorithms together with, for example linear classification. We wish to use them all and check their performance and results.

However successful we are in the processing of data through predictive and learning technologies, we are in danger of neglecting the semantic stored in our data, if we do not pay attention to it. Data powers all algorithms and its semantics affects algorithm's performance [1]. If we do not understand data semantic fully or if we misinterpret it for any reason, or if we ignore problems that might be hidden in data semantics, we might get unreliable results without knowing it.

This paper illustrates an example of applying ML algorithms for running predictions in mechanical engineering, which revealed hidden problems when structuring and creating training data sets. We had to question the suitability of ML not only for this problem domain, but also across similar problems in engineering. Being aware of the semantic of data which is needed for a training data set, and focusing more on the data than on choosing the best ML algorithm, revealed how easy it is to forget about data and misuse its semantics.

The paper is primarily written by computer scientists, for practitioners and students, who would like to start using learning and predicative technology on a larger scale. They might be interested in

URI: https://hdl.handle.net/10125/63850
978-0-9981331-3-3

investigating the role data sets may have in the process of defining ML algorithms, ML classifiers and their features. However, it is important to re-iterate that the journey of semantic of data, from its source to a training data set, would determine the success of our algorithmic predictions.

The paper is organized as follows. Section 2 introduces the task from the mechanical learning field, and sets up a scene for collecting data and it's semantic for the purpose of creating a training data set. It will be used in ML classifiers, which should predict slippery driving conditions in rail transport. The section has four subsections, which describe the process of finding data, finding factors which may influence driving conditions in rail transport, defining the structure of the potential training data set from discovered factors, and entering data values for all factors. The problems encountered in this process are highlighted in the same section, but Discussion and Conclusions comment on research discoveries, outline options for resolving detected problems and pave the future work.

2. Structuring a Training Data Set

If we wish to create a data set, which would train an ML algorithm, in order to categorize conditions between rails and tracks as slippery or non-slippery, we would need to find out the definition of slippery and non-slippery conditions, how they are calculated and which factors influence them.

Slippery conditions in traffic are measured through the level of adhesion or friction, which exists between two different surfaces. The term adhesion describes the tendency of dissimilar particles or surfaces to stick one to another. It is calculated as *force* required to separate two surfaces in contact [2]. In engineering, the term is widely used to describe the tangential force induced in the wheel–rail contact. The adhesion is the relation between the longitudinal tangential force and normal force in wheel-rail interaction [3]. However, friction is the force that appear as resistance when one body is sliding/moving on another [4].

The adhesion force coefficient $\mu_{adhesuin}$ is calculated as a ratio between adhesion force and normal force

$$\mu_{adhesion} = F_T / F_N$$

It is related to the friction coefficient as:

$$\mu_{adhesion} \leq \mu_{friction}$$

This relation is important because the adhesion coefficient can hardly be measured directly, but friction coefficient can be measured by special equipment [5].

Therefore the process of creating a training data set must concern the factors which may influence the level of adhesion/friction (a/f) between rail tracks and wheels. It is expected that the training data set will have all the data values, for all possible situations in which any combination of factors determines exactly the level of adhesion and friction. This expectation comes from the data science point of view.

The next 4 subsections explain the process of creating a training data set.

2.1. The Process of Finding Factors Which Influence Adhesions and Frictions

The calculation of the level of a/f, for various conditions, where numerous factors influence them, requires another calculation, for a/f coefficients, which in turn depend on data values we have for all relevant factors. Therefore the problem of predicting slippery conditions between rails and wheels is now moved towards the problem of finding factors and their data values, which affect a/f coefficients.

We had to define our own process of finding relevant factors, which may influence a/f between hard surfaces, because there is no available source of data which lists such factors and helps in calculating a/f coefficients for all possible known factors and their combinations.

Therefore, this paper involves finding all relevant factors and their impact on a/f, through a) peer reviewed literature, available for the last 50 years and b) collected results of described experiments.

The process consisted of four steps:

1) Extracting experiments and factors, from each experiment, which may cause slippery conditions in rail transport
2) Collating experiments/factors/group of factors into a table in order to find potential overlapping or exclusions between the experiments
3) Recording data values for a/f coefficients from each experiments in order to have them ready for our potential training data set
4) Analyzing the collected factors in order to outline potential problems and inconsistencies which could adversely affect our training data set
5) Defining the semantic of collected data in terms of its role and potential in structuring and defining a training data set.

However, the first problem appeared even before reviewing the literature and executing step 1): *There is no consensus on how to measure a/f.*

In rail traffic, the most accurate measurement of adhesion is obtained by using instrumented train that measures adhesion between actual wheels and rail [6]. However, this method is complicated, expensive and not easily controllable. A cheaper solution is a tribometer, a hand-pushed device with steel wheels that

roll over the rail [7]. This device measures adhesion, but unfortunately it is not measured between the train wheel and rail track. It is measured between the device's wheel and rail surface instead. Thus, the obtained data is interpreted in the literature as *maximum available level of adhesion measured,* and it will not correspond to exact coefficient of adhesion between the wheel of the rain and rail [5].

2.2. Creating a Table of Factors (Group of Factors) for Each Experiment

Finding all relevant factors which can cause slippery conditions and calculating the level of a/f is far from being a trivial task. Table 1 lists factors and collection of factors, which appeared as verified factors in experiments extracted from the literature.

Table 1: Factors which influence a/f extracted from the published experiments

Factors and set of factors	ref.
changes of speed for dry, damp, damp leaf and wet rail	[8]
Herzian contact pressure	[9], [10]
humidity level	[11]
level of humidity changes and rail surface temperature (constant, 3 values)	[12]
humidity and the presence of leaf contamination or dry rail head condition	[13]
dry and clean rail surface or dry rail surface with sand applied on it	[14]
wet and clean surface or wet surface with the applied sand	
greasy contaminated surface	
moisture on the rail-head	
light snow or light snow with sand on the rail surface	
wet leaves on the rail	
water lubricated conditions at certain temperature, two values of contact pressure, two types of surface roughness and changing speed	[15]
synthetic ester oil conditions at certain temperature, two values of contact pressure, two types of surface roughness and changing speed	
water lubricated conditions at two values of temperature, the contact pressure, two types of surface roughness and changing speed	
synthetic ester oil conditions at two values of temperature, the contact pressure, two types of surface roughness and changing speed	
water and oil on the rail surface conditions at two types of surface roughness, constant speed and pressure	
different types of lubricants: dry, wet, oil, dry leaves and wet leaves	[16]
dry and dry, sanded conditions at low speed	[17]
wet and wet, sanded conditions at low speed	
dew or foggy conditions at low speed	
sleet and sleet, sanded conditions at low speed	
light snow and light snow, sanded conditions at low sp eed	
wet leaves on the rail condition with speed was marked as low	
contaminated water contamination tested at two values axle-load and different speed values	[18]
machine oil contamination tested at three values of speed and different axle-load	
dry, wet, oil and leaves contaminated conditions	[19]
water lubricated track with the temperature, the surface roughness, three values of wheel load and different speed	[20]
water lubricated track with four values of temperature, the surface roughness, the wheel load and different speed	
water lubricated track with the temperature, four values of surface roughness, the wheel load and different speed	
dry, wet and oil contaminated rail conditions at certain speed, and load/contact pressure	[21]
water and oil(grease) contaminations with two values of axle loads as speed range	[22]
watered rails with temperature, the surface roughness, two speed rates with Hertz pressure variation	[23]
watered rails with temperature, two types of surface roughness, the speed with Hertz pressure variation	

Table 1 illustrates the scale of the problem which appeared while trying to collect data which could possibly be used as a training data set for any ML classifier. The left column of the table contains the factors which are taken into account in each detected experiment. Therefore it is expected that some sources report on one experiment (such as [16,19, 21,22]) and some report on a set of experiments, when trying to calculate the adhesion coefficient.

It is important to note that in all these experiments, when more than one factor was taken into account (as

in [14]) the coefficient of adhesion was calculated by focusing on one or two factors, and sometimes keeping all other factors constant during the experiment. Therefore source [14] really shows a good set of data which is used for the calculation of the adhesion coefficient, but in each of these experiments only a limited number of factors were monitored, and their data values are entered in the calculations for the adhesion level. Unfortunately, at the time writing, there were no available peer reviewed papers which could be juxtaposed to this research and thus Table 1 remains a unique source of information on a/f factors.

2.3. Problems Encountered

The complexity of predicting slippery driving conditions is now becoming obvious. It is NOT that we do not know how to calculate adhesions between two surfaces for known factors, which affect it. The problem is that there are so many known factors which can affect the friction between the rails and wheels, but there is no scientifically proved method which could define a universal way of calculating a/f when all these factors are involved. In summary:

a) There is an evidence that real time measurements of the friction on real tracks and wheels does not necessarily create correct data used in calculations. This signals that i. data set and its values should be created where each of these factors is "measured" according to knowledge available in the published work, or ii. laboratory experiments are used, in which humans control what has been measured and why. Option ii. is favourable in the literature.
b) These numerous factors listed in Table 1 are of different nature. Some of them are impossible to measure and we do not have published research which explores possible co-relation between them. Do they influence each other, why and when?
c) There are no experiments, which overlap in terms of which factors are taken into account when calculating adhesion and friction. We could not compare results be-tween the experiments.
d) Some experiments deal with one or two factors and some keep having a set of factors in their focus of interest, which could affect the data values we entered in the data set.
e) There is no consensus from the literature on which factor(s) should be more important than the other(s), and in which combination of these factors may influence adhesion and friction more than in any other.
f) There are no publications, which cover the overall problem of "creating a minimal set of factors, which can guarantee a certain level of accuracy when calculating a/f".

It appears that it will NOT be reasonable to believe that we can collect all possible data for all factors, for the purpose of predicting a/f between rail tracks and train wheels, and assume that a reliable training data set is being created.

Therefore, bullets 1)-3) summarize the findings:

1) The calculation of adhesion and friction coefficient is not very simple. It is not a linear function and it is practically almost impossible to determine co-relation between all these factors, which is valid in any circumstances.
2) This information is not transparent, i.e. it is hidden in the literature, not available explicitly and not systematized, in order to assist in this research.
3) We should be extremely careful in the process of creating the training data set in this problem domain. The literature review and the data available from scientifically proven experiments do not guarantee a reliable training data set.

Therefore, if for any reason the initial perception that it would be possible to use learning technologies to predict slippery rail condition, might proof wrong.

2.4. Creating a Training Data Set

The process of structuring and creating the training data set is iterative and consists of:

(i) deciding upon factors from Table 1 which could be a good starting point in the calculation of a/f coefficient
(ii) finding values for a/f coefficients, based on chosen factors in (i), which are available in the literature
(iii) assessing the semantics of the content of the training data set and its factors in terms of repetition and overlapping and
(iv) defining a rationale behind decisions to either merge/split factors or re-categorize certain factors defined in (i) into training data set values, or both.

The process from (i)-(iv) deliberately keeps the problems from a)-f) in our mind. They have not been completely resolved, but the data set, which has serious issues with the lack of scientifically proven content, does need a user centered process for creating semantics which is essential in the *training*. This is exactly what has been defined in (i)-(iv). No automated software tools could help us in this particular case. They might be dangerous.

2.4.1 Defining Factors and Data Values for the Training Data Set

Figures 1 and 2 show the first attempt to choose the factors for the training data set, available from Table 1, and to enter their data values into the spreadsheet. Calculations for the adhesion coefficient is placed in

Column A. The top row in these figures outline the specific way of choosing factors from the left column in Table 1. However, each row in Figures 1 and 2 should corresponds to one experiment in Table 1.

	A	B	C	D	E	F	G	H	I	J
1	μ	dry	rain	layer Fe2O3	humid weather	T air, C°	T rail, C°	T lubr, C°	sunshine	oil
2	0.31	+								
3	0.265	+								
4	0.23	+								
5	0.2	+								
6	0.17	+								
7	0.15	+								

Figure 1.: The Data set, left side

K	L	M	N	O	P	Q	R
grease	organic contamination (leaf)	sand	snow	sufrace roughness	Speed, km/h	Herzian contact pressure, Gpa	Herzian contact pressure, Mpa
					5		
					25		
					50		
					75		
					100		
					125		

Figure 2.:The Data set, right side

Factors which were collected in the left column of Table 1 may overlap between experiments and therefore, all experiments were inspected together in order to find out if the overlapping factors give new insights into the calculations for adhesion coefficients. This had to be done before data values, for tables shown in Figures 1 and 2 were entered.

The rationale for choosing factors are discussed in the next four paragraphs.

Factors such as rain, the layer of Fe_2O_3, humid weather, sunshine, oil, grease and organic contamination (leaf) are easy to detect. The factor which defines dry condition is named 'dry' in our set of factors, but it does not mean "opposite to rain" condition. It indicates the **absence of water** that can appear on the rail head surface. However, it was found that in the UK railway experiments, where the changes of the coefficient of adhesion appear because of the train speed modification [9] a new **speed** factor has become important which was added to Fig. 1.

The next factor, isolated from Table 1 is Hertzian stress, or Hertzian contact pressure [9], [10] and calculations for its adhesion level were available. . Factors 'air humidity' and 'surface temperature'(of rails), 'sand' and 'snow' 'contact pressure', 'surface roughness', and 'lubricant temperature' had impact on the adhesion level with available calculation for the coefficients in the literature.

There was something very interesting found in [16], where the authors described the relation between the level of adhesion and different types of contamination, which can be dry, wet, oil, dry leaves and wet leaves They are all considered to be a contaminant of rail tracks. However, it was difficult to add all these different types of contaminants to our selection of factors because their number grew significantly and we do not know their correlation. Therefore, it was decided not to separate these factors. For example, we assumed that the condition for wet leaves would be equal to rain, i.e. we wanted to show the presence of "water" in the wheel/rail contact when we have wet leaves as contaminants.

The data-set contains 14 columns and 315 rows. The columns with the adhesion and speed values were filled by their numerical values. For factors which cannot have quantitative data values, symbol "+" was used to show the presence of that factor in the experiment, and symbol "-" for its absence. The factors which have data values or status (are they present or absent) not described in the experiment, were left empty, without their data values.

2.4.2 More Problems Encountered

One of the first problems encountered when inspecting the potential training data set was the absence of data values. However careful we were when defining the factors, we could not avoid this problem: the absence of values in the data set signals that they can not be found in the literature, i.e. in scientifically proved experiments. This in turn creates another problem: the lack of data is an obstacle for defining features and obtaining precision of ML classifiers.

The lack of data from the literature is not unusual. All these experiments have to be done by monitoring maximum 1-2 factors and keep all others constant. This is how these experiments could give viable results. They do not have an issue of missing information: they just keep a few factors "constant" in order to monitor a chosen factor for its impact on adhesion.

Therefore, this problem is aggravated with the following discoveries:

a) Most of the rows contain information with coefficient of adhesion value in column A and the value of one factor which influences it. The rest of the data values for other factors are unmarked (factors have no data values).
b) Values for some of the factors in some experiments might be irrelevant. They are sometimes either not measured or are unknown.
c) Some data values for factors are numbers, some are ranges of numbers, some data values for identical factors are given in different

measurement units and some are just simply NOT PRESENT.

Examples of missing values in the data set are numerous, but legitimate. Therefore all options where carefully explored in order to minimize the number of missing values, without damaging the semantic of data.

We give one example where our discretion was used, while not interfering with the semantic of data. For example, Figure 3 shows that the row which contains a data value for the coefficient of adhesion has only one data value which is "+" in the *grease* column. On the one hand, the values for temperature for air and rail, surface roughness, speed, and contact pressure are REALLY UN-KNOWN for this value of adhesion. This is not all. Oil and organic contamination factors are not applicable to this example, and in our data set, it looks like information about them is missing! This is a significant anomaly in the data set and quite dangerous to use in any learning algorithms.

sunshine	oil	grease	organic contamination (leaf)
		+	

Figure 3: "Empty space" problem

Figure 4. shows a chosen "solution", which might eliminate this problem and possibly populate the data set with more values. When the experiment stated that for dry railway condition the values for rain, humid weather, oil, grease, and organic contamination are not applicable, these values were marked as "-" with a yellow background color. This is how far we could go with a potential labeling of data without impacting the semantics of data taken from the experiments.

sunshine	oil	grease	organic contamination (leaf)
	-	+	-

Figure 4 Potential solution for a missing data value

Figure 4 allows for a better visualization of the content of the data set and may help in the definition of the classifier. The values for the Fe_2O_3 and sunshine conditions (factors) were not often found in the literature and we found only one publication, but with the values for friction, not for adhesion. There were also numerous problems with the definitions of contaminants, but no co-relation was found between their data values. Do contaminants interfere with each other and in which conditions?

3. Discussion and Conclusions

The previous section illustrates the process of creating a training data set for predicting slippery conditions between hard surfaces applicable to rail traffic. However precise and determined we were to create a coherent and correct semantic of the data set, numerous problems were detected. Apparently, every step carried out in the process of creating a training data set, ended up with a list of problems.

It started from the beginning, when it was discovered that it was not possible to measure adhesion and friction between wheels and rails from real life experiments. Therefore no live generated data, created by trains in motion, can be used for any data set. This means that it can not be assumed that rail transport can generate data, while trains are running, which could be used in calculations of a/f coefficients.

In this study, the process of data collection created a data set which is accurate, but it is half-empty, due to numerous missing data. Under no circumstances could they be substituted with values because there are no scientifically proven experiments for such an action.

Therefore, most of the time problems were collated, which could not help in proposing solutions for creating a reliable data set.

However, the following options can be debated.

Option 1 - Small manual improvements of the data set could be carried out without affecting the semantic of data, by adding various labels, symbols and tags to the data set and possibly merging columns to minimize the impact missing values may have on ML classifiers. There are commercial software tools and data science environments which encourage us to do so.

Option 2 - This data set and experiments with all possible options could be used when defining a ML classifier and, manipulating feature selections may detect if it can help in minimizing the impact of deficiencies in our data set [24,25,26].

Option 3 - Follow the philosophy from Mechanical Engineering practices and mirroring each experiment with a particular settings in/fraction of the data set, should be tested. For example, a (the) main factor(s) used in one experiment are to become the only feature(s) of a particular instance of the classifier. How would feature selection impact classifier precision upon our data set?

Option 4 - A classifier could be used just as an indication for potential danger in rail transport: the data set may just raise awareness of potentially slippery conditions in transport without making firm conclusions. This may be useful in situations when changes in conditions in transport are either clear, or obvious or frequently described in the data set, and the

classifier would recognize them. Would rail companies need this?

Option 5 – There is a possibility of investigating the active learning model [27] which has been used in improving the quality of a training data set in unsupervised learning [28]. The process of creating the training data set is not automatic and relies on human involvement. Therefore it would be still cost effective to pre-build an ML classifier which can label missing data values through human intervention. This newly "trained" data could be merged with the current training data set. Similar ideas have been used for unsupervised deep learning for image classifications [29], and SVM for text classification [30]. Similarly, transfer learning might be of help here [31,32].

Option 6 – Various models of data collections for the purpose of running ML algorithms [33] could also be investigated. However, the sensitivity of this problem domain and the fact that we need the highest level of classifier's accuracy upon the training data set might not open a door to data from shared sources.

3.1. Could Full Scale Exploratory Data Analytics Help?

The competing worlds of data science and statistics have not been properly debated in publications and it is no secret that both sides have different methods and goals. In computational algorithms, which are underpinned by learning and predictive software technologies, we focus on their accuracy, and the accuracy of prediction models we create for a particular problem domain. Data scientists walk from one model to another and keep an eye on their accuracy. This is not the same as in the world of statistics. Statisticians carefully verify their models, by looking if the data, chosen for a model, does not violate its conditions and assumptions (i.e. the model's semantics), before they run it. Therefore, it does not surprise us that data scientists advocate the process of exploratory data analytics (EDA) [34,35], which uses summary statistics and visualization in order to extract insights from the data which could help in modelling ML algorithms.

Practices of EDA from the industry, which claim to guarantee the extraction of meaningful insights from raw data, using various software tools, suggest using descriptive statistics, correlation, analyses of variance and simple semantic grouping of data, which could also address missing values in the training data set, which obviously make sense. This study uses a specific EDA, which is partially described in section 2.4.2. It included only correlation, grouping of data and addressing missing values. Descriptive statistics and analysis of variance was not considered because throughout the process of collecting data, it become obvious that calculations of a/f coefficients, essential for training algorithms were not always reliable, as debated in section 2. Data entries into the training data set were restricted to values which are guaranteed and CORRECT, but

a) the data set could not grow and

b) the quantity of missing data did grow.

Consequently in order to address missing values, through data grouping and correlation, as shown in Figures 3 and 4, nothing "new" was discovered, which was not previously, or intuitively obvious. There was no "insight" from data through EDA which could not have been obtained through the process of creating the training data set.

It is important to note that correlation, which is considered to be a part of EDA, proved to be more useful in the definition of a classifier, particularly when measuring if feature selections affect the precision of predictions [36]. However, feature selections should be performed upon a relatively reliable training data set, and it is debatable if, during the process of classifier definition, it is wise to go back to the training data set and try to "get more insight" from it. For computer scientists, it is prudent to distinguish between

(i) defining the semantic of data and

(ii) defining a computational model which uses the data.

In other words there should be a subtle difference between *defining a training data and defining a ML classifier.*

For readers interested in results of co-relation between selected features of the ML classifier and a/f levels, the visualization of potential co-relation between data values of a selection of features and the average value of adhesion for the same features, was used. It was important to find out if changes in slippery to non-slippery conditions, and vice versa, appear when the data values for each feature change. Apart from a few known features: axle load in kN, Hertzian contact pressure, speed, surface roughness and lubricant temperature, which have also been described in the literature on the calculations of a/f coefficients, there were no further or "new" insights found within the semantic of the data set.

To summarize, if EDA is supposed to be performed as a part of initial investigation on data, to detect patterns and anomalies in particular, then in this particular problem domain, EDA failed to give us any results which could be interpreted as a new insight in the data set.

3.2. Conclusions and Potential Future Work

The main outcome of this study is that there is no evidence in scientifically published papers that ML classifiers may predict accurately slippery conditions in rail transport, because we can not guarantee the reliability of its training data set. Only by obtaining data from real life and scientifically proven experiments could a reliable, i.e. trustworthy training data set for ML classifiers be created.

However there are publications which address the problem of low adhesion in rail transport by looking at adhesion estimation through model-based and multiple-model based methods [37], model based condition monitoring [38], vibration [39], motor current differences [40] and hybrid slip control methods [41], [42], [43], to mention just a few. It is unfortunate that in most of these cases, low adhesion is actually detected AFTER it occurs, which makes these solutions impractical to use in real life and thus its results are unsuitable for a training data set. Furthermore, some of these methods require, for example, precise measurements of rotational wheel speed and absolute vehicle speed, which are both problematic to obtain in real time. Some of these methods detect adhesion only during movements, and some have problems with the definition of velocity, which aggravates the problem.

Due to the complexity of the problem, it is not possible directly to measure all factors which create low adhesions, and the lack of real-time information which fully describes areas of low-adhesion between hard-surfaces exuberates the problem.

An interesting study for the detection of adhesion and its changes in [44] exploits the variations in the dynamic behavior of the railway wheelset caused by the contact condition changes. The authors use a bank of Kalman filters, designed at selected operation points for the adhesion estimation and fuzzy logic in order to identify the contact conditions by examining the residuals from the Kalman filters. The increased number of these filters may improve the accuracy of calculations, but they require time, which is impractical in real life and in turn, may affect a/f coefficient values. Also, Kalman filter accuracy is limited in cases of low friction and large track irregularities [45].

Therefore, the results of these experiments were not considered to be suitable for the training data set in this study.

What could be proposed?

In the first place, the scientific community might need around table talk, where mechanical engineers, physicists, computer scientists and statisticians would sit together and analyze the problem of and agree on the way of determining adhesions and friction levels. We are in the third decade of the 21st century, where computational models accommodate current ML algorithms, which differ from the traditional process of performing calculations for detecting slippery conditions in rail transport, which existed more than 100 years ago.

Therefore, for computer scientists the following question still holds. If the process of creating a reliable training data set for this problem domain is troublesome, with numerous obstacles encountered, should not this process have been abandoned earlier?

There are two possibilities which should be considered in order to address the question above, but they do not rely solely on predictive and learning technologies. If data scientists still wish to predict slippery levels in rail transport, based on all known factors which influence them, they should start looking at software solutions outside statistics and predictions. Just by assuming that a vast amount of data, which may be potentially generated by modern sensor technologies in rail transport, would guarantee the success of predictions for slippery driving conditions, through ML classifiers, is simply wrong.

The first possibility considers the power of semantic technologies, their manipulation of semantics through description logic and reasoning with rules, we could use our training data set in the definition of a semantic classifier, which would be able to understand the meaning of missing data. In other words, a dynamic definition of a classifier and its features would be governed by ontological reasoning and we could guard the classification and learning from the data set through reasoning.

The second possibility focuses on algorithms. It appears that all of them, currently used in ML represent very well tried and old probabilistic or linear models. Has the time arrived to start thinking differently in terms of algorithms, which would support what we wanted to do here? Should we start thinking about data centric ML instead of focusing on algorithms?

There is absolutely no way of knowing exactly, what the effectiveness of the ML classifiers in this study would be, even if all available ML classifiers are run upon the data set. They might score high in their precision level [36], [46], but would it be enough for sitting comfortably on a train managed by predictions based on the created training data set? Without thorough analysis of data, while preparing the training data set, how would anyone know WHERE a danger is? It is not that data scientists might not be aware of deficiencies of the data. They will not be aware of potentially dangerous results.

Should we blame algorithms for this?

4. Acknowledgment

We wish to thank Professor Mehdi Gebreil Mousavi, from University of South East Norway in Kongsberg, Department of Science and Industry Systems, for his constructive comments on the content of our training data set.

5. References

[1] A., Halevy, P., Norvig, O. Pereira, "The Unreasonable Effectiveness of Data", IEEE Intelligent Systems, March/April 2009, pp 8-12.

[2] L.E. Buckley-Johnstone "Wheel/rail contact tribology: Charactering low adhesion mechanisms and friction management products", Doctoral thesis, University of Sheffield, May 2017.

[3] M. Malvezzi, L., Pugi, S. Paipini, A. Rindi, P. Toni, "Identification of a wheel–rail adhesion coefficient from experimental data during braking tests", Proceedings of the Institution of Mechanical Engineers Part F Journal of Rail and Rapid Transit, 2013, 227(2), pp.128-139

[4] Y. Zhu, U. Olofsson, K. Persson, "Investigation of factors influencing wheel–rail adhesion using a mini-traction machine", *Wear*, 292, pp. 218–231, 2012.

[5] Y. Zhu "Adhesion in the wheel–rail contact" Doctoral thesis, Royal Institute of Technology, Stockholm, Sweden, November 22, 2013.

[6] J. Lundberg, M. Rantatalo, C. Wanhainen, J. Casselgen "Measurements of friction coefficients between rails lubricated with a friction modifier and the wheels of an IORE locomotive during real working conditions", Wear, Volumes 324–325, 15 February 2015, pp 109-117.

[7]Y. Zhao, B. Liang, S.D. Iwanicki, "Estimation of the friction coefficient between wheel and rail surface using traction motor behaviour", Journal of Physics: Conference Series, Volume 364, May 2012.

[8] ORE report 2, question b164. Adhesion during braking, and anti-skid devices, Utrecht, Netherlands, 1990, at https://www.shop-etf.com/en/technical-documents/technical-and-research-reports/b-b153-b49-rolling-stock-brakes-diesel-engines-running-gear/b-164-adhesion-during-braking-and-anti-skid-devices

[9] H. Harrison, "The development of a low creep regime, hand-operated tribometer",*Wear*, 265(9-10), 2012, pp. 1526–1531.

[10] K.S.Baek, K, Kyogoku, T. Nakahara, "An experimental study of transient traction characteristics between rail and wheel under low slip and low speed conditions". *Wear*, 265(9-10), 2008, pp. 1417–1424. 2008.

[11] R. Lewis, U. Olofsson,"An alternative method for the assessment of railhead traction".*Wear*, 271(1-2):62–70, 2011.

[12] T.M. Beagley "Severe wear of rolling/sliding contacts" *Wear*, 36(3):317–335, 1976.

[13] U. Olofsson, K. Sundvall, "Influence of leaf, humidity and applied lubrication on friction in the wheel-rail contact: pin-on-disc experiments", Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit, 218(3):235–242, 2004.

[14] Y. Zhu, "Adhesion in the wheel-rail contact under contaminated conditions", PhD thesis, KTH Royal Institute of Technology, Sweden, 2011.

[15] Y. Zhu, U. Olofsson, K. Persson, "Investigation of factors influencing wheel–rail adhesion using a mini-traction machine", *Wear*, 292:218–231, 2012.

[16] F.A. Gallardo-Hernandez, R. Lewis, "Twin disc assessment of wheel/rail adhesion". *Wear*, 265:1309–1316, 2008.

[17] Moore, D.F. Principles and applications of tribology, Elsevier, 1975.

[18] X.V. Xuesong, J.W., Zhang, J. Chen, "Wheel/rail adhesion and analysis by using full scale roller rig". *Wear*, 253:82—88, 2002.

[19] K. Nagase, "A study of adhesion between the rails and running wheels on main lines: results of investigations by slipping adhesion test bogie". Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit, 203(1):33–43, 1989.

[20] M. Ishidaa, T. Nakahara, H. Chena, T. Bana, "Adhesion between rail/wheel under water lubricated contact", *Wear*, 253:75–81, 2002.

[21] R. Lewis, E.A. Gallardo-Hernandez, "Twin disc assessment of wheel/rail adhesion. *Wear*, 265:1309—-1316, 2008.

[22] J. Zeng, Z. R. Q. Zhou, Y. Liu, Z. R. W. Jin, H. Zhang, Z.F. Wen "Adhesion experiment on a wheel/rail system and its numerical analysis. *Journal of Engineering Tribology, Proceedings of the IMechE Part J*, 218(1):293–303, 2004.

[23] H. Chen, A. Namura, M. Ishida, T. Nakahara, "Influence of axle load on wheel/rail adhesion under wet conditions in consideration of running speed and surface roughness", *Wear*, 366:303–309, 2016.

[24] N. Danilchanka, R. Juric, "The Process of Creating a Training Data Set: Lesson Learned from Mechanical Engineering", In SDPS 2018 Workshop on Accountability of AI, Bologna, Italy, December 2018, available at https://www.sdpsnet.org/sdps/documents/sdps-2018/SDPS%202018%20proceedings%20ver%205.pdf#8

[25] L. Li, H. Lin, "Challenges of Feature Selection for Big Data Analytics", in IEEE Intelligent Systems, March/April, 2017, pp 9-15

[26] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," in Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence), (C. N_edellec and C. Rouveirol, eds.), vol. 1398 of Machine Learning: ECML-98, Springer, Berlin, Heidelberg, 1998.

[27] B. Settles and M. Craven, "An analysis of active learning strategies for sequence labeling tasks," in Proc. Conf. Empirical Methods Natural, Lang. Process., 2008, pp. 1070-1079.

[28] G. Joo, C. Kim, "MIDAS: Model-Independent Training Data Selection Under Cost Constraints", *ACCESS, 2018.2882269,* December 27, 2018.

[29] A. J. Joshi, F. Porikli, and N. Papanikolopoulos, ``Multi-class active learning for image classification," in Proc. IEEE Conference on Computational Vision and Pattern Recognition (CVPR), Jun. 2009, pp. 2372-2379.

[30] S. Tong and D. Koller, ``Support vector machine active learning with applications to text classification," J. Mach. Learn. Res., vol. 2, pp. 45-66, Nov. 2001.

[31] S.J. Pan, J. Yang, "A survey on transfer learning", IEEE TKDE, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[32] H. He, E. A. Garcia, "Learning from imbalanced data," IEEE TKDE, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[33] J. Roh, G. Heo, S. E. Whang "A Survey on Data Collection for Machine Learning: a Big Data - AI Integration Perspective, May 2019 https://arxiv.org/pdf/1811.03402.pdf

[34] D. Motin, M. Lissandrin, Y. Velegrakis, T. Palpanas,"New Trends on Exploratory Methods for Data Analytics", Proceedings of the VLDB Endowment, Vol. 10, No 12, 2017.

[35] E. Karageorgiou, "The logic of exploratory and confirmatory data analysis, Cognitive Critique. 2011; 3(Winter):35-48 available at http://www.cogcrit.umn.edu/docs/karageorgiou_11.shtml

[36] N. Danilchanka , R. Juric, R. (2020) The effectiveness of ML Classifiers in Determining Slippery Condition in Rail Transport, under review in ASME Journal of mechanical Design.

[37] G. Charles, R. Goodall, R Dixon, "Model-based condition monitoring at the wheel–rail interface", Vehicle System Dynamics, 46(S1):415–430, 2008.

[38] C. Christopher P. Ward, R. M. Goodall, R. Dixon, G.A. Charles, "Adhesion estimation at the wheel–rail interface using advanced model-based filtering", Vehicle System Dynamics, 50(12):1797–1816, 2012.

[39] I. Hussain, T.X. Mei, R.T. Ritchings "Estimation of wheel–rail contact conditions and adhesion using the multiple model approach", Vehicle System Dynamics, 51(1):32–53, 2013.

[40] Y. Mei, Wilson, "A mechatronic approach for anti-slip control in railway traction", IFAC Proceedings, 41.2:8275–280, 2008.

[41]T. Watanabe, M. Yamashita, "Basic study of anti-slip control without speed sensor for multiple motor drive of electric railway vehicles". In Proceedings of the IEEE Power Conversion Conference-Osaka 2002 (Cat. No. 02TH8579), volume 3, pp. 1026– 1032, 2002.

[42] K. Xu, G. Xu, C. Zheng, "Novel determination of wheel-rail adhesion stability for electric locomotives", International Journal of Precision Engineering and Manufacturing, 16(4):653–660, 2015.

[43] D.-Y. Park, M.–S. Kim, D.-H. Hwang, J.–H. Lee, Y.-J. Kim, "Hybrid re-adhesion control method for traction system of high-speed railway". In ICEMS'2001. Proceedings of the Fifth International Conference on Electrical Machines and Systems (IEEE Cat. No. 01EX501), volume 2, pages 739–742., 2001.

[44] T.X. Mei, I. Hussain, "Detection of wheel-rail conditions for improved traction control", Proceedings of IET Conference on Railway Traction Systems (RTS 2010), April 2010.

[45] C. Li, S. Luo, C. Cole, M. Spiryagin, "An overview: modern techniques for railway vehicle on-board health-monitoring systems", Vehicle system dynamics, 55(7):1045–1070, 2017.

[46] R. Juric, "How Biased Can AI Be?" In Proceedings of the SDPS 2018 Workshop on Accountability of AI, Bologna, Italy, December 2018 available at https://www.sdpsnet.org/sdps/documents/sdps-2018/SDPS%202018%20proceedings%20ver%205.pdf#8