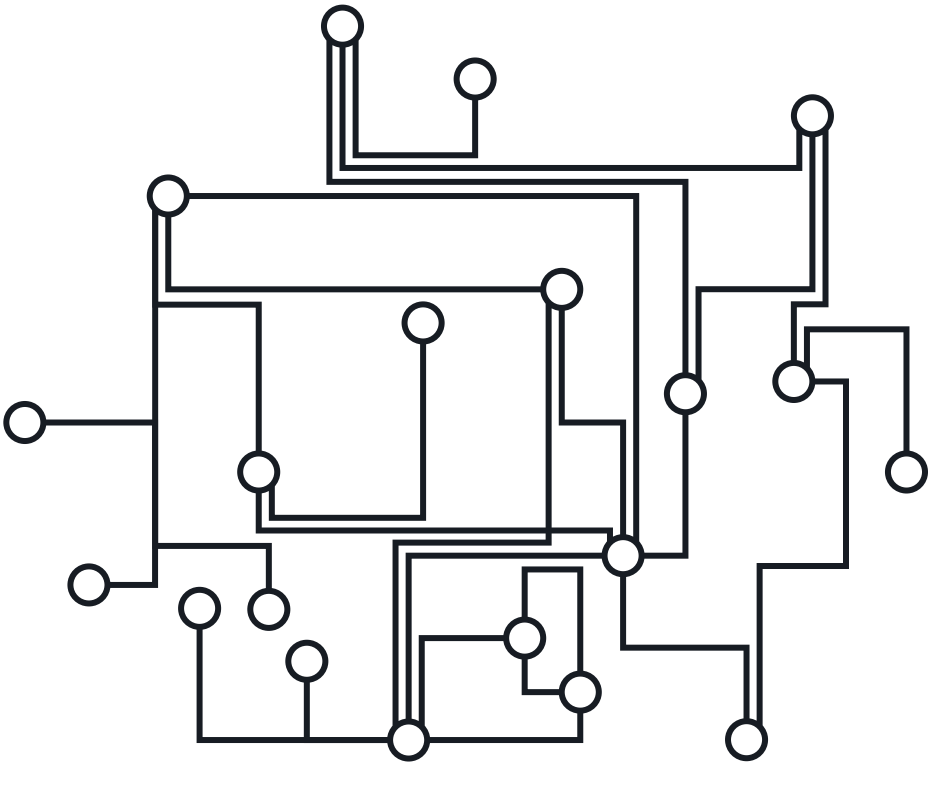
Analysis for a possible implementation of a machine learning algorithm in the packaging process



Paul Christiaan Alexander Yoe

04.08.2018

DPM-nr. 1543

Preface

I would like to than Mr. Frank Stahlhut, Dr. Ir. Roy Damgrave, Ms. Erika Cammi and Mr. Benjamin Schehrer for their coaching and engagement.

I owe my thanks to all colleagues which supported me during the conduction of the thesis.

Abstract

This research aims to enhance the digital product development process. One of the product development processes is the digital mock-up (DMU) process (Hirz, Dietrich, Gfrerrer, & Lang, 2013). A task of the DMU operation makes sure all Three Dimensional Computer Aided Design (3D-CAD) parts fit correctly into the vehicle. 3D-CAD parts are assembled to develop the product virtually. The 3D-CAD parts can clash. The assignment of the clash severity is aimed to automatically assess. This will be completed by using a machine learning algorithm.

The goal of machine learning algorithms is to develop a machine as if it is has knowledge (Ertel, 2016). Machine learning algorithm is a technique which is used for pattern recognition. The pattern will be 'remembered' by the algorithm. To recognise a pattern, the algorithm is trained on data features with corresponding outcome. This is known as supervised learning. The data features existed of the Hausdorff distance and material data.

The machine learning algorithm was able to predict the severity of the collisions by making use of a classification approach. This classification approach is explored in an artificial neural network (ANN) and a Random Forest algorithm. The model for the ANN was the multilayer perceptron (MLP) method.

The first tests proved that a machine learning algorithm could become a valuable asset. After several tests, the decision was made to enhance the features in order to improve interpretation for the machine learning algorithm. The results did not improve on this enhancement. Afterwards, different machine learning algorithm configuration are tested. A small neural network configuration proved to be most successful. Further research need to be conducted in order to successfully apply the machine learning algorithm.

Table of content

1 Introduction:

The digitalisation of the world is growing at an ever accelerating pace. It increases the amount of data generated drastically. All this data is a driver behind machine learning technologies such as artificial intelligence, natural language processing and machine learning (Reinsel, Gantz, & Rydning, 2017). The technology of machine learning algorithm is a tool to recognise patterns. These algorithms are trained to be highly accurate in their prediction of the pattern. They are deployed in many industries. The medical industry can predict the flow of patients (Jiang, Chin, & Tsui, 2018) and the automotive industry uses this for autonomous driving (Yoon & Kum, 2016).

The development of a machine learning algorithm is complicated. Most algorithms are trained on certain data features. The features are the base to recognising the pattern and final decision of the algorithm. Though, the best input features are unknown and are often derived from human decision making. The degree of impact of features on the algorithm can be obtained by analysation and exploration.

This research aims to gain insight in an opportunity to enhance the digital product development process in the automotive industry. This process virtually construct the product. One of the product development processes is the digital mock-up process (Hirz, Dietrich, Gfrerrer, & Lang, 2013). A task of the digital mock-up operation makes sure all Three Dimensional Computer Aided Design (3D-CAD) parts fit correctly into the vehicle and can operate correctly to fulfil their function. Here all vehicle components are digitally constructed and assembled such as the leather for chairs, cables for electric connections and bodywork parts. Collisions can occur during the development of these assemblies. Later in the development process, these errors could have a huge impact on the ability to build prototypes or production vehicles.

At the moment, assessment of the collisions is done manually. This is a time consuming activity. It is estimated that this process could be assisted by a machine learning algorithm. The manual assessment of collisions indicated that assessment decisions are occasionally based on meta-data. Meta-data could function as the data features to train the machine learning algorithm. If the machine learning algorithm is capable of assessing collision correctly. The machine learning algorithm can reduce collision assessment to a comprehendible operation if the algorithm is able to correctly assign an extensive amount of collisions. It will produce an overview of the salient collisions as a result. Therefore, the goal of the thesis is to find the optimal meta-data features for a machine learning algorithm which allows to automatically assess collisions by their severity.

2 Research approach

This research will add to the literature of functional machine learning algorithm possibilities. It is an approach how a business process can be supported. The Cross Industry Standard Process for Data Mining (CRISP-DM) is used in order to conduct the research in a structured approach. The CRIPS-DM methodology consist of six processes: Business understanding, Data understanding, Data preparation, Modeling, Evaluation, Deployment (Shearer, 2000). An overview of the CRISP-DM methodology is shown in figure 1. The business understanding process will be elaborated. The business process itself with the different calculation processes will be described. Before data understanding, the generally encountered problems for designing a machine learning algorithm will be explained. This is due to the non-sequential design process of a machine learning algorithm. Afterwards, the combination of the business process with the algorithm overview is further elaborated. For example, possibilities to present the data of the packaging process to the algorithm will be presented. Then the processes of data understanding, data preparation, modelling and evaluation of the machine learning algorithm to will be described in three iterations. Each iteration will compare the correlation of the input data and the results of the algorithm. A general conclusion of the three iteration will be described. Afterwards, a recommendation on future deployment and further research will be given.

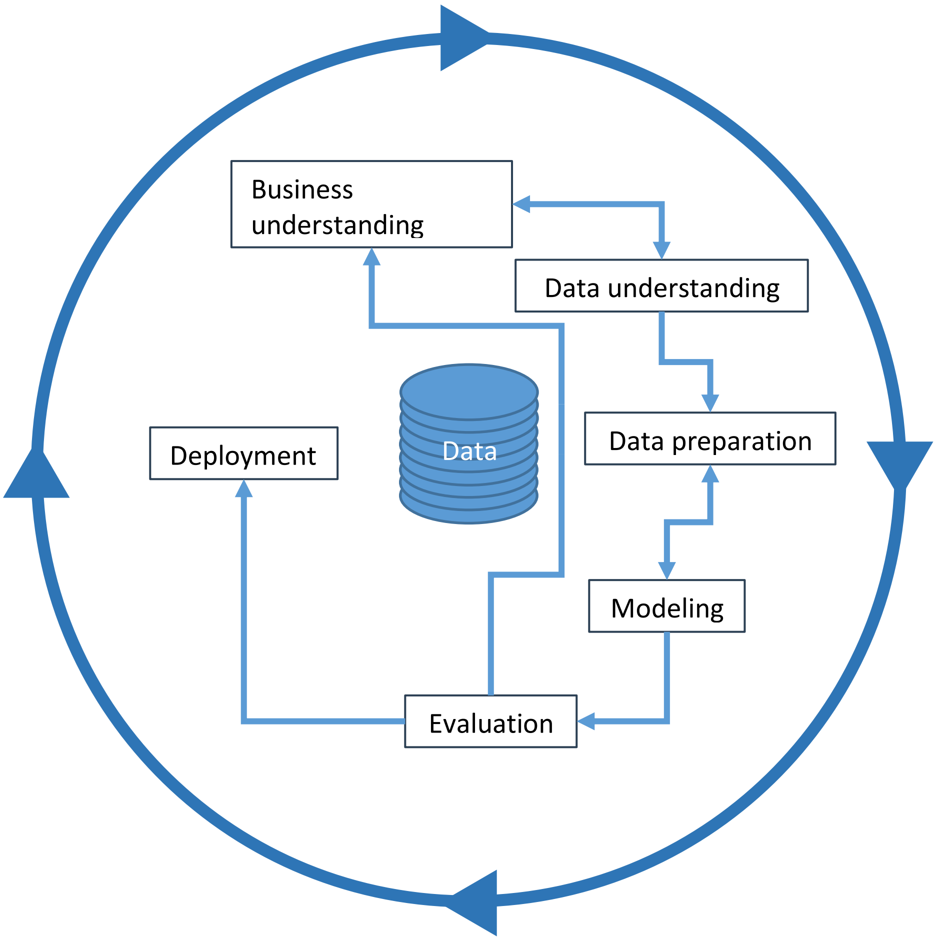


Figure 1: Cross Industry Standard Process for Data Mining (CRISP-DM) methodology

3 Packaging process

A product documentation structure exist out of three level (Groll, 2008). The geometric level holds the parts which are still in the design stage and can be altered. All parts are modelled into 3D-CAD by constructors. 3D-CAD models describe the form of the part and are assembled when an optimal geometry is achieved. The connection between the assembly and single/rough part is the connection which is studied. Each level could be a starting point of vehicle development. The development can evolve into both direction. The research will be conducted in the lower levels of the structure, single part and rough part. Figure 2 represents the product documentation structure.

Product	Product
Technical	Mayor component
	Assembly
	Single part
Geometric	Rough part

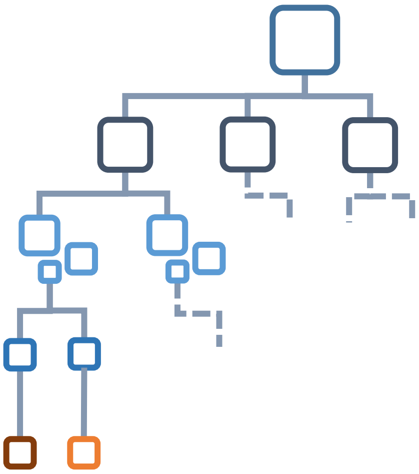


Figure 2: product documentation structure

3.1 Digital mock-up

The virtual product development is conducted through three process stages (Hirz, Dietrich, Gfrerrer, & Lang, 2013). The first stage is the 3D-CAD design. It includes the creation of the geometry and product structure. The second stage is the environment were the machine learning algorithm will be functioning in. This stage is called the digital mock-up (DMU) process (Hirz, Dietrich, Gfrerrer, & Lang, 2013). Figure 3 shows a zoomed in version of the intergraded virtual product model. DMU is a digital process which copes with the variant possibilities of the full-scale product. Each part in the product is controlled on its functioning and clearance. DMU process functions to govern the product possibility (Josip Stjepandić, Nel Wognum, & Verhagen, 2015).

During the DMU process all parts are checked if the parts clash through each other's geometry (Josip Stjepandić, Nel Wognum, & Verhagen, 2015). The packaging process ensures crucial collisions are detected and resolved. Though, not all collision are relevant while it is hard to model the behaviour of material characteristics into CAD geometries. An example of a not relevant collision would be if a rubber part and a steel part intersect each other in CAD while the rubber part can bend around the steel part in the real world. Though, it could be relevant when a steel plate part interconnects with another steel plate part. This cannot occur in the real world and is thus a relevant collision. This data is relevant while it can make a difference in whether the vehicle can be build. These material properties are added to the CAD-geometry as meta-data. Metadata describes additional data such as material and the collision status.

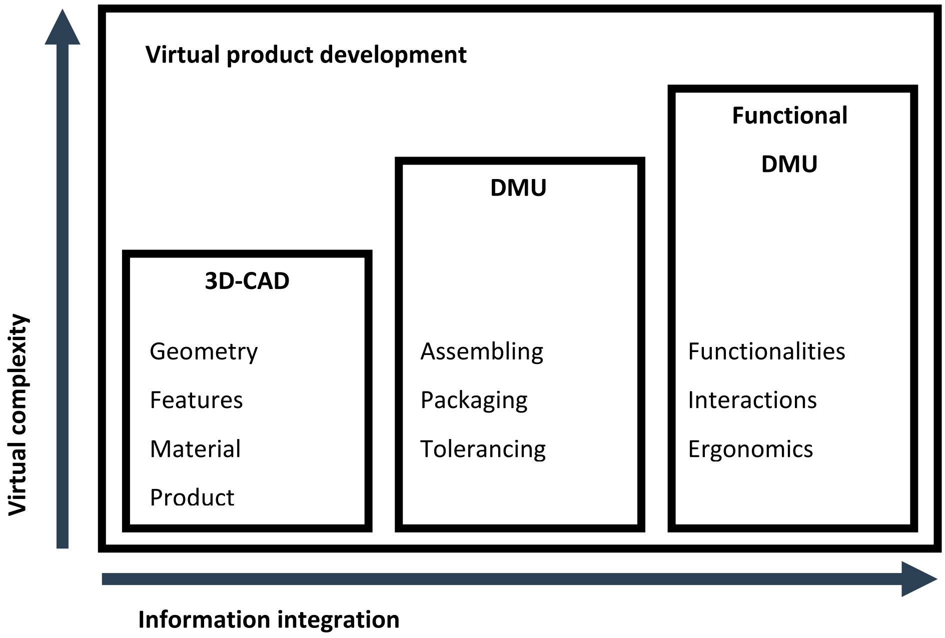


Figure 3: Virtual product development (Hirz, Dietrich, Gfrerer, & Lang, 2013)

3.2 Packaging structure

The packaging process makes sure all parts fit into the digital vehicle assembly. Each packager governs its own construction space and is the specialist in the specific space of the vehicle. They check which geometrical changes occurred and where they should perform a new evaluation while the 3D-CAD parts can collide with each other. Every collision pair exist out of two 3D-CAD parts. One 3D-CAD part can hold multiple 3D-CAD geometries. The different 3D-CAD geometry is calculated against the other CAD-geometries of the opposed rough part. Hence, a collision pair can result in multiple collisions. The calculation process which will be elaborated in the next chapter. This information is stored in a new Product Data Management (PDM) item which includes every collision. PDM is the system where all part information is stored (Groll, 2008). The new PDM collision item is shown in figure 4.

The collision results are provided back to the packager. These results consist out of an extensive list of all collisions. Packagers will go through each of the intersection and manually alter the geometry or assess the geometry. The assessment consist out of four statuses and are presented in Table 1. The collision assessment is not changed until any geometrical change occurred. In addition, these assessment statuses could be used a means of communication medium between packagers. There is no clear guideline or expertise of the packager required to evaluate collisions.

Relevant	The geometry of a component causes a serious conflict with respect to manufacturing, tooling, runtime etc., and needs urgent revision.
Relevant 2	The geometry of a component causes a non-critical conflict but must still be revised.
Not relevant	A conflict or a component of a conflict is correct in terms of *packaging* and does not have to be edited.
New	A conflict is new when two components that were conflict-free now have a conflict.

Table 1: Collision statuses

The assessment of collisions is time consuming and is not optimal while time is spend on not relevant collisions. During this process the constructer will lose focus if multiple not relevant collision occur which can lead to mistakes. Assessment prioritisations are aimed to provide insight in the severity of the collision. This is difficult due to the lack of consistency in collision assessment while there are different requirements in other construction spaces. A collision could occur relevant in a construction space but could not occur relevant in another. In addition, the list is too extensive to comprehend. Ideally, the collision list could be reduced by making use of a machine learning algorithm which would provide a prediction, based on already assessed collisions. Therefore, packagers are the stakeholder while collision assessment is aimed to be enhanced.

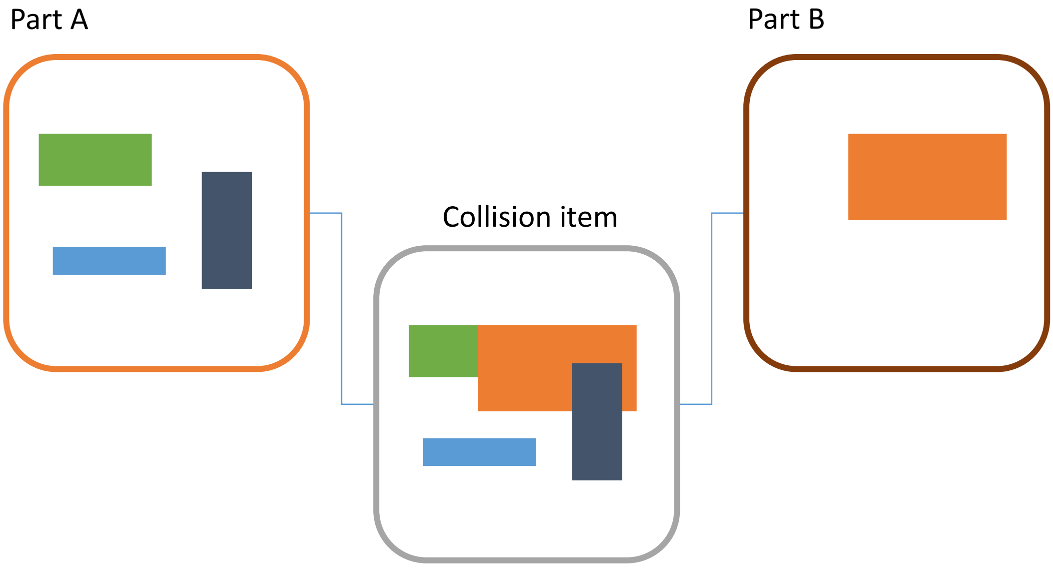


Figure 4: Creation of collision item

3.3 Collision calculation

The calculation process begins with the calculation which provides the insight if geometry is colliding. Every geometry in one part is calculated against all geometries of the opposite part. This is shown in figure 5. The result value of the collision are calculated in order to determine if there are collisions. The result value is the longest distance between the parts. The value is negative if there is a penetration between two geometries. The result value is positive if there is a space in between the geometries.

The result value is marked red in figure 6. The vertical distance could be more relevant to the assessment of collisions while the blocks are stacked vertically. In order to acquire the most relevant distance another calculation method is used after the result value is calculated.

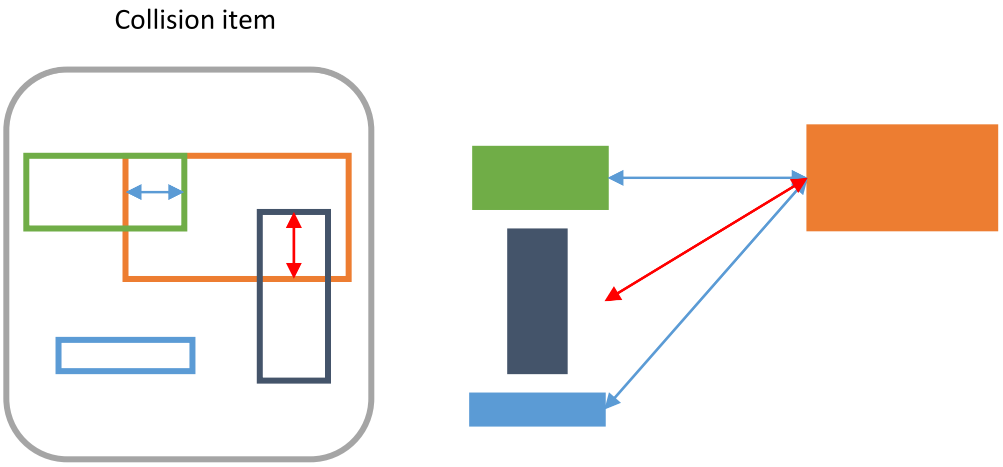


Figure 5: Explanation of collision calculation

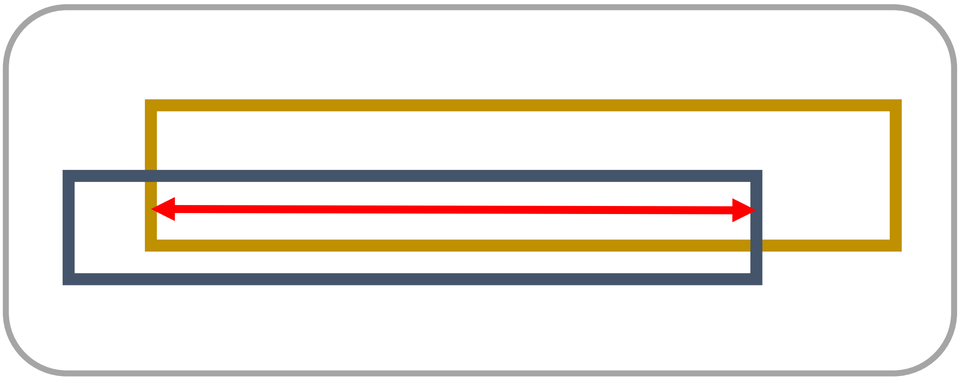


Figure 6: Collision with the result value marked red

In order to gain more insight in each collision, a new calculation method is applied. This calculation is able to calculate the geometrical intersection volume between the intersection partners. An important result of this calculation is the Hausdorff distance. The Hausdorff distance is defined as the maximal value of all the minimum distances (Guéziec, 2001).

An example of the Hausdorff distance is shown in figure 7. The Hausdorff distance must be determined from each point on the first, star shape, to the closest point on the second triangle shape (Guéziec, 2001). The distance with the largest value is returned. This calculation must also be conducted vice versa. Then the maximal value is determined by comparing the largest values. This value is the Hausdorff distance. For example a star, shape one, and a triangle, shape two, collide. The orange distance is the result value distance. All other distances are minimal distances. The red distance is the maximum distance of all. Therefore, this distance will be returned as the Hausdorff distance.

$$H(S,T) = \max(\max_{v \in T} d(v,S)\,, \max_{v \in S} d(v,T))$$

This Hausdorff distance is relevant for the packager while this shows and how the collision can be resolved with minimal effort. It is the most accurate value to measure how critical the collision is. Each collision calculation returns the result value and the Hausdorff distance of all collisions. The status of new is acquired when the collision is never assessed or geometrical changes occurred.

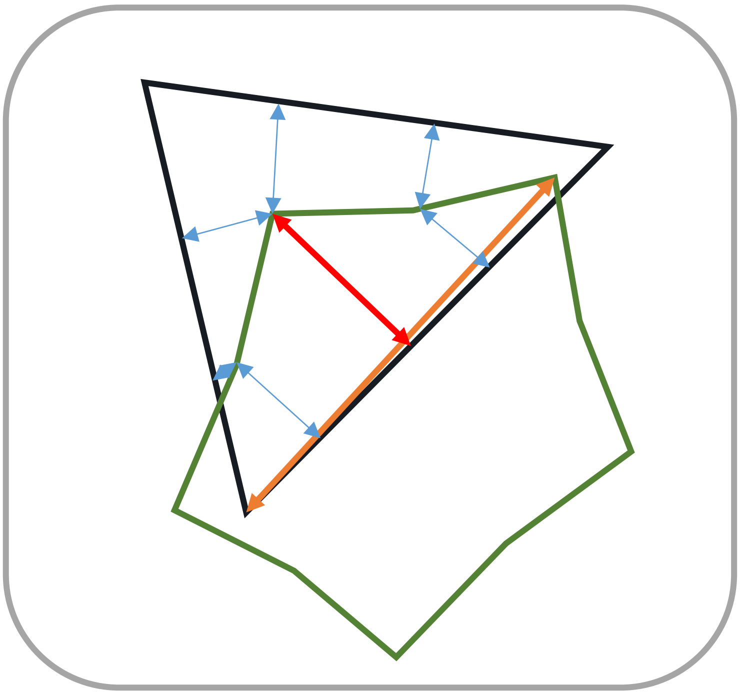


Figure 7: Hausdorff distance

3.4 Data origin

The available data sets consisted out of two vehicles. One was a case study on a passenger car and the other a productive Van project. The collisions assessment in each data set was different. The passenger car collision assessments were altered according to a specific guideline. Though, the Van dataset was assessed by packagers. Therefore, passenger data includes more consistency because all collisions were treated in the same way. The Van data set includes more expert know how while the collisions are assessed with their expertise. The Van data set reflects reality because it takes into account different packagers with their expertise and construction spaces.

The collision data originates from temporary SQLite database and the PDM system. Each collision calculation process produces more metadata for the SQLite database. This is because each collision partner can be modelled out of multiple CAD models. Only the collision with the highest result value will be transferred to the PDM which assures the correctness of the PDM data.

It proved to be difficult to exclude wrong attributes of metadata to the assessment of the collision in a SQLite database. Obtaining data out of the PDM system did not have comparable problems. The passenger car originates from both the PDM system and a SQL-databases. The Van dataset originates from SQL-databases.

4 Part I: Algorithm overview

The possibilities of machine learning algorithm in order to enhance the packaging process is researched. This is based on the indication that a pattern in meta-data can be recognised for collision assessment. Machine learning algorithms can be trained to recognise these pattern in data (James, Witten, Hastie, & Tibshirani, 2013). In addition, a machine learning algorithm is chosen while it can cope well with small differences in data without losing consistency in its decision making. Machine learning solutions are always a problem specific solution. This technology is applied for the first time on the collision assessment process. Therefore, it is expected to encounter bottlenecks in the processes. Furthermore, it could deem that a machine learning algorithm is not suitable while it does not fit the problem.

The aim to enhance the construction process with a machine learning algorithm is to minimise the effort of the packager in collision assessment. The algorithm should decide whether it is sure about his prediction or should learn more to reinforce the decision. The prediction accuracy of the algorithm will be enhanced by a confirmation of the prediction by the packager. It will drastically reduce the collision assessment workload for the packagers if the algorithm will be able to predict assessment accurately

The challenge is to research the machine learning algorithm in a way the optimal results will stand out. Therefore, the approach to gain the optimal output is to start small and add parameters or change settings during the process. It is anticipated that the algorithm makes mistakes in the beginning. Every change should further optimise the machine learning algorithm. It will provide insight in the functioning of the machine learning algorithm. This decision accuracy will be gained by testing and analysing the results of the algorithm.

The amount of quality data has a direct impact the performance of the system (David Camilo Corrales, 2018). The system will not be able to recognise the pattern in the data sufficiently if a small amount of data is used and will perform poorly. Furthermore, if the quality of the data is inadequate the algorithm cannot fully rely on the dataset and will also perform poorly. Thus, both quality and amount are key for a well performing cognitive system.

4.1 Solution determination

The packaging process should be closely examined to determine the best objective. The analysation should give insight in the structure of the process, the bottlenecks of the process and the key indicators for the assessment of a collision. The best solution is closely related with the input variables, preparation of the variables, algorithm problem solver, tuning and the preferred output. Here the challenge is to cope with all variables. Whether one of the variables changes the impact of other variables changes as well. It is a dynamic process. A structured approach should be applied to gain insight in the optimal distribution.

4.2 Mathematical problem

The mathematical problem for solving the objective should be determined in order to find the best machine learning algorithm. There are different problems to choose form in a machine learning algorithm. Each is suited for different use cases. It is key to know what the essence of the problem is to find the suitable problem solver. Three main mathematical problems which could be assigned are: Regression, Clustering and Classification.

During the assessment process, each collision will be assigned to either relevant, relevant 2 or not relevant. These are three classes the machine learning algorithm could choose to assign a collision to. This indicates a classification problem (Hastie, Tibshirani, & Friedman, 2008). This can be underlined by examining the other problems. The outcome of a regression problem would exist out of quantitative data. Clustering finds pattern in data by clustering the observations. Thus, the decision is made to use the classification problem for the machine learning algorithm.

4.3 Pattern recognition

The goal of machine learning algorithm is to develop a machine as if it is has knowledge (Ertel, 2016). Machine learning algorithm implies that a machine could gain knowledge and reproduce it. This would make the machine intelligent. Machine learning algorithm is a technique which is used for pattern recognition. The pattern will be 'remembered' by the algorithm. To recognise a pattern, the algorithm must first be trained.

The learning of a pattern can be trained by already available conclusions in observations (Hastie, Tibshirani, & Friedman, 2008). This is known as supervised learning. It is also possible predict the pattern by the algorithm itself. This is known as unsupervised learning. This study uses the method of supervised learning while the data is provided with the desired output.

The data set will be split to train the algorithm. One set will be used for the training and the other set will be used to compare the prediction of the algorithm with the actual observation. This will result in a prediction accuracy. The high prediction accuracy algorithm should then be able to assign a similar observation correctly. The prediction accuracy depends largely on the amount of observations. A rule of thumb is that more observations will improve pattern recognition accuracy and will open more possibilities to tune the algorithm.

4.4 Feature selection

Feature selection is a techniques used to improve algorithm performance (Pérez-Rodríguez, Haro-García, Castillo, & García-Pedrajas, 2018). Each feature or input parameter should have as much relevance to the objective as possible. The key to choose the useful parameter is to know what the data reflect. This reflection is difficult to acquire while the human brain does not reveal the key decision making features. The challenge with the selection is that different parameters have a different effect on the performance of the algorithm. Therefore, the parameter selection is based on intuition and human decision making. This makes it an educated guess.

The approach used to find the best mix of parameters is to begin which as few parameters as possible and add various parameters during the process. The mix will be analysed and compared with the previous attempt.

Furthermore, data can have causality with other variables (Chen, Hao, Cai, Zhang, & Hu, 2016). The data applied to the algorithm only included direct features. Related features are already acquired implicitly. For example: If material is chosen as input parameter, it is not useful to add the e-modulus of the material as second parameter. While the material parameter already takes into account the e-modulus.

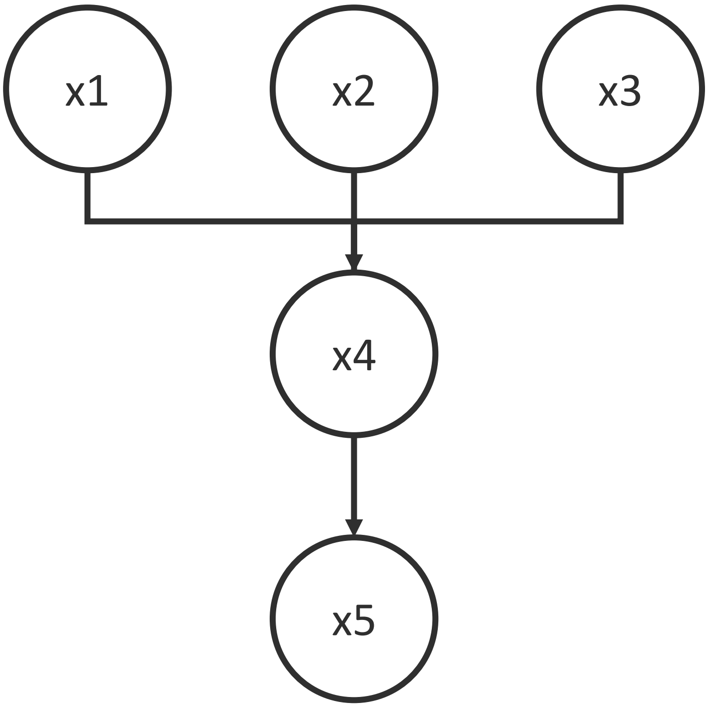


Figure 8: Parameter causality

5 Part II: Combining algorithm with business process

This chapter will provide a deeper understanding of the collision process in combination with the machine learning algorithm. First, the origin and pre-processing of the data will be described. Then the parameter preparation is elaborated including the different parameters used and its technique to make it usable for AI. This process makes the data usable for data analytics. Data analytics is used to gain insight in the data. Then the Pipeline and the cross-validation technique will be outlined. These techniques accelerate the process and ensure the accuracy of the algorithm. All instruments described will be used in the hypothesis verification.

The data extraction and programming of the machine learning algorithm is done in the programming language Python. This language is most suitable for data preparation, analysis and programming machine learning algorithms. This is done using the package Scikit-learn. This package is suitable for the programming of the machine learning algorithm. Packages are additional programming instruments which can be used in the programming language. This package is chosen due to its functionality and quickly generation of results. The downside is that the algorithm provides restricted control of the machine learning algorithm.

5.1 Data preparation

Two available data sets were used. A data set of a passenger car and a Van data set. Both data sets differ from each other in amount, collision status, expert knowhow, depth of collisions and materials. There is chosen to use these two datasets in order to compare the outcome. This will gain insight in the functioning of a machine learning algorithm in combination with the packaging process.

Both data sets were available in different SQLite databases which were identically structured. Therefore, Both SQLite databases must be pre-processing by a specific algorithm to be usable for the algorithm. This meant data could not be extracted directly.

The other possibility was to extract the data directly out of the PDM system. The PDM system would show an error if a mistakes would occur in the algorithm. This made the extraction out of the PDM system more reliable. The disadvantage was that only the passenger car data could be acquired.

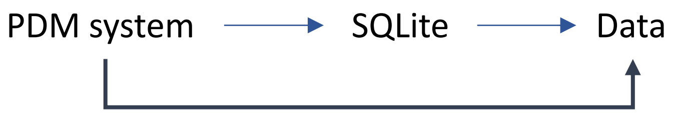


Figure 9: Overview data acquisition

5.2 Machine learning algorithms

The design of a machine learning algorithm begins with the determination of the objective. There was chosen to focus on the detecting of not relevant collision with high certainty. The packager can rely on the detected not relevant collisions. It is uncertain if the remaining collisions contain only relevant collisions. Here the packager reduces time spend on not relevant collisions. The objective was determined in discussion with IT and collision experts.

The opposite option would be to have a high certainty all relevant collisions would be detected. This approach was rejected while the packager would have to invest time in the remaining not relevant collisions. These would include relevant collisions because the algorithm would not be sure where to address the collision. Hence, the packager would go through both lists of relevant and not relevant collisions.

5.2.1 Artificial Neural Network

Artificial neural network (ANN) is convenient as a forecasting technique (Jiang, Chin, & Tsui, 2018). One of the different ANN models is the multilayer perceptron (MLP). MLP is a neural network which exist out of hidden layers and nodes and edges and is suited for pattern recognition (Yoon & Kum, 2016). MLP model is chosen as the machine learning algorithm.

MLP is a supervised learning algorithm (Yoon & Kum, 2016). Supervised learning requires pre-defined output value. These are also referred to as labels. The algorithm will train itself by optimising the network in order provide the desired output value. This process should be applied often to acquire an accurate algorithm. Therefore, the amount of data is crucial to train the neural network sufficiently. It is recommended to include 10.000 observations during training in order to obtain a well performing neural network.

The edge functions as an input for their attached node (Ertel, 2016). Each edge possesses a specific weight value in a fully trained network. The input value of the node is the summation of the weight values multiplied by the according output value. The node will apply this value on a formula. The model will provide the resulted value to the next edge if the output of the formula exceeds the threshold. The weight value is represented as w_{ij} and the input value of the edge is x_j .

$$x_i = f(\sum_{j=1}^{n} w_{ij}x_j)$$

The MLP can be configured in any form. The configuration settings of the algorithm, which were altered during the research, were: Amount of hidden layers and amount of nodes. Other configuration settings were a default value. Figure 10 presents a neural network.

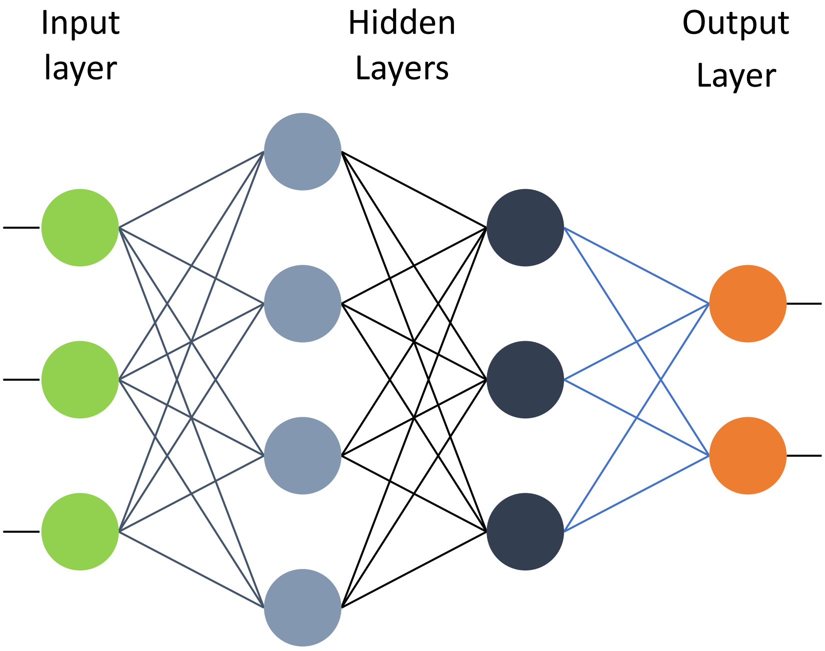


Figure 10: Overview artificial neural network

5.2.2 Random Forest

Random Forest is another approach and is making use of decision nodes, also known as trees (James, Witten, Hastie, & Tibshirani, 2013). A random calculation method for the node is chosen each time a split occurs. The algorithm is not permitted to assign the calculation method. These nodes make a decision which is either true or false. This ensures the tree will function without taking any consideration of a strong feature influence. Random Forest is a simple method and suits small data sets.

Random Forest will provide the final decision by a majority vote when using the classification method (Hastie, Tibshirani, & Friedman, 2008). Each tree in the forest will provide a decision. Random Forest will provide the final output by a majority vote of all trees. A simplified representation of a forest existing out of two trees is presented in figure 11.

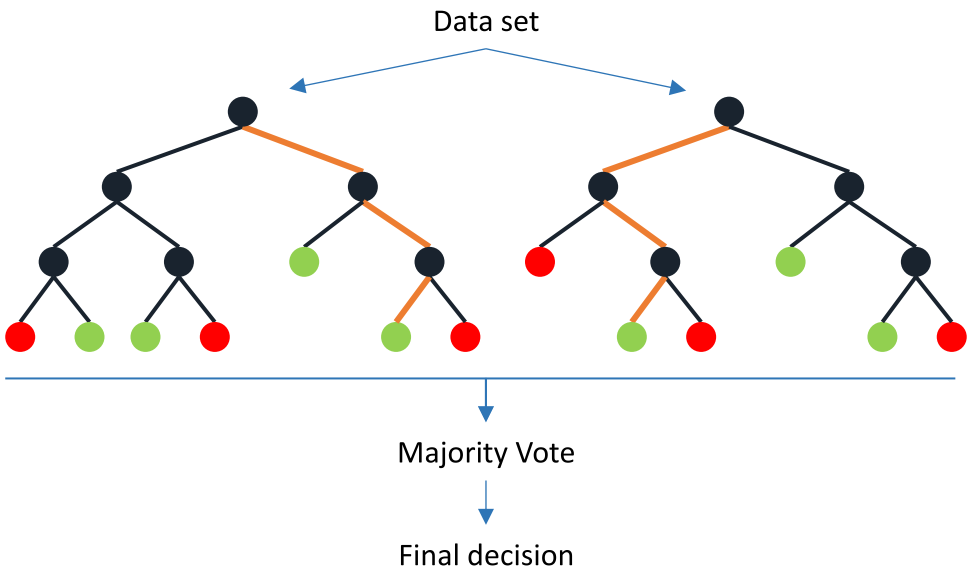


Figure 11: Simplified representation of Random Forest

5.3 Parameter selection

The parameters or feature selection is related to the determined objective. However, the parameter selection cannot be based prior to the training of the algorithm. The chosen parameters are based on human decision making and are derived from expert knowhow and intuition. In addition, the amount of parameters selected is dependent on the amount of available data. The machine learning algorithm can handle more input parameters with more data because the algorithm can optimise itself more to find the best pattern. Each input parameter provided to the machine learning algorithm can only consist out of numbers. At the moment MLP cannot understand letters and text. All text input for the MLP algorithm will be transformed into numbers.

A data observation exist out of Hausdorff distance, material and status of a collision. The Hausdorff distance already exist out of a number. Especially, material and status of collision are categorised in different groups with a corresponding number. The categorisation of the status of a collision exist out of two variables. It is either a relevant collision, corresponding with 1, or a not relevant collision, corresponding with 0. The categorisation of materials is derived from the frequency of the material name in the material list. A high frequency was assumed to be of more importance. Data analysis would reveal if this approach was effective. The categories could be redesigned if the approach proved to be ineffective. The exact material categorisation is elaborated in 6.2.2 Parameters.

5.4 Data analysation

The data should be analysed before providing the observation to the algorithm. Data analysation could provide the insight that the observations do not provide assistance to the intended objective. It must be proven that the data is usable. It provide an enlarged insight in the data. It could lead to redesign of the categorisation of the input parameters which cannot be detected immediate. Furthermore, the data analysis could be linked to the origin of the data. This would provide understanding in business processes. The results of the algorithm could be underpinned by the insight provided by data analysation.

5.5 Process optimisation

The input data is normalised and scaled after the data analysis is completed and the data is considered convenient for the machine learning algorithm. This means all observations have a value between 0 and 1. The observations are weighted equally to make sure the ANN cannot train on a preferred parameter. Though, the influence of each parameter can differ and should be tested whether the parameter is of any impact on the performance of the algorithm.

Furthermore, it is of crucial importance the data is randomised in order to train the train. This should prevent the algorithm from enhancement of recognising only specific cases. The downside of this method is that the output of the AI is also randomised and scaled. Therefore, analysing the observation after the training is harder to analyse. A transcription method should be in place.

In order to acquire the best configuration the pipeline method is executed. Configuring a neural network is more an art than academically confirmed (Erik Bochinski, 2017). The pipeline method is a method to verify the results of multiple configuration. The pipeline method will be used in combination with cross-validation. This is a method which analyses if the performance of the machine learning algorithm on overfitting (James, Witten, Hastie, & Tibshirani, 2013). This phenomenon occurs when the algorithm is unable to predict unseen observation correctly. In addition, the algorithm is trained to precisely on flaws in the data. Hence, the pipeline outcome predicts which configuration is would to make the best performance.

6 Hypothesis verification

It is expected that during the study the prediction accuracy of the algorithm will increase. The decision was made to initiate with a small amount of input parameters. This is due to the lack of orientation in the algorithm performance. It is the intention to increase input parameters and algorithm effectiveness during each iteration.

6.1 Iteration framework

Each iteration is conducted following the framework shown in figure 12. The frameworks shows a clear relation of the input data with the final algorithm. Especially, the quality and amount of data determines the quality of the final algorithm. In the framework the steps in between the input data and final algorithm cannot enhance the quality nor amount of data.

Data cleaning is conducted to gain insight in the data quality. Data quality is the extent of data to meet the specified requirements (David Camilo Corrales, 2018). Quality requirements consist out of amount of missing, duplicate or outlier values. This quality of the data is projected by the processes of data cleaning and data analysis. This data quality process is shown in figure 12.

Several adaptation are made to the data cleaning process framework (David Camilo Corrales, 2018). Missing values in data are deleted while the imputation of data will have an unknown effect on the collision. This is not desirable in order to guarantee data quality. Then duplicate values are removed and afterwards the dimensionality of the data is reduced. Dimensionality reduction makes sure the data can be interpreted by the algorithm (David Camilo Corrales, 2018). Outlies in observations is considered a result form the data analysis.

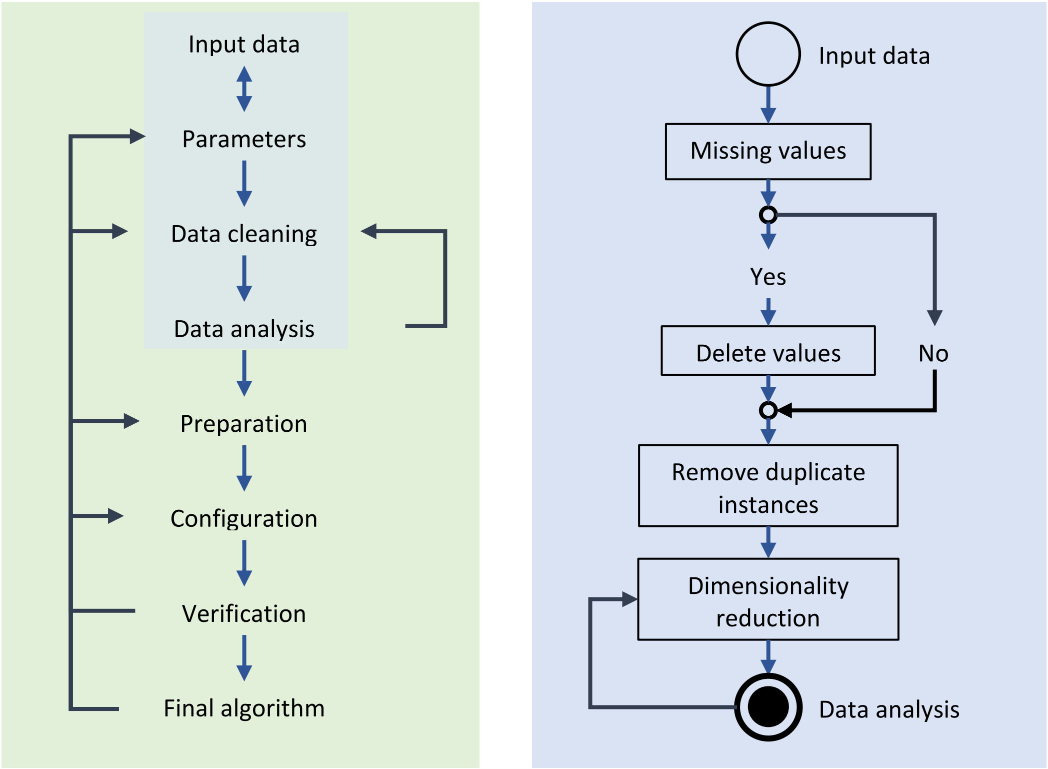


Figure 12: Overview of iteration framework with adapted Data cleaning process (David Camilo Corrales, 2018)

Each iteration is a dynamic process due to the many loops in the framework. This is further enhanced while within each step there are possibilities to vary. It is not known what the optimal performance of the variable settings is because of the many configurations. Each iteration should point in a direction which parameters and settings have a positive effect. Another configuration could potentially be better than the optimal found configuration.

6.2 Iteration 1

Iteration 1 aimed to determine to gain insight in the possibility of apply a machine learning algorithm. It would also provide the advantage to expose bottlenecks in the observations or the collision process. This iteration will provide the benchmark which could be optimised in further iterations. Passenger vehicle and Van observations are extracted out of a SQLite database. A small amount of parameters is selected to train the machine learning algorithm.

6.2.1 Data set 1

The passenger car and Van dataset were used during this iteration. Both data set existed out of several SQLite databases. The SQLite database proved to be the first bottleneck. Each SQLite database included more produced metadata as described in 3.4 Data origin. It is estimated that this will influence the input data set. The SQLite data was pre-processed in order to acquire correct observations. This proved to be challenging due to a complicated systematic in the database.

6.2.2 Parameters

All input parameters were available in the dataset but had to be processed. The label is the collision status, relevant or not relevant. This is described in 5.2.1. Artificial Neural Network

The relevant 2 collisions are added to the relevant collisions because their amount was insignificant. This iteration uses the material and Hausdorff distance as input parameters. The Hausdorff distance is a distance and is directly usable while it is expressed in a value. The other parameter is the material.

Material categorisation

The original data contained a list of material names. These names should be transcribed while text cannot be interpreted as input. After examination of the material list, it was concluded that the list included far more materials then expected while material nuances and material demands were implemented in the name. It was assumed that the most occurring materials would give an indication of the importance of the material. The nuances and material demands were eliminated from the name while this was information which considered to precise. Relevant for collision calculation is the solidity and compressibility of the material itself. The list of material is reduced in order to give the algorithm the possibility to increase informative value of a smaller list of possibilities. The decision was made to group the material list on the frequency in occurrence and of the basic material element. The list contains 21 categories in order to be flexible to shrink the list in following iterations and is presented in table 2.

0	Remaining	11	Cast-iron
1	Aluminium	12	Glue
2	Composite	13	Leather
3	Copper alloy	14	Foam material
4	Copper-zinc alloy	15	Lubrication fluid
5	Sealing material	16	Steel
6	Thermoset plastics	17	Textile
7	Elastomer	18	Thermoplastic
8	Electric component	19	TPL Elastomer
9	Foam	20	Composite material
10	Glass	21	Zinc-alloy

Table 2: Material categorisation

6.2.3 Data analysis

The data analysis is used to acquire insight in both data sets. It is aimed have a clear understanding of the observations within the data set. It is also used to compare the differences in the data sets. Data analysis will provide the demographics of the Hausdorff distance and the material categorisation. It will also expose possible unbalance in the data. Furthermore, the insight could be used to explain the training performance afterwards. There is looked at the distribution of amount of data, status of collisions, Hausdorff distance and materials used in collisions.

Figure 13 shows the total amount of data with the distribution of relevant and not relevant collision statuses. The amount of data is well below the recommended amount of 10.000 observations. This indicates that the training quality of the neural network will not be optimal. Furthermore, the distribution on both sets is different. The Van data hold much not relevant data with only a few relevant collisions. This is assumed to be of impact for the machine learning algorithm while it can only learn minimal exceptions. The data in the passenger vehicle data is more evenly distributed.

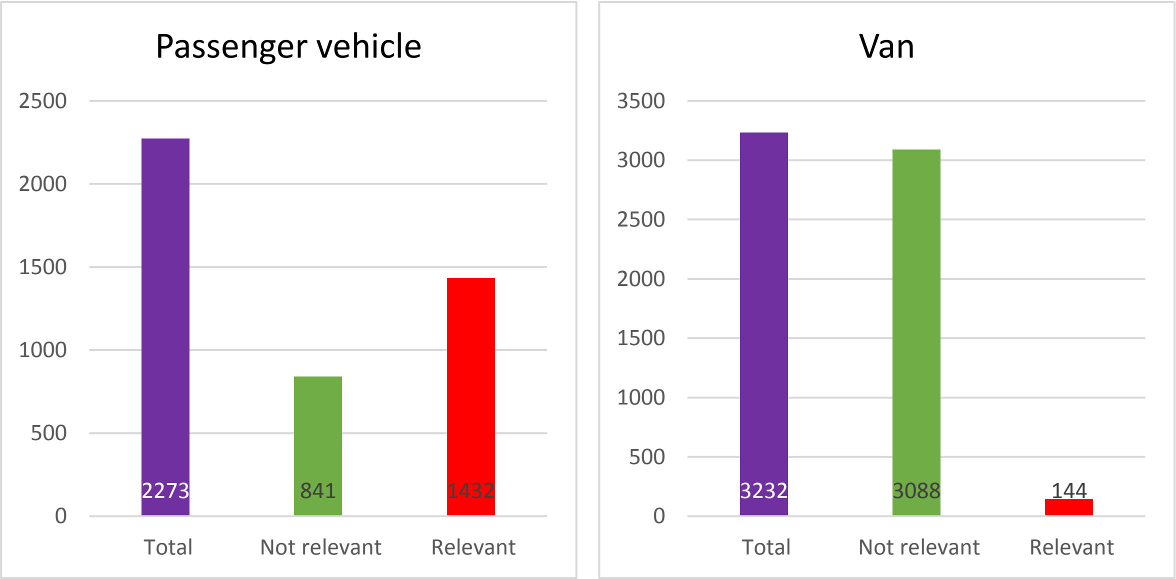


Figure 13: Passenger vehicle (left) and Van (right) collision status distribution

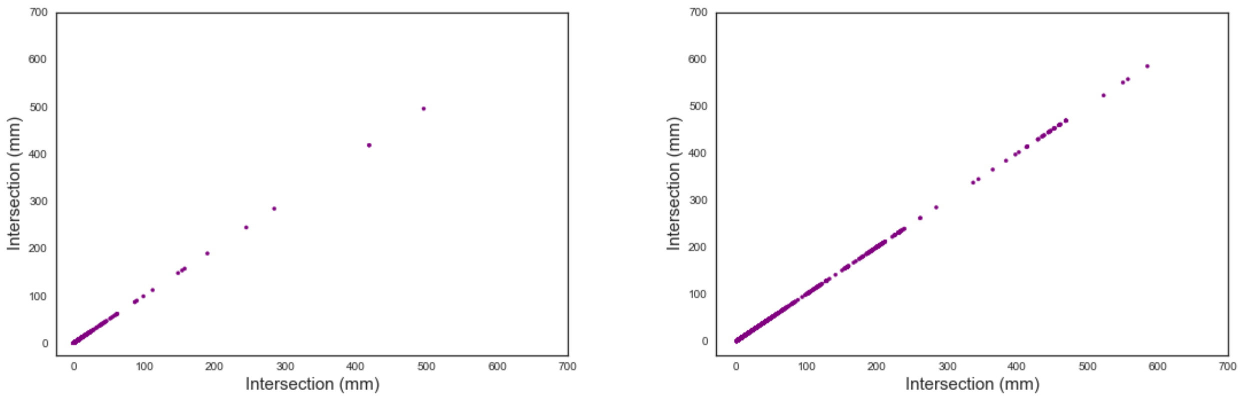


Figure 14: All observations (left: passenger vehicle, right: Van)

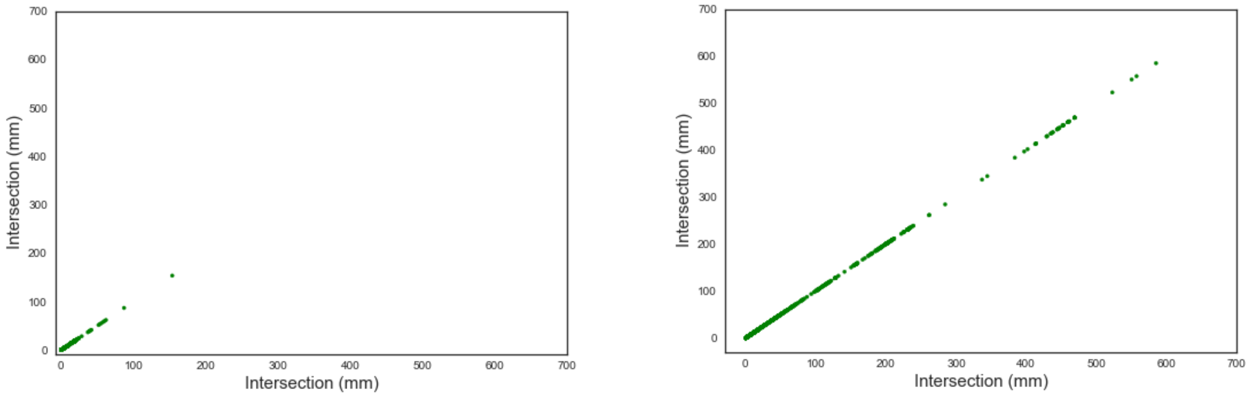


Figure 15: Not relevant Hausdorff observations (left: passenger vehicle, right: Van)

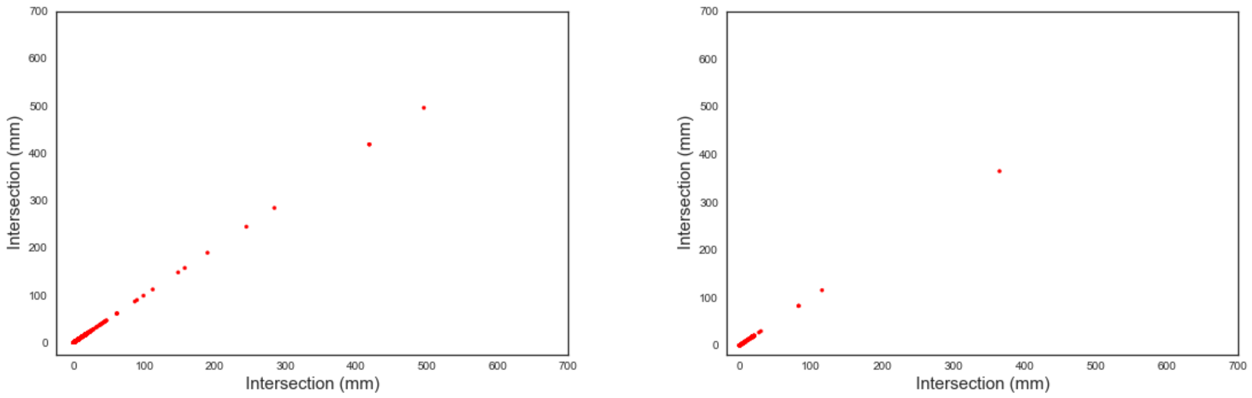


Figure 16. Relevant Hausdorff observations (left: passenger vehicle, right: Van)

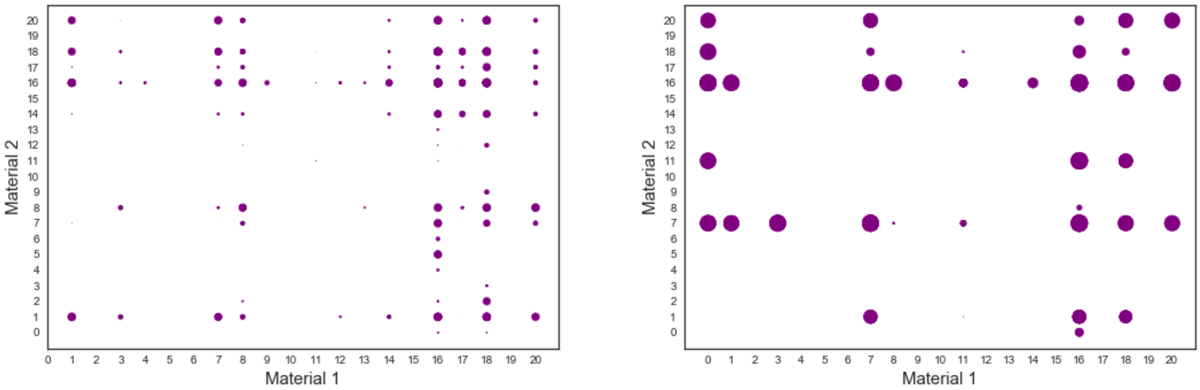


Figure 17: All collision combinations (left: passenger vehicle, right: Van)

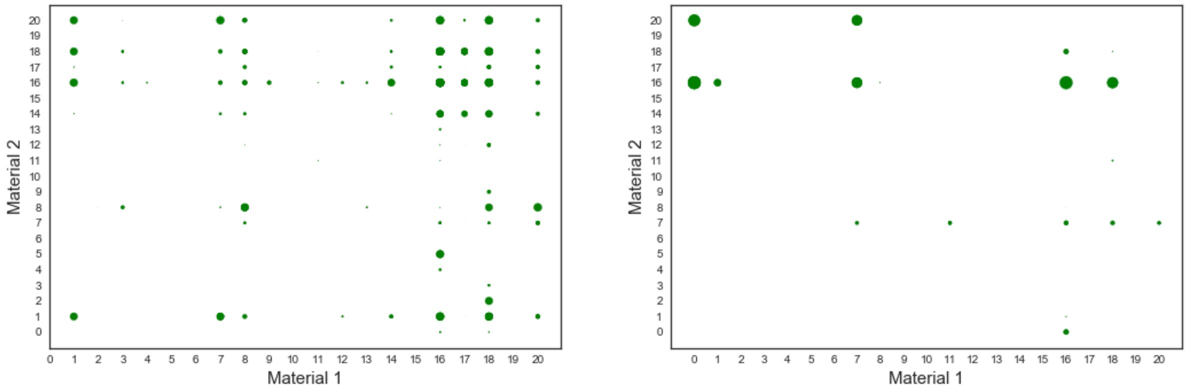


Figure 18: Not relevant collision combinations (left: passenger vehicle, right: Van)

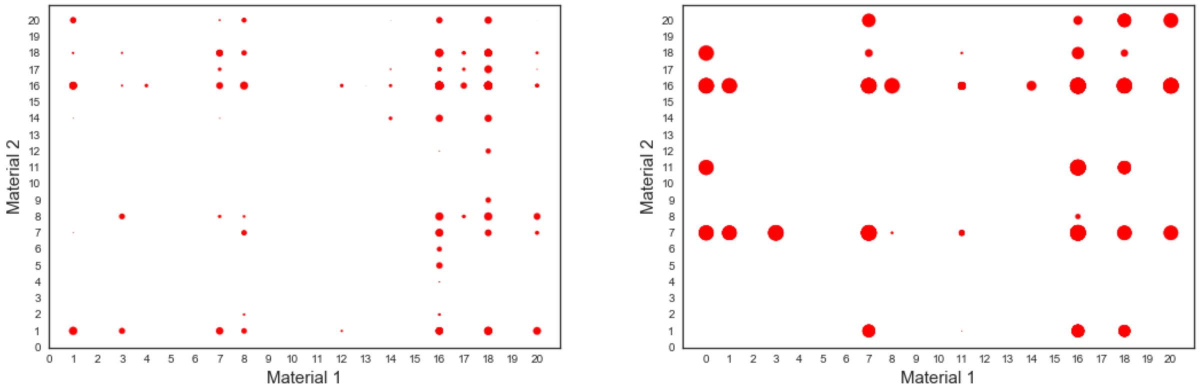


Figure 19: Relevant collision combinations (left: passenger vehicle, right: Van)

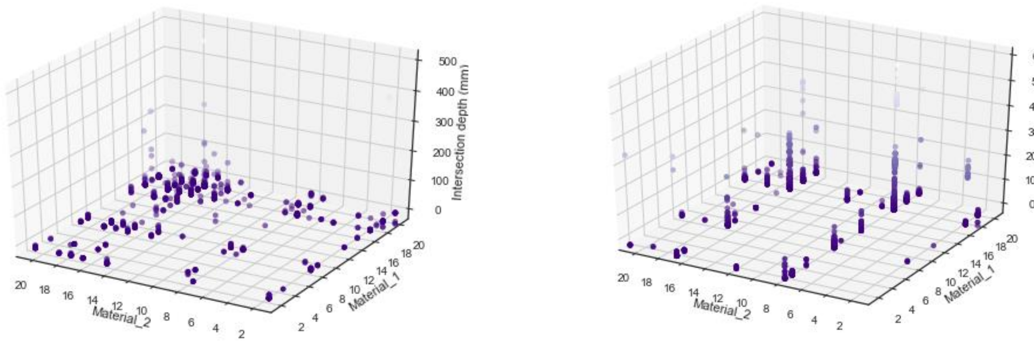


Figure 20: All observations combined analysis overview (left: passenger vehicle, right: Van)

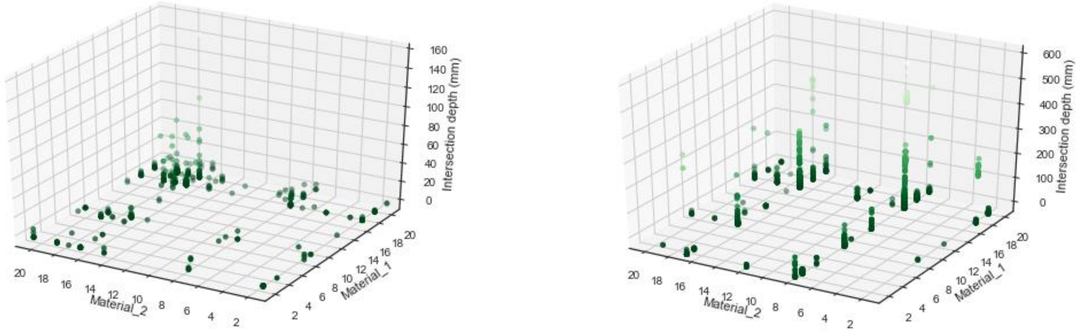


Figure 21: Not relevant observations combined analysis overview (left: passenger vehicle, right: Van)

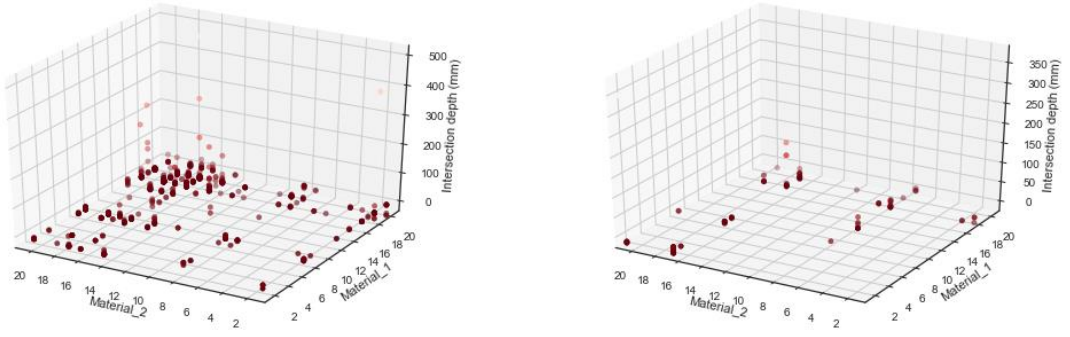


Figure 22: Relevant observations combined analysis overview (left: passenger vehicle, right: Van)

The distribution of collision statuses can be recognised in the Hausdorff distance data shown in figures 14, 15, 16. It is perceived as uncommon that a Hausdorff distances would be higher than 300 mm. Every collision above 300 mm was expected to be a relevant collision. The not relevant Van collisions shows that this assumption could be false. Furthermore, the not relevant collisions in the passenger vehicles are clustered in the lower range of the overall distances.

The material combination of collisions is explored in figure 17, 18 and 19. Each material is transcribed to the material group elaborated 6.2.2 Parameters There is concluded that the passenger vehicle holds more materials than the Van vehicle. This could be due to the high amount of different luxurious parts used in the interiors of a passenger vehicles. Another insight is the occurrence of materials. Some materials do not or rarely occur such as numbers 4 and 5, respectively copper zinc alloy and sealing material. Especially, sealing material was expected to occur while for example rubber used in door portals should have a collision. Furthermore, the higher materials numbers such as 16 and 18, respectively steel and thermoplastic occur often. These materials are a key material in car products.

The material combination grid can be combined with the Hausdorff distance. This resulted in the 3D images in figures 20, 21, 22. Each point is a collision and the brightness indicates the depth of the collision. It gives an indication of the distribution of the collision status with the relation of Hausdorff distance.

6.2.4 Algorithm configuration

Choosing the machine learning configuration settings was complicated due to the lack of a reference point. Furthermore, data analysis showed the diversity in the data sets. The decision was made to configure both nodes and layers in the algorithms differently. The different configurations where chosen in order to obtain an insight in the reaction of the neural network. The activation function was held the same. The following configurations were used:

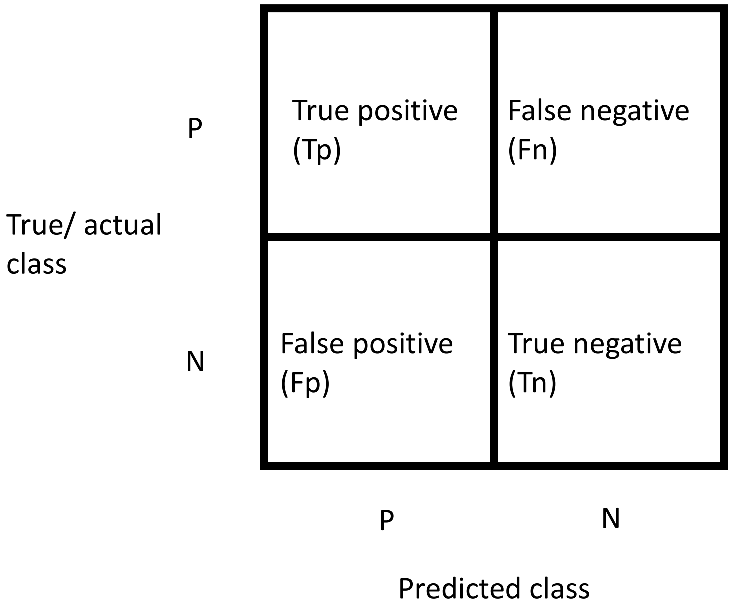
Passenger vehicle:	
Activation function:	ReLu
Layers:	10
Nodes in layers:	60,600,600,600,600,600,600,600,600,4
Test size:	25%
Van:	
Activation function:	Relu
Layers:	8
Nodes in layers:	80,600,600,600,600,600,600,4
Test size:	25%

Table 3: algorithm configuration of configuraiton 1

6.2.5 Results

The key indicators of the results can be derived from the confusion matrix. This matrix represents the comparison between the outputs of the algorithm and the actual/true label. This is shown in figure 23. The quadrant shown the correct predictions as True and the wrong predicted as False. This quadrant is useful to gain insight in the accuracy of the algorithm. Furthermore, different analysis can be conducted such as the percentage of true predicted labels. The confusion matrix in further reading only presents the outcome of the test data.

Figure 24 shows the total distribution of the Van data with its according confusion matrix. It can be concluded that the result is not adding value. The total test data included 42 relevant collisions and 766 not relevant collisions. The algorithm decided correctly with an overall accuracy of 95% and a not relevant accuracy of 96%. However, this high accuracy is not representable for a functional algorithm while the distribution of the total input data showed an extreme high amount of not relevant collisions. The Van algorithm is trained to estimate all collisions as not relevant. The algorithm would not be able to assign other observation correctly. This is because the algorithm did not receive any anomalies nor sufficient relevant observations from the Van data set. This is negative because the packager should consider the not relevant collisions as well.



		Predicted class	
		P	N
True/ actual class	P	True positive (Tp)	False negative (Fn)
	N	False positive (Fp)	True negative (Tn)

Figure 23: Overview of configuration matrix

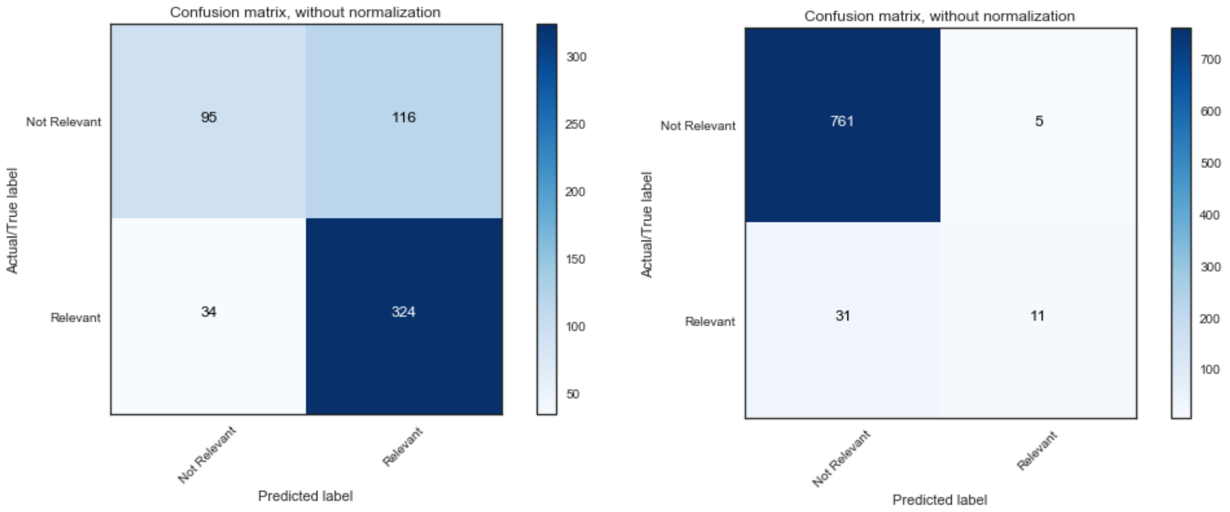


Figure 24: Configuration matrix result, iteration 1 (left: Passenger vehicle, right: Van)

The passenger vehicle data included 358 relevant collisions and 211 not relevant collisions. This distribution is better to train the algorithm with. The passenger vehicle algorithm was able to recognise relevant and not relevant collisions. This is a contrast to the Van algorithm. The passenger vehicle performed better due to the fact that the data is a case study collision job and all collisions are assessed by a guideline.

At first sight the passenger vehicle confusion matrix looks promising while it recognised 74% correctly. Though, after analysing the matrix more precisely, the algorithm failed to correctly evaluate 10% for relevant collisions and 45% of not relevant collisions. This percentage is too high for a packager to rely on while it means only 55% of all not relevant evaluated collisions are correctly assigned. Furthermore, a larger amount is assigned as relevant collisions. It is estimated that this is due to the larger amount of relevant collisions in the test data.

Both neural network were not functional examining the relation with the confusion matrixes and the configuration of the neural network. The amount of hidden layers and nodes were considered to extensive in comparison with the used data size and parameter amount. It is estimated that a smaller neural network could provide the same results. Furthermore, based on expert best practices the neural network should first expand in smaller steps in order to capture nuances in the data.

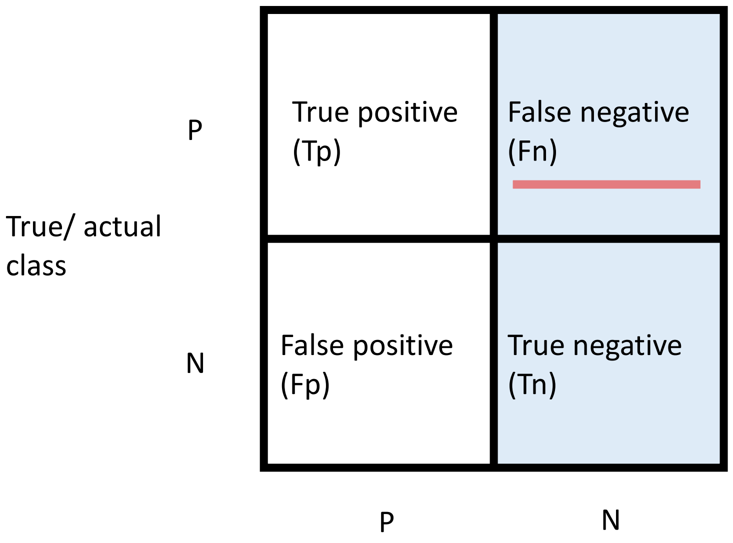
6.2.6 Verification

The first indication of the passenger vehicle results are promising. The overall correct assignment of 74% were considered as a positive sign that a machine learning algorithm can be applied on the process. Further improvements could enhance this percentage. Key to this improvement is the amount and quality of available data. The data was considered to present an inaccurate representation of the collisions due to the complicated processing of the SQLite database. Therefore, there is chosen to conduct further research with data directly out of the PDM system. This data would not be processed which would increase the soundness of the data set.

Another optimisation could be the design of the algorithm configuration. This could be enhanced by making the neural network smaller. It is estimated that the algorithm can achieve the same results with less resources. This results in a smaller neural network with smaller steps building up the network.

Additional insight data analytics provided enhanced understanding in the collision process. This showed that there is no leading part in a collisions. This impacts the analysation of materials while materials in both parts are equally valued. In further iteration the material order will change to a combination of material with the lowest material number in front. Furthermore, there will be looked if the material categories could be reduced while some categories were considered empty and therefore inadequate. It is assumed this would increase the ability of the algorithm to take the material of both collision partners into account.

The confusion matrix can be used as tool which data would be presented to packagers. The blue quadrants represents all collisions which will be assigned as relevant by the algorithm and will be shown as relevant collisions to a packager. The aim is to optimise the detection of not relevant collisions with high certainty. It is not stated which standpoint should be taken to optimise the not relevant collisions. Logically, the actual label of not relevant collision is aimed to optimise. Hence, the false positive quadrant is aimed to be optimised and is presented with red underling. Here all not relevant collisions are assigned as relevant.



		P	N
True/ actual class	P	True positive (Tp)	False negative (Fn)
	N	False positive (Fp)	True negative (Tn)

Figure 25: Optimisation configuration matrix overview (P= not relevant, N= relevant)

6.3 Iteration 2

This iteration focused on increasing the ability to interpret the data by the algorithm and increasing the prediction accuracy of the algorithm. The key changes in this iteration is the input data, adding a new material combination and configuration to the machine learning algorithm. There is chosen to use alternative data which was extracted directly out of the PDM system. This method of data acquisition improve the quality of the data, since it avoids errors due to wrong identification of the data. These collisions were assessed by a guideline which should provide consistency and reliability throughout the data. Data analysation should gain insight if this assumption is correct.

6.3.1 Data set

The data set consisted of original data from of the PDM system. The data set only contained passenger vehicle data and contained more observation than iteration 1. The data could only be compared with itself while Van data was not available. Three test were conducted which were aimed to enhance the data and algorithm.

6.3.2 Parameter

This iteration focussed on the optimisation of the material parameter. It was chosen to apply a new processing method due to the restricted amount of available data and extended insights in the collision process. The new implemented combination of materials is shown in table 4. The materials which are included in the new categorisation are stated in appendix B. Iteration 1 represents a collision with steel on thermoplastic and the second collision represents thermoplastic on steel. In the first iteration the two example collisions could be perceived as two separate collisions with no relation. New packaging process insight proved this assumption wrong. Iteration 2 structured the lowest material number first. Now both example collisions can be perceived as the same collision while the material combination is the same. This should provide a more structured presentation of the material categorisation and improve the understanding of the material input for the algorithm.

Another optimisation was the further reduction of material categories to eight most commonly used materials. Each category includes a wider range of materials and reduces the categories which failed to include any collision. It is estimated the alteration would impact the weighing between material categorisation and Hausdorff distance.

Collision	Iteration 1		Iteration 2	
	Part 1	Part 2	Part 1	Part 2
1	16	18	16	18
2	18	16	16	18

Table 4: Represented material combination

1	Metal
2	Plastic
3	Sealing material
4	Electronics
5	Glass
6	Fluid
7	Leather
8	Wood
9	Remaining

Table 5: New material categorisation

6.3.3 Data analysis

The data analysis started with determining the distribution of relevant and not relevant collisions. The data reduced drastically by removing missing values and duplicate observations. This is considered curious while it implies that multiple collisions appear multiple times which should not occur. Then the data is further processed in the material categorisation of iteration 1 and the new material categorisation. The amount of collisions did not correspond between the original and new material list. This is again a strong signal that collisions occur multiple times with different material combinations and equal Hausdorff distances.

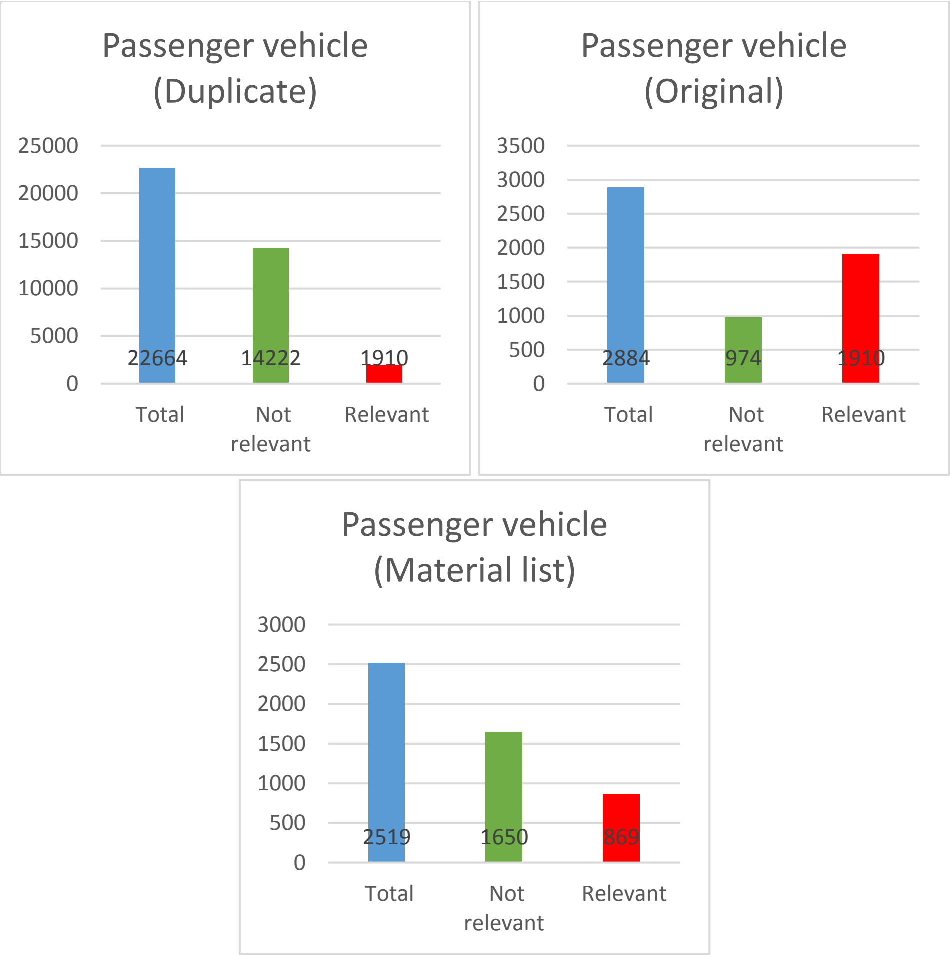


Figure 26: Collision status distribution, iteration 2

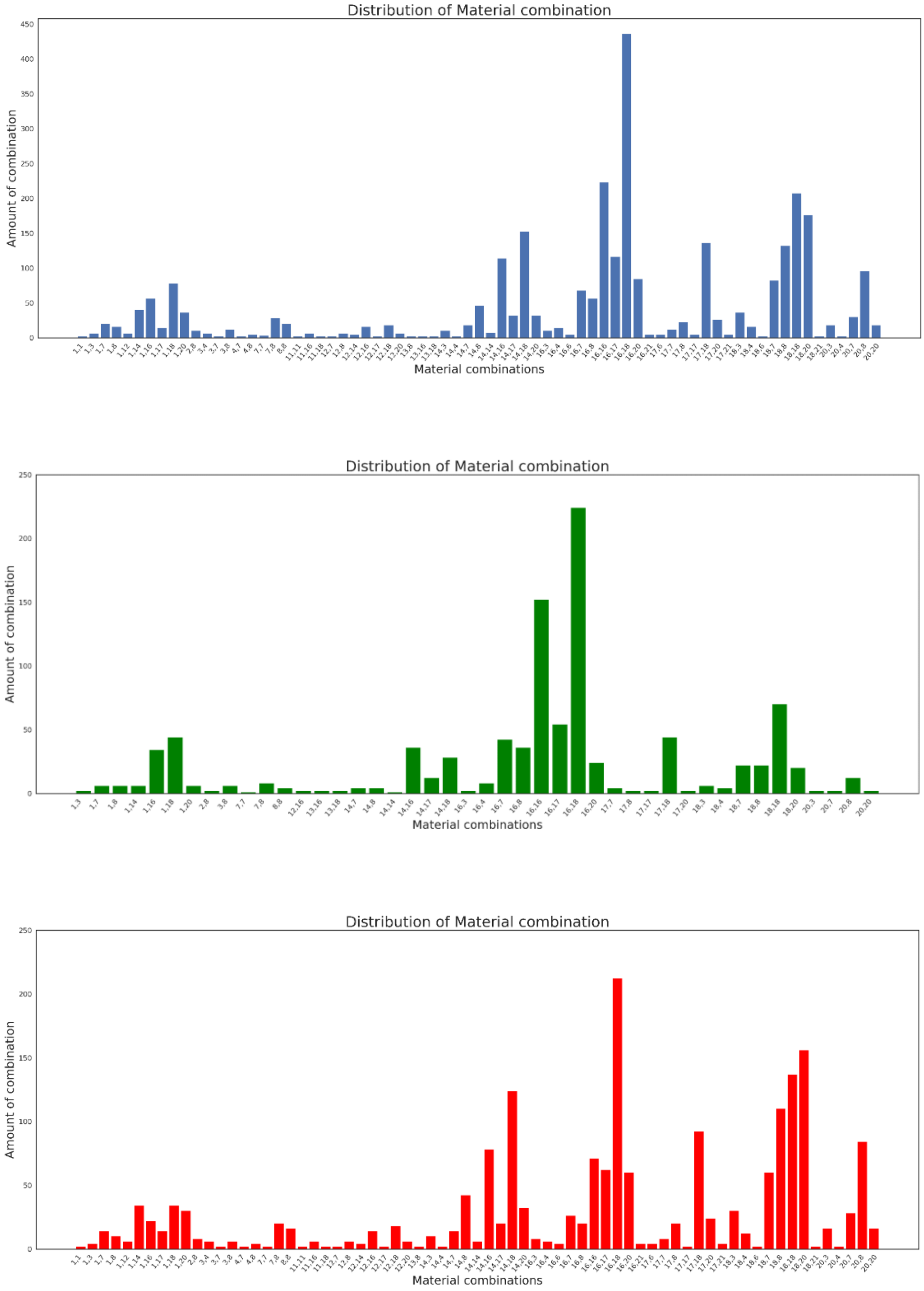


Figure 27: Material combination iteration 2 (top: all observations, middle: not relevant observations, bottom: relevant observations)

The distribution of the material combinations with the original material categorisation provides a better overview of the material combinations in collisions. A larger figure size of figures 27 and 28 can be found in appendix A. Again it shows that collisions with steel, thermoplastic occur often. This trend is also occurring in the distribution of the new material categorisation. It should be taken into account that the total amount of unique collisions could be less with the potential of duplicates in the data set.

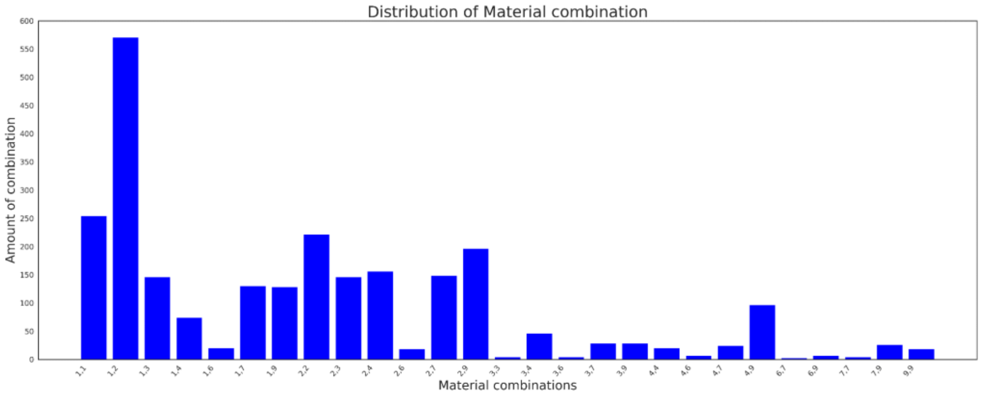


Figure 28: All observations, new material categorisation

6.3.4 Algorithm configuration

The configuration of the machine learning algorithm was the same for all three variants. This is due to the strong signal that collisions could occur multiple times and could impact further researched. It is expected to influence the algorithm negatively while the multiple occurring collisions will be rooted deeper into the algorithm. This will provide false information in the configuration matrix. The collision could be arranged by the algorithm with 100% accuracy while the equal collisions is already rooted in the neural network. This makes that the confusion matrix will provide a falsified positive output.

Passenger vehicle	Duplicates, Original and new categorisation:
Activation function:	ReLu
Layers:	3
Nodes in layers:	80,350,1000
Test size:	25%

Table 6: Configuration iteration 2

6.3.5 Results

An equal trend in all three confusion matrixes is noticeable. This trend is already noticeable when looked at the colour schemes. Looking at the numbers of the trend, the algorithms performed worse than previous iterations. The accuracy of the new materials is 75%. This is only 1% better than the previous iteration. Though, the algorithm failed to correctly assign 20% of the relevant collisions.

It is perceived unusual that the outcome of the algorithm perform equally. It is unlikely to occur while the input data is distributed differently and the algorithm is configured similar. As described it is estimated that the input data includes duplicates. The performance of the algorithm could be explained in combination with the neural network. The neural network could be trained on many duplicate data with an image in the test data. This would root the specific collisions deeply into the network. It is estimated that the configuration of the neural network would not make any difference to the performance as long as the data included sufficient duplicates. Any further iteration with new parameters should not have any influence on the predictability of the algorithm itself. This would not result in a successful algorithm. The impact severity of duplicate data on the performance of the algorithm should be researched.

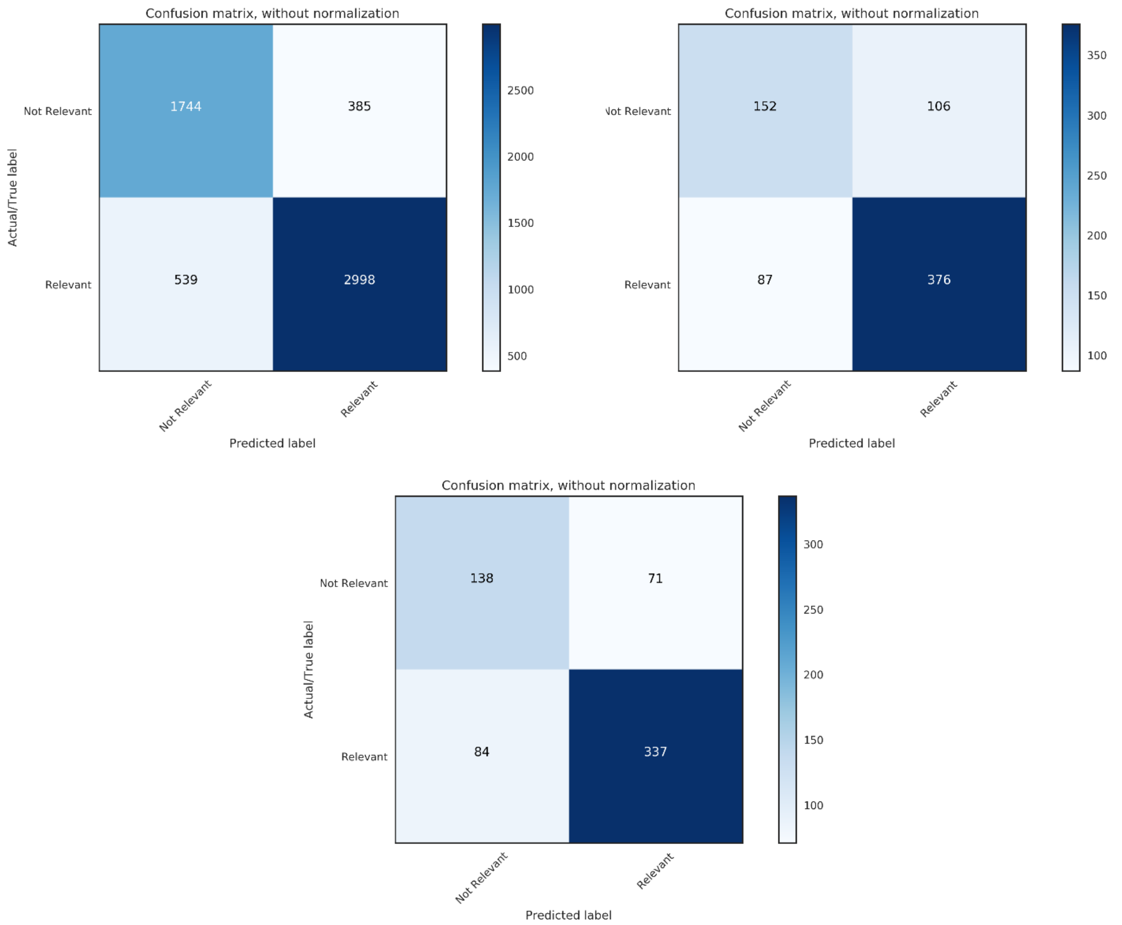


Figure 29: Results configuration matrix (top left: all observations passenger vehicle, top right: passenger vehicle old material configuration, bottom: passenger vehicle new material combination)

6.3.6 Verification

Extracting directly out of the PDM system should give the advantage that original data is correctly extracted. The collisions are changed of status following a guideline. It was assumed this should provide consistency and reliability throughout the data. Though, choosing to acquire the data directly out of the PDM system proved to be a misconception. Duplicate data is estimated to be provided in the dataset instead of a unique data set.

Another conclusion is that the data in combination with the algorithm is at its maximum capability. The results did not improve in comparison with the first iteration. Though, in order to add more parameters the amount of data should be increased.

The relation with the quality input data and the performance of the algorithm is underlined in this iteration. The amount of data is not the only element in designing a well performing algorithm. The results could show the impact of an impure data set. This should be researched further. It is estimated that using the same data set with another configuration should not enhance the outcome of the algorithm. Each slightly different configuration should result in the same performance.

6.4 Iteration 3

The consideration that the data includes duplicates or that the algorithm performance is at its optimum is elaborated in this iteration. The data is further explored in MLP and Random Forest which are described in chapter 5.2 Machine learning algorithms. Random Forest is applied while the data set is considered too small to train a neural network. Additionally, the pipeline principle is applied on both algorithm to accelerate insight in the best configuration.

6.4.1 Data set

The decision was made to only use the passenger vehicle data set with the additional new categorisation due to the equal performance in previous results. The outcome will be compared with the output of iteration 2. It is estimated to provide the same output trend as in iteration 2. The data analysis and parameters are described in 6.3.2 Parameter and 6.3.3. Data analysis.

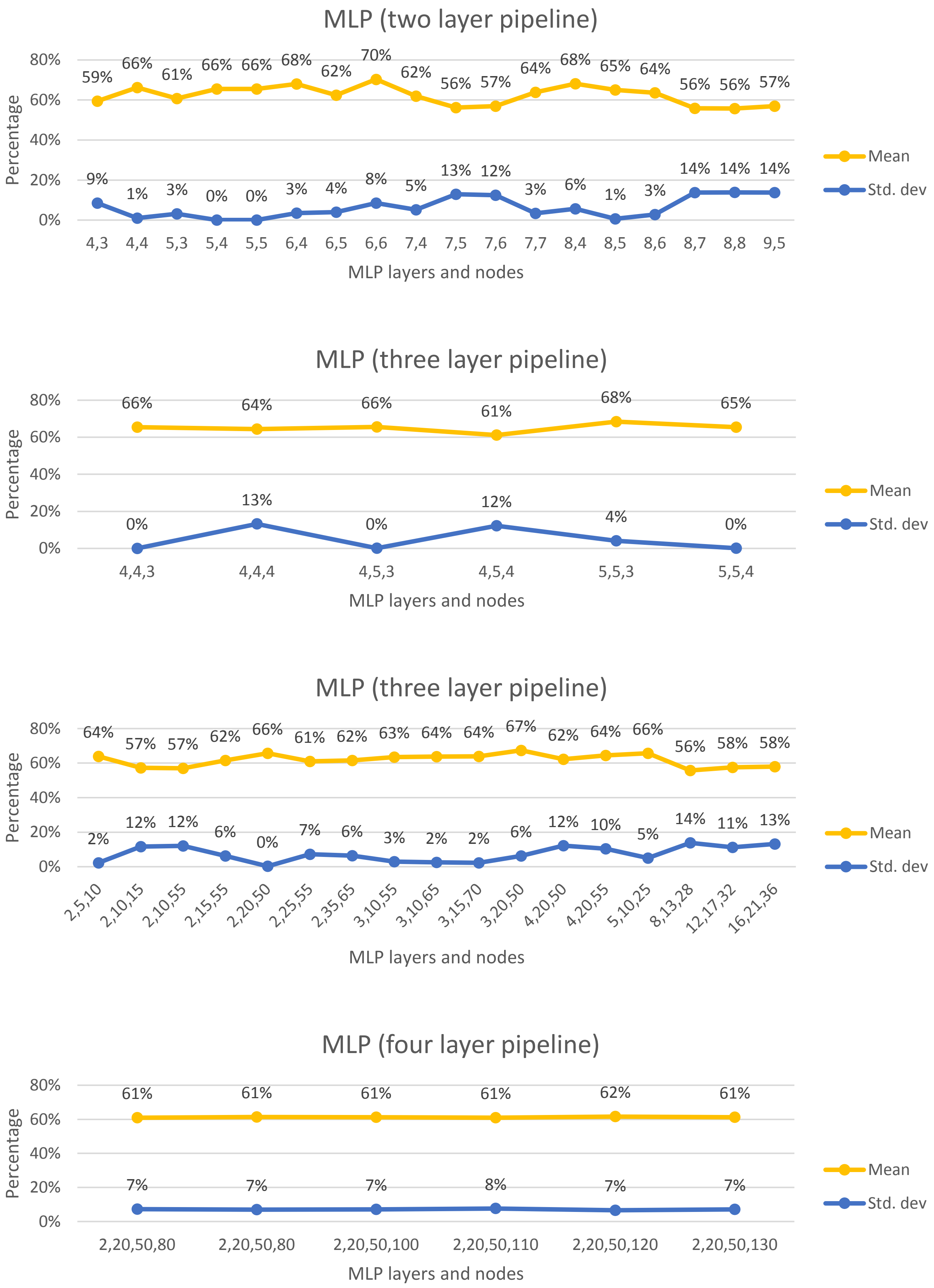
6.4.2 Algorithm exploration

The aim of the algorithm exploration is to acquire the optimal algorithm configuration. The exploration is divided into two machine learning algorithms: MLP and Random Forest. A technique called Pipeline is used in order to accelerate the configuration process. The pipeline process made use of cross-validation while the possibility of overfitting is still there. Overfitting occurs when the trainings data set corresponds with the test training set. This was assumed to appear in iteration 2.

Cross-validation makes use of three sets instead of a training and test set. The last set is used to verify the algorithm. Cross-validation is used to verify the model in order to reduce the impact of duplicates. It return the percentage of correctly predicted observations. Afterwards, the highest scoring configuration is trained in order to confirm the configuration.

6.4.2.1 MLP

The MLP pipelines are configured with many possible neural networks. The decision was made to differentiate between two, three and four layers. A small neural network was chosen while it is estimated that it will suffice. Each configuration is represented separately. The layers are presented as their corresponding amount of nodes. Larger neural networks with more layers will have an increasing amount of nodes. The MLP pipeline exploration resulted in a constant output of between 55% and 70%. High scoring configurations with a different amount of layers are trained to see the effect on the performance.


MLP (two layer pipeline)

MLP (three layer pipeline)

MLP (three layer pipeline)

MLP (four layer pipeline)

Figure 30: MLP configurations

6.4.2.2 Random forest

Random forest is applied on the data with a small amount of trees. This is chosen with perspective to the amount of data and parameters. The exploration performed consistently around 40%. It shows the algorithm is unsure about its accuracy. All configurations are trained in order to compare their operating capabilities. Training should provide a refined insight in the best configuration

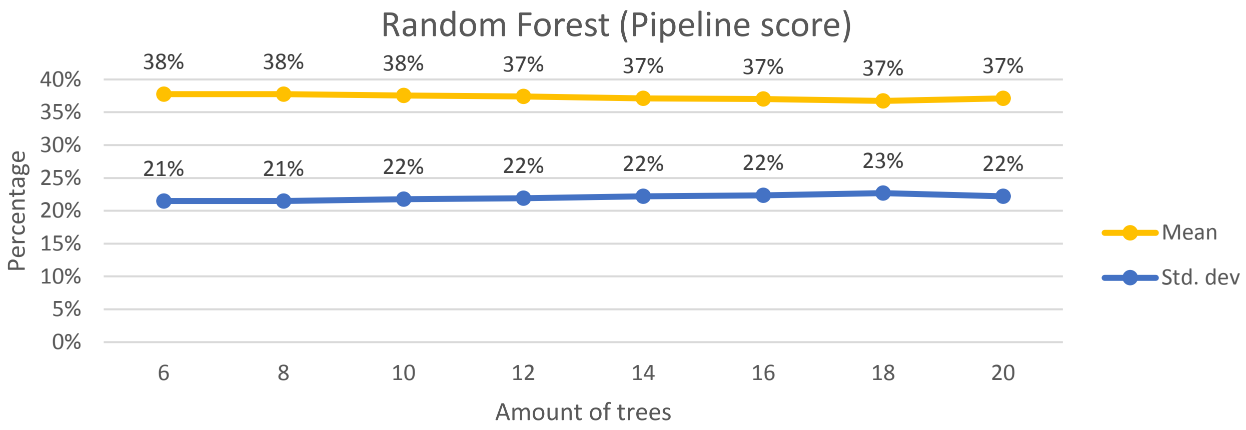


Figure 31: Random forest configurations

6.4.3 Results

The cross-validation improved the insight in the optimal configuration of the algorithm. Training of these configurations provided insight in the performance trend. The results of MLP and Random Forest show a better accuracy than expected in the cross-validation. Both algorithms will be elaborated separately.

The data exploration of MLP already showed a slight decline in accuracy of assigning collisions with a larger neural network. This trend can also be recognised in the trained neural network. The results of the trained MLP is shown in figures 32 and 33. The second configuration of MLP, two layers with each 6 nodes, showed a high decline in performance. The false negative assigned collisions increased in contrast to the false positive assigned collisions which declined. This unexpected decline could be explained by the 8% deviation in the data exploration. The configuration scored a 62% when the 8% deviation was subtracted. The other networks showed a similar trend in declining performance.

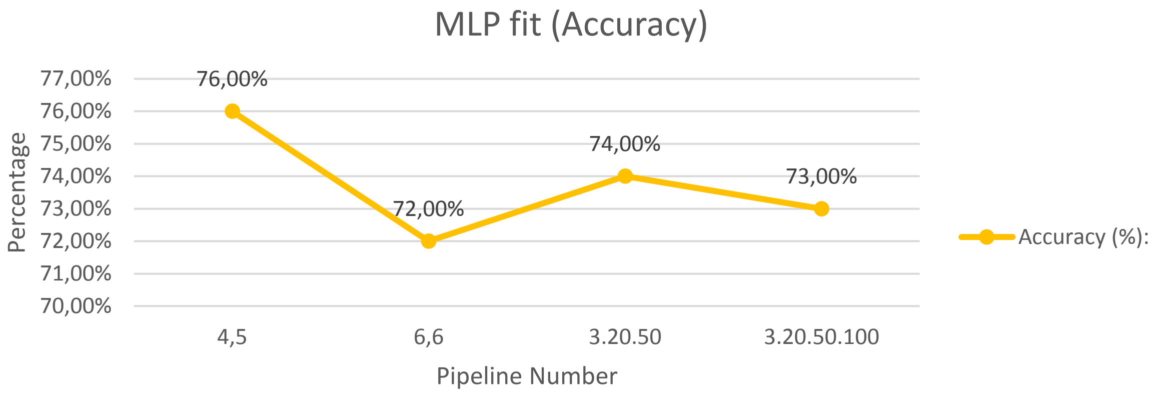


Figure 32: MLP fit (accuracy)

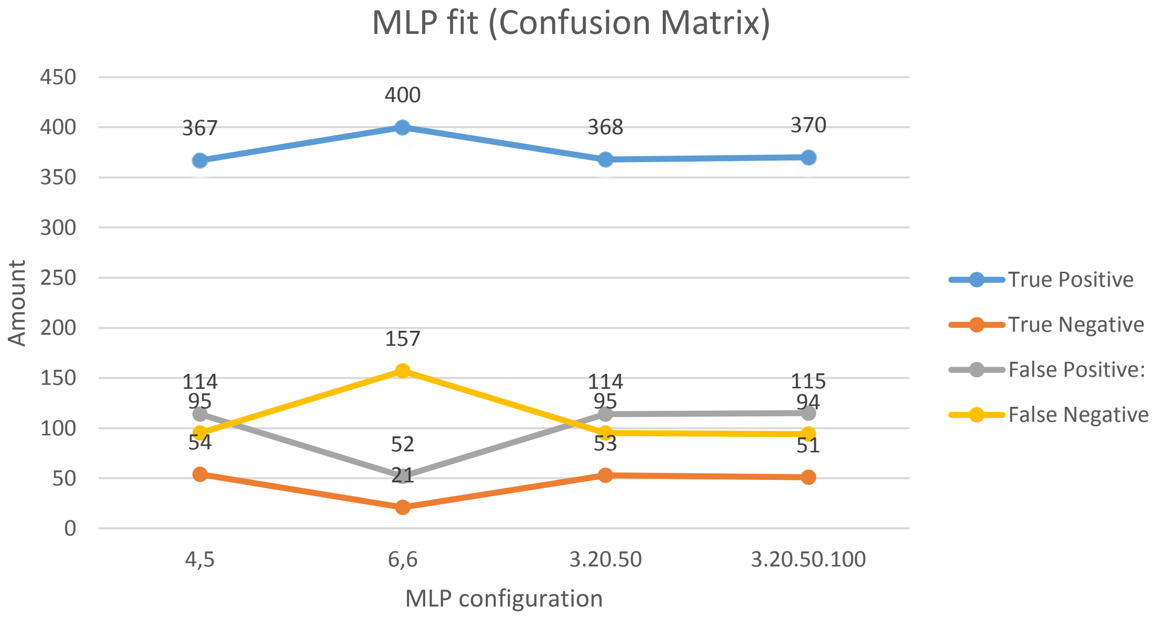


Figure 33: MLP fit (confusion matrix)

The increased accuracy of Random Forest was not expected while the data exploration only showed a 38% accuracy. This could hint at duplicates in the data and overfitting of the trained Random Forest algorithm. The performance of the trained algorithm is consistent for both accuracy and scores in the confusion matrix. The algorithm performed best between 12 and 16 trees. This would be recommended if there would be chosen to conduct further research on Random Forest.

Overall iteration 3 did not improve the performance of iteration 2. This could point at several causes. The cross-validation was lower than the outcome of the trained algorithm. This could be due to the lower amount of data to validate the model in a cross-validation process. Though, it could also hint at duplicates. This would be underpinned in the exceptional better performance of the Random Forest algorithm.

It could also hint at a performance barrier. The algorithm is at its optimal and could not be further improved with the current data set. Adding parameters would also not suffice while the amount of data does not permit this. Substitution of a parameters would be an option for improvement. Though, there the importance of the parameter must be determined.

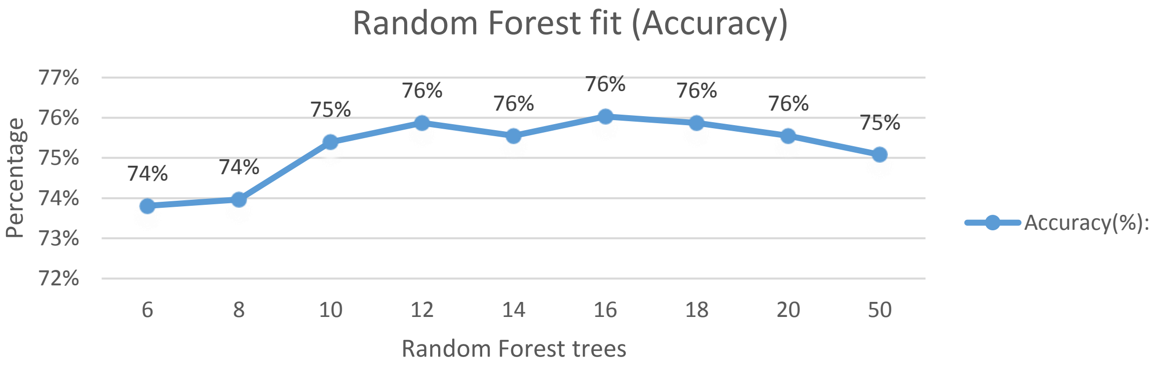


Figure 34: Random Forest fit (accuracy)

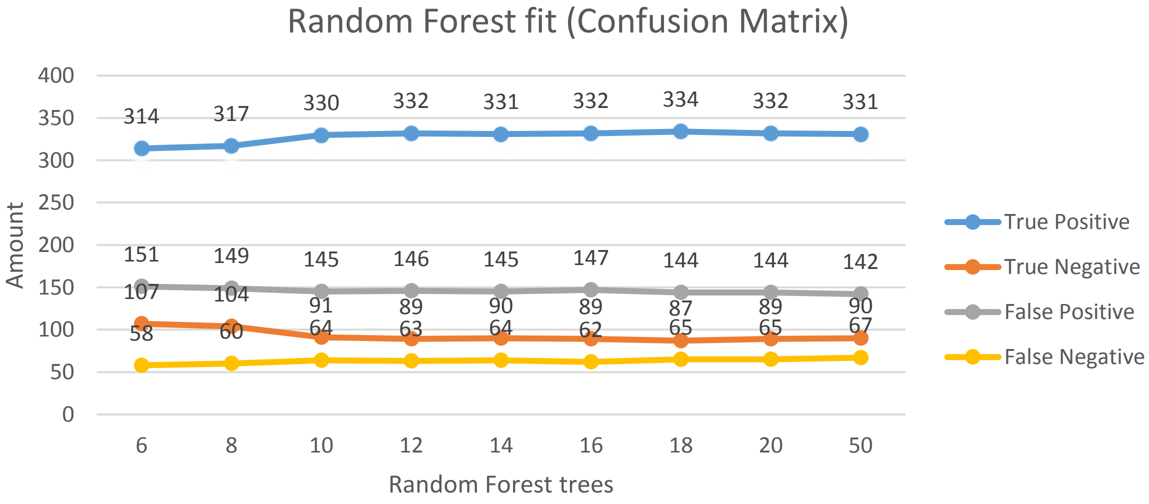


Figure 35: Random Forest fit (confusion matrix)

6.4.4 Verification

The configurations of the algorithm impacted the performance poorly. Further exploration on configuration is expected to have no impact on a more accurate performance. Optimisation should be investigated in other areas.

There are several opportunity to optimise the current performance. One of the opportunities would be to analyse the core data and manually determine the occurrence of duplicates and its magnitude. The other opportunity is the substitution of a parameter. The importance of the chosen parameters in relation to the algorithm should be determined in order to substitute the correct parameters. Another option would be to research the incorrect assigned observations. This could provide insight in the cause of incorrect assignment. Though, this could be difficult while the data is normalised and randomised before transfer to the machine learning algorithm.

Furthermore, there could be looked at opportunities outside the current data set. The collision process could be optimised in order to produce a sufficient amount of data. This is expected to be one of the key elements to acquire a higher performing algorithm. Parallel could be determined what the origin of duplicates are if duplicates occur.

7 Conclusion

The development of the machine learning algorithm started with comprehending the packaging process. This process secures the ability of all parts to fit and function correctly. There was aimed to enhance this process by a machine learning algorithm. The algorithm should assign a status to a collision, either relevant or not relevant. Afterwards, the process of developing a machine learning algorithm was described.

The iteration were aimed at optimising the available parameters and explore the relation between the configuration and output performance. The first iteration resulted in a 74% accuracy which looked promising. Though, during the following iteration a performance barrier was encountered. This iteration did not significantly improved the performance. This was mainly influenced by the amount of available observations. This is the key reason why there could not be conducted research on an extensive amount of parameters. Furthermore, the quality of the data was estimated to be of influence on the results. The third iteration showed the current data set could not be improved much more. The results showed a consistency of around 73%.

The aim to research several parameters not achievable due to the insufficient amount of observations. The decision was made to enhance the available features in order to improve the features interpretation for the machine learning algorithm. The results did not improve on this enhancement. Afterwards, different algorithm configuration are tested. The configurations with a small neural network proved to be more successful than deeper neural network. This is estimated to be influenced by the amount of observations and input parameters. A machine learning algorithm will be able to assign the collisions based on meta-data is possible but further research need to be conducted.

8 Discussion and recommendation

The key for a successful implementation of a machine learning algorithm is to acquire insight in the relation between the operating environments, choosing of parameters and the algorithm performance. All three fields must be known well while each field requires his own expertise. Furthermore, each decision made influenced other facets. In order to insure a well performing machine learning algorithm, a high amount of qualitative data must be available.

Iterating more on the current data set is estimated to be of no significant increase in performance. This is due to the low amount of available observations. It is recommended to look at the collision process while the process generates the data. It should be clear what the data processes are. Then the collision process could be optimised in order to produce a sufficient amount of data. This should be a priority to ensure an increasing stream of available data. Parallel could be determined what the origin of duplicates are if duplicates repeat to occur. The optimisation of the collision process is expected to be one of the key elements to acquire a higher performing algorithm.

Furthermore, there are opportunities of the current data set. The incorrect assigned collisions should be determined. This could help to gain insight in a possible pattern of the failure to assigning the observations. The algorithm can be altered accordingly after gaining this insight.

Another opportunity could be to add or substitute a parameter. This would require to determine the weighing of the parameters. A parameters would be obsolete if a parameter would fail to weigh significantly in the decision process. The substitution of a parameter could be underpinned in addition to the insight in the pattern of inaccurate assigned observations. A new parameter would then be chosen. A recommendation would be to add the collision geometry itself or the volume of the collision geometry. This would be estimated to be relevant information for the algorithm.

List of figures

List of tables

References

Chen, W., Hao, Z., Cai, R., Zhang, X., & Hu, Y. (2016). Multiple-cause discovery combined with structure learning. *Springer*.

David Camilo Corrales, J. C. (2018). *How to Address the Data Quality Issues in Regression.* Basel: MDPI.

Erik Bochinski, T. S. (2017). Hyper-parameter optimization for convolutional neural network commitees based on evolutionary algorithms. *IEEE*.

Ertel, W. (2016). *Grundkurs Künstliche.* Springer Fachmedien Wiesbaden GmbH.

Groll, M. (2008). *Interconnection Based Product and Process Documentation.*

Guéziec, A. (2001). "Meshsweeper": Dynamic Point-to-Polygonal-Mesh Distance and Applications. *IEEE*.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of Statistical Learning.* Springer.

Hirz, M., Dietrich, W., Gfrerrer, A., & Lang, J. (2013). *Integrated Computer-Aided Design in Automotive Development.* Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning.* Springer.

Jiang, S., Chin, K.-S., & Tsui, K. (2018). A universal deep learning approach for modeling the flow of patients under different severities. *Elsevier*.

Josip Stjepandić, Nel Wognum, & Verhagen, W. J. (2015). *Concurrent Engineering in the 21st Centruy.* Springer International.

Pérez-Rodríguez, J., Haro-García, A. d., Castillo, J. A., & García-Pedrajas, N. (2018). A general framework for boosting feature subset selection algorithms. *Elsevier*.

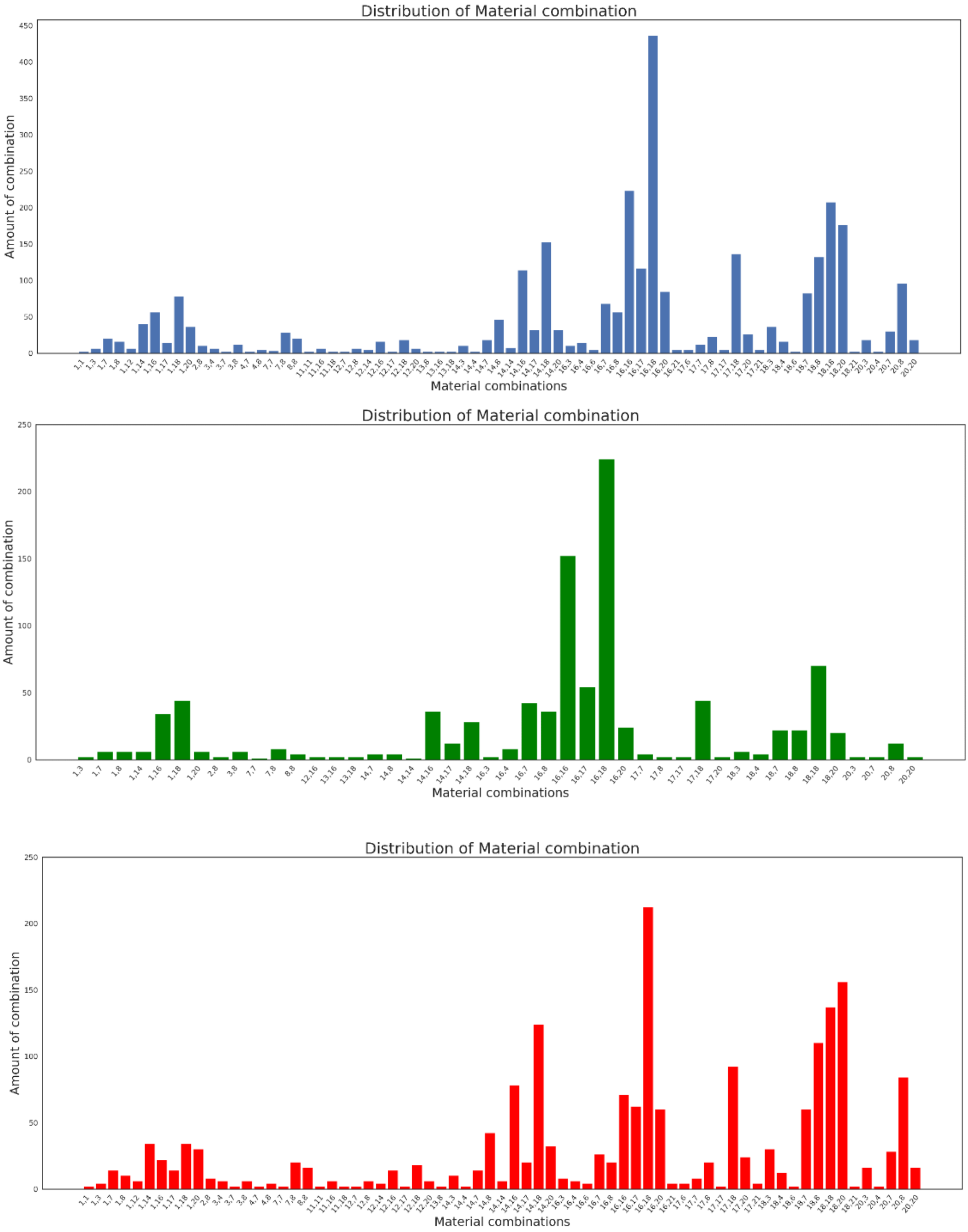
Reinsel, D., Gantz, J., & Rydning, J. (2017). *Data age 2025: The Evolution of Data to Life-critial.* IDC.

Shearer, C. (2000). *The CRISP-DM Model: The New Blueprint for.* THE DATA WAREHOUSING INSTITUTE.

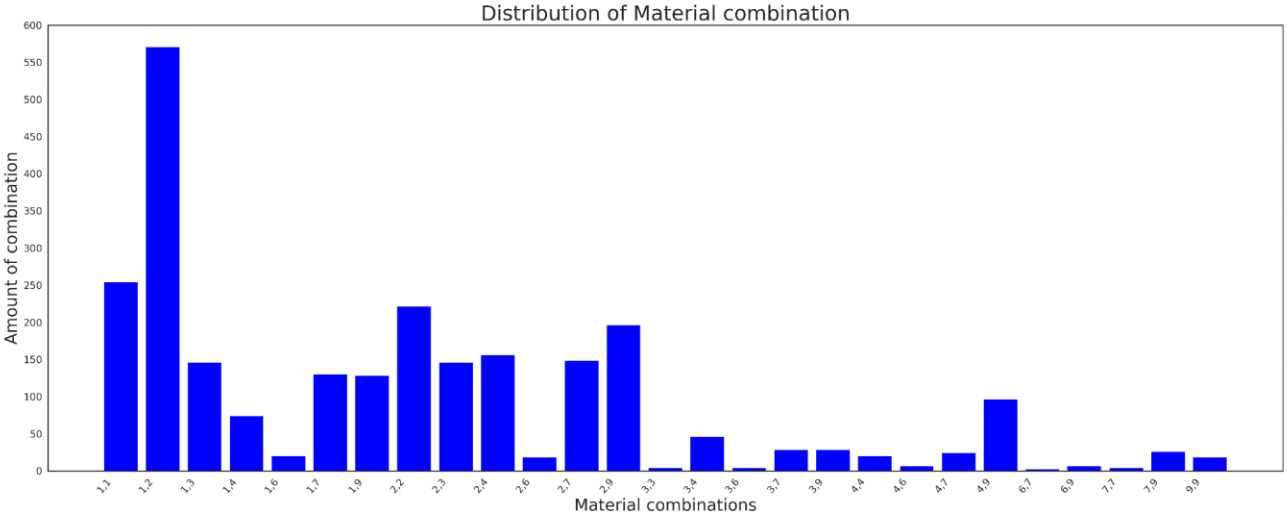
Yoon, S., & Kum, D. (2016). The Multilayer Perceptron Approach to Lateral Motion Prediction of Surrounding Vehicles for Autonomous Vehicles. *IEEE*.

Appendix:

A. Distribution material combination



Large version of figure 27: Material combination iteration 2 (top: all observations, middle: not relevant observations, bottom: relevant observations)



Large version figure 28: All observations, new material categorisation

B. New material categorisation

New category:		Old category:	
Number:	Name:	Number:	Name:
1	Metal	incl.	
		1	Aluminium
		3	Copper alloy
		4	Copper-zinc alloy
		11	Cast-iron
		16	Steel
		21	ZN-Leg
2	Plastic	incl.	
		2	Composite
		6	Thermoset plastics
		7	Elastomer
		18	Thermoplastic
		19	TPL Elastomer
3	Sealing material	incl.	
		5	Sealing material
		9	Foam
		14	Foam material
4	Electronics	incl.	
		8	Electric component
5	Glass	incl.	
		10	Glass
6	Fluids	incl.	
		12	Glue
		15	Lubrication fluid
7	Leather	incl.	
		13	Leather
		17	Textile
8	Wood	incl.	
		-	Wood
9	Remaining	incl.	
		20	Composite material
		-	Lack

New material categorisation