# TPOT's performance for Biomedical Data

*A Data Mining Seminar*

T.P.A. Beishuizen (0791613)
Biomedical Engineering - Computational Biology
Data Engineering - Information Systems
Eindhoven, University of Technology
Email: t.p.a.beishuizen@student.tue.nl

December 4, 2017

# Contents

# 1 Introduction

At the Computational Biology department (cBio) of Biomedical Engineering (BME), many requests are made to analyse gathered data. This data usually stems from research in hospitals, but can also be from other BME groups and publicly available data. Currently a standard is missing to efficiently analyse those data sets. With the vast number of data sets that are available, such a standard in the form of a framework on data analysis would be valuable. It would speed up projects and give them a higher chance to succeed the goal, due to improved efficiency. Before a framework can be made however a research must be done on all aspects that influence data analysis.

An example of biomedical data sets stemmed from the Catharina Hospital in Eindhoven. This extensive data set was a combination of two data sets. The first is a data set filled out by a doctor that analysed basic human features as well as the presence of co-morbidities. The second data set consisted of 41 markers measured pre- and post-surgery for 2367 patients that underwent gastric sleeve or gastric bypass surgery, also known as bariatric surgeries. The number of bariatric surgeries is increasing worldwide. Although initially thought otherwise, this type of surgery has added benefits on top of losing weight. Among those benefits the remission of metabolic co-morbidities can be named. Due to binary labelling of those co-morbidities, valuable information is lost. On top of that the labelling is not clearly defined either. To obtain more and better results, this binary labelling could be replaced by a continuous severity score. Ruben Deneer conducted a research on trying to achieve a successful replacement. This data set is a prime example of a data set that is not trivially preprocessed, when looking at multiple data sets, missing values and erroneous data.[1]

A possible solution for partly providing such a framework can be found in automated machine learning (AutoML). This relatively new extension to machine learning tries to automatically search the best combination of preprocessing, feature selection and machine learning algorithms to efficiently describe a data set. CPU or actual time and memory usage are the two main constraints for autoML, due to testing different methods with varying parameters taking up time and space. These autoML algorithms usually are made with the combined algorithm selection and hyperparameter optimization (CASH) challenge in mind.[2] The first autoML approach was published as `Auto-WEKA` that focused on classification methods, spanning 2 ensemble methods, 10 meta-methods, 27 base classifiers and their hyperparameter settings.[3] An upgrade was published that added regression and parallellism.[4]

Olson and Moore proposed an alternative tool for autoML, a Tree-based Pipeline Optimization Tool (TPOT).[5] This tool also tries to find the best classifier by creating pipelines for the different algorithms. The pipelines are evaluated and branched or altered according to the evaluation. It makes use of genetic programming, a technique used for evolutionary computations.[6] Evaluation of several TPOT approaches are proven better than doing a basic machine learning analysis and are therefore a promising approach for implementation. [7] Several projects used layering and meta-learning techniques to tackle the time and space constraints.[8]

To test if TPOT is also suitable for biomedical data, it must be tested by a specific set, such as the bariatric data set. This data consists of several biomedical data challenges only some of which seem solvable by TPOT. A research must be done to find out whether TPOT can manage to retrieve good results from this data set and may benefit of some improvements.

For the data seminar, first the background will be given. This background will be about the biomedical data sets, the exemplary data set used for the seminar, a topic of automated machine learning and at last a part about TPOT. Secondly the research question is given, together with a hypothesis.

# 2 Background

## 2.1 Biomedical data

Biomedical engineering can be seen as a specific part of engineering with a wide variety of topics. These topics can be theoretical, non-experimental undertakings, but also state-of-the-art applications. Not only research and development can be used, but also implementation and operation. Combining all of these different parts in one definition is hard. [9] For this project, the focus is mainly on research and development, also known as knowledge discovery. [10]

When a biomedical engineer starts a project, at the start usually only a data set and the research goal are known. To achieve that certain goal from the data set, four aspects influence the project's course and development. At first obviously the data itself is a big part of such an influencer as the research is restricted to limitations from it. Examples of such restrictions are multidimensionality, set size, data heterogeneity, missing feature values and population handling. The other obvious influencer is the main research goal. Since the biomedical engineer wants to achieve a certain goal, the approach outcome must match that goal for the research to be successful. Most goals are focused around either data mining, extracting relations from available data, or modelling, creating a model within data features. A third influencer is the availability of data analysis tools. The steps to take from data to goal do not only include an approach, but also a tool to execute it. The choice of a certain tool has a big impact on the project, as each one of them has its own advantages and disadvantages. The two most well known tools within BME are MATLAB and Python, however some engineers have used R, Java or C++ and there are still other possibilities. A last big influencer is the biomedical knowledge. What experience the scientist already has with similar projects can greatly influence the choice of approach and framework. Knowledge of the supervisor and publicly known information on the research subject from books and articles also influence the approach, as already known outcomes do not have to be researched again.

For data engineers the main focus lies in trying to find patterns in the data. Therefore in this seminar the focus will mainly be on the data driven aspect of a biomedical project. Characteristics of such a data driven approach (as mentioned before) are mainly focused around data volume, dimensionality, complexity, heterogeneity and quality.[11, 12]

Collecting data because it is possible can make data sets bigger than needed. Both in number of instances and features, data sets can be harder to understand or analyse when more is available.[11] This volume problem usually is tackled by taking sub-populations of the complete set. These subsets can either be focused around a part of the population (gender, age, race) or taken at random to still represent all of it. Due to the efficiency of analysis techniques and the rise in computational speed of servers[13], volume on its own becomes less of an issue. Volume does however become an issue when combining with heterogeneity and quality. [14, 15]

Not all data sets have a high number of instances that cause a big data volume. Sometimes there are relatively few instances, while the number of features is proportionally high. [16] Usually many of those features are not relevant enough for the research, however are still used for testing. Trying to remove features that are not important, will greatly help finding relations between the others and create more knowledge about the research topic. Lowering the number of features also makes the data volume go down, so analysis should be easier. Mainly an optimal features set should be selected to obtain the best results. [17]

Biomedical data can also be very complex. Useful results may be present, however it can be very hard to obtain it. Examples of complex data are images, several biomedical signals and temporal data. Details of the useful results that are present in images can for example be very hard to detect, the temporal data can vary quite much over time and the biomedical signals can be hard to combine with static biomarkers. [18] This aspect can benefit from exchanging knowledge with other research areas that specialize in mining of those complex data sets. [14, 19]

The biggest challenge encompasses aligning different data sets. No standard for data sets is available and therefore data sets differ greatly from each other. Data is weakly structured or even unstructured [15] and variables are processed differently due to other protocols or the collectors'

preference of representation.[20] Also the variety of data is hard to combine when sources are fundamentally different. When parts of the data are images, another part is a table from the laboratory and a third part is textual remarks of the doctor, standardizing merging those three is much harder than merging three lab sets. Those merges are also very prone to errors, as imprecisions can be vastly different between those data sets. No tool can work directly with these raw data sets and preprocessing must almost definitely occur beforehand.[14, 21]

A last challenge is about data quality. The data is usually gathered by doctors and laboratory workers. Since the data is manually gathered by humans, the data have a relatively high error rate. Therefore the data can be quite noisy, values can be inconsistent, wrongly entered or even missing.[21] Not only human errors cause the data quality to drop, but the heterogeneity, as well. Two hospitals might have different protocols for the same treatment and sample different biomarkers for that protocol. Due to that difference, biomarkers may be missing for some of the entries. The time of data gathering is also a big factor as some biomarkers change greatly over time. The databases are usually also built for financial purposes and not for research, which can hurt the quality. [18]

These challenges within the data are greatly discussed.[19] Many proposals to tackle them are made, however none is actually widely adopted, yet, as a global standard for databases. Also, with the uncontrolled growth in biomedical data, it will become hard to have such a standard recognised all over the world. [20, 22, 23, 24, 25, 26]

## 2.2 Bariatric Data Set

To test the autoML tool, data based on bariatric patients are used. This data consist of two separate data sets. The first one is called "The Dutch Audit for Treatment of Obesity" (DATO). This data set is a national database that houses all registrations and health statuses of pre- and post treatment bariatric surgery patients in the Netherlands. Several basic variables are noted, such as height, weight, BMI, age and date. Before surgery, the co-morbidities T2DM, hypertension and dyslipidemia were given a binary label of "Yes/No". After surgery they were given one of the following labels:

1. **Cured** No co-morbidity any more
2. **Improved** Less affected by co-morbidity
3. **Same** No change in co-morbidity status
4. **Worse** More affected by co-morbidity
5. **Denovo** Diagnosed co-morbidity while not present before surgery
6. **Not present** No co-morbidity present

The second data set came from a laboratory database, stored in health records. This extensive data set consisted of 3 clinical and 38 blood markers measured pre- and 6, 12 and 24 months post-surgery. The tests pre-surgery had some additional markers on top of the 41 ones. These markers can be divided in the following categories: (Table 2.2) Complete blood count, liver function, kidney function, inflammation, lipid spectrum, coagulation, glucose metabolism, thyroid function and at last minerals and vitamins. The data sets of the patients that underwent bariatric surgery can be extracted from these.

A first challenge of the two data sets is to combine these two data sets. Some challenges arise when doing so. Such a challenge is obviously to find the right connection between them, using the survey and lab data of the same patient. Since most likely they are not always made available on the same day, possible extrapolation needs to be used to properly link them. These challenges must be solved before asking the question what markers can say about the severity of co-morbidities.

Table 1: The markers present in the bariatric laboratory data set [1]

| | Before Surgery/Pre-Op/Screening | After Surgery/Post-Op/Follow-up |
|---|---|---|
| Complete blood count | hemoglobin | hemoglobin |
| | hematocrit | hematocrit |
| | erythrocytes | erythrocytes |
| | mean corpuscular hemoglobin | mean corpuscular hemoglobin |
| | mean corpuscular volume | mean corpuscular volume |
| | thrombocytes | thrombocytes |
| | leukocytes | leukocytes |
| Liver function | bilirubin | bilirubin |
| | aspartate aminotransferase | aspartate aminotransferase |
| | alanine aminotransferase | alanine aminotransferase |
| | lactate dehydrogenase | lactate dehydrogenase |
| | alkaline phosphatase | alkaline phosphatase |
| | gamma-glutamyltransferase | gamma-glutamyltransferase |
| Kidney function | urea | urea |
| | creatinine | creatinine |
| | potassium | potassium |
| | sodium | sodium |
| | calcium | calcium |
| | phosphate | phosphate |
| | albumin | albumin |
| Inflammation | C-reactive protein | C-reactive protein |
| Lipid spectrum | total cholesterol | total cholesterol |
| | high-density lipoprotein-cholesterol | high-density lipoprotein-cholesterol |
| | total/high-density cholesterol ratio | total/high-density cholesterol ratio |
| | low-density lipoprotein-cholesterol | low-density lipoprotein-cholesterol |
| | triglycerides | triglycerides |
| Coagulation | prothrombin time | prothrombin time |
| Glucose metabolism | hemoglobin A1c (IFCC) | hemoglobin A1c (IFCC) |
| | glucose | glucose |
| | insulin | - |
| | C-peptide | - |
| Thyroid function | parathyroid hormone | parathyroid hormone |
| | thyroid-stimulating hormone | - |
| | free T4 | - |
| | cortisol | - |
| Minerals and vitamins | iron | iron |
| | ferritin | ferritin |
| | folic acid | folic acid |
| | zinc | - |
| | magnesium | - |
| | vitamin A | - |
| | vitamin B1 | vitamin B1 |
| | vitamin B6 | vitamin B6 |
| | 25-OH vitamin D | 25-OH vitamin D |
| | vitamin B12 | vitamin B12 |

A second challenge lies in the absence of several values. Some blood markers pre-surgery were not measured post-surgery and some of them were discarded from the marker panel after doctors disregarded them as not useful. This missing values can be preprocessed different ways, for example changing their values the median value of the set, or to the value of a or multiple nearest neighbours. Also a decision must be made when to discard measurements if too many values are missing.

A third challenge is to actively cope with human input errors. The hospital estimated 5% of all values added by medical staff has an error, which means one in twenty values is erroneous, a non disregarding amount. An example of this would be the weight of a person. It was 632kg, which should actually be 63.2 kg. More of those errors are present, which can be partially filtered by for example outlier detection.[1]

A Graduation project of Ruben Deneer was done to add a severity score to the co-morbidities of this data set. This project used the following pipeline to obtain results:

1. *Merging data sets.* The lab and DATO sets were matched when they were data of the same patient and in timespan closer to each other than three months. For possible double matches only the closest one was taken and matchin data set outside the predefined measurement dates (pre-surgery and 6, 12 and 24 months post surgery) were removed as well.

2. *Removing missing values.* Data features were removed when roughly more than 30% of the data was missing. Data entries were removed when missing values were present.

3. *Feature preprocessing* Several data features were preprocessed to closer resemble values that are known useful for biomedical research.

4. *Logistic regression* Logistic regression was used to create a model that gave a severity score to the bariatric values. Two types of logistic regression were used, proportional odds and continuation ratio. This logistic regression was done with a 10-fold cross validation and its assessment of fit was measured by an ROC-curve.

5. *Model visualization* The model was at last visualized by a nomogram, an old fashioned table that doctors can quickly understand how the model can be explained (Figure 1).

## 2.3 Automated Machine Learning

Before automated machine learning (autoML) existed, a dataset was mined by hand. First a preprocessing algorithm was chosen and used to prepare the data. Next a (machine learning) algorithm was chosen to mine the desired results out of the data. At last the hyperparameters of the chosen algorithm were tuned to optimize the desired results. These three steps are vastly different and significant issues arise when combining these. Several ideas arose to combine the steps, called Combined Algorithm Selection and Hyperparameter optimization (CASH).[3] After some time, when preprocessing was added in the mix as well, the name autoML was being used.[8]

As explained shortly before, to go from data and results several steps must be taken: Preprocessing, algorithm selection and hyperparameter optimization. This sequencing is called a machine learning pipeline. Such a pipeline can consists of zero, one or multiple preprocessing steps for data preparation, can be one of many different machine learning algorithms which on their part have wide ranges for multiple hyperparameters. The explosion of possible pipelines makes it hard to choose the right one. Knowing successful combinations is useful, however every data set has different features that ask for different pipelines. [8]

AutoML tires to find the best machine learning pipelines to compute which algorithms must be selected, combined with tuning the hyperparameters and preprocessing. This algorithm selection usually is done in a meta-learning approach, which focuses on finding how the machine learning algorithms perform for a task interval. Hyperparameter optimization has challenges on his own to find the right ones and there are many different approaches to tackle preprocessing. All of these are discussed in their own subsection to briefly explain them.

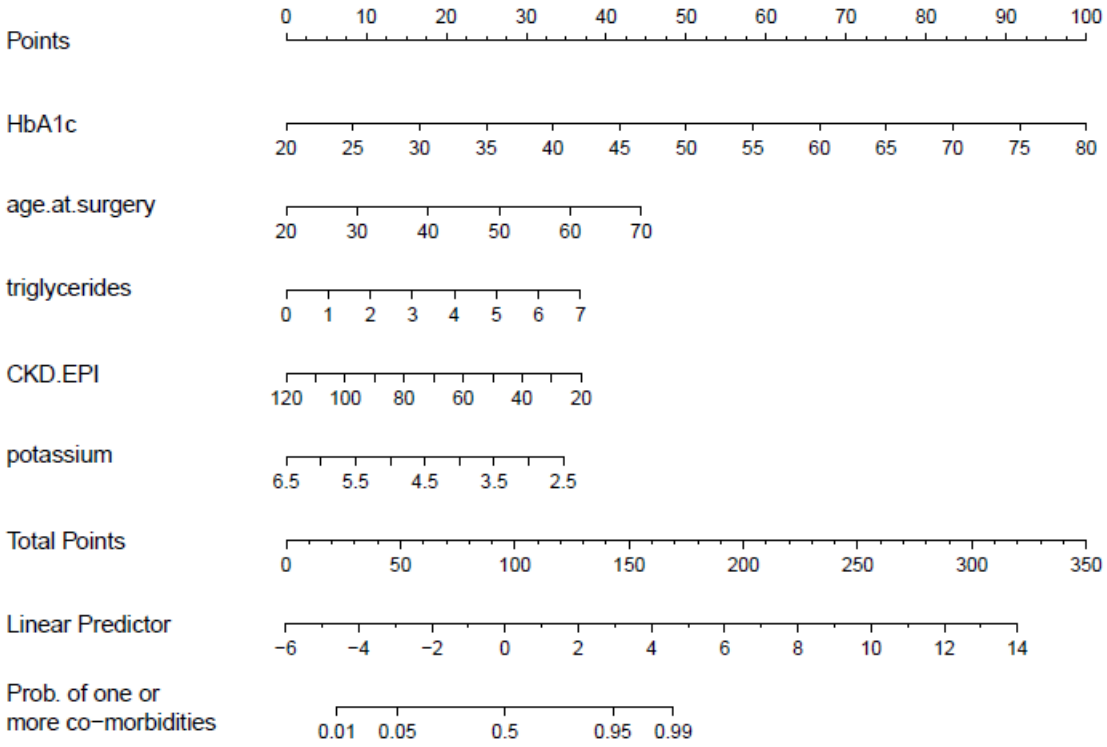


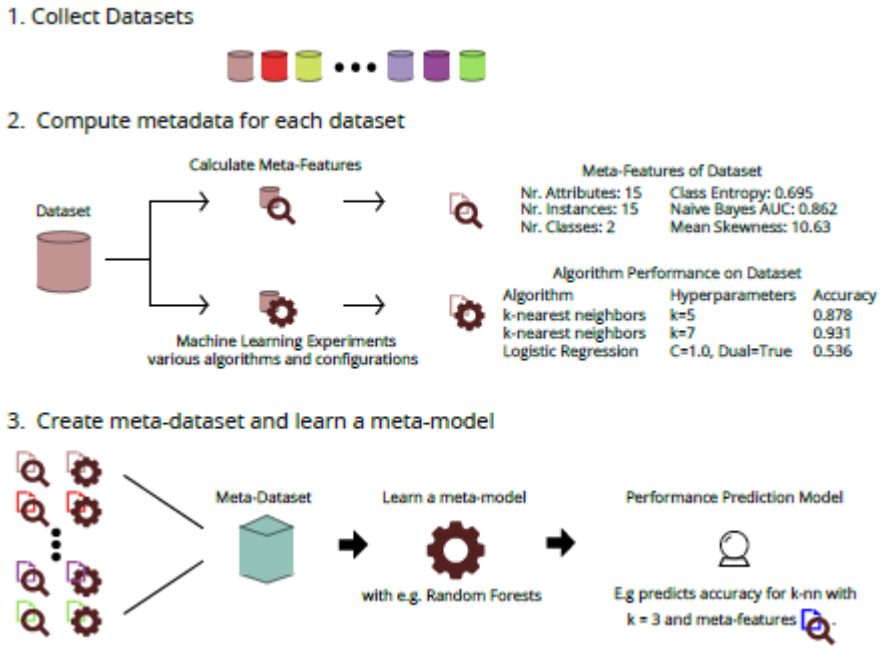

Figure 1: The nomogram that explained the created comorbidity severity score of bariatric patients.[1]

### 2.3.1 Meta-Learning

Machine learning algorithms show different behaviour for different tasks. Meta-learning tries to find out how their performance changes between those tasks (Figure 2.3.1). It tries to link the algorithm with data sets it would do good for and tries to find which hyperparameters give a good performance. In combination with the machine learning pipelines, meta-learning would try to find the best ones available. Since not only algorihtms must be selected, but also hyperparameters must be optimized and preprocessing must be done, the time and space needed for meta-learning explodes. This can be lessened when removing bad pipelines and limiting the range of hyperparameters, machine learning- and preprocessing algorithms as much as possible.

Features of the meta-learning phenomenon are used to predict the performance. There are three types of these meta-features. The first type is simple, statistical and information-theoretic. They can be a basic feature of the data set, as well as a value after a statistical computation or a specific theoretical value. The second meta-feature type can be called landmarks. Landmarks give the performance of algorithms, how well they are doing with the given data set. The last meta-feature category is model-based. Specific characteristics of the used model can be used as meta-features as well.[27, 28]

For using those meta-features in picking the best machine learning algorithm meta-learners can be used. Meta-learners are algorithms that choose between the possible choices. There are four ways of doing that. The first is plainly choosing the best algorithm in the set, this choice speeds up the process but is prone to being a bad choice. Second a subset of good algorithms can be chosen, which is slower, but has a higher chance to give a good outcome. Thirdly the algorithms can be ranked, which makes the chance of picking a good algorithm quicker starting at the top. Fourth is to use estimations of performance which gives information expectations over




[h!]

Figure 2: A layout of how meta-learning works. 1. Data sets are collected. 2. Meta-data is computed for each dataset. 3. A meta-dataset is created and a meta-model is learned.[8]

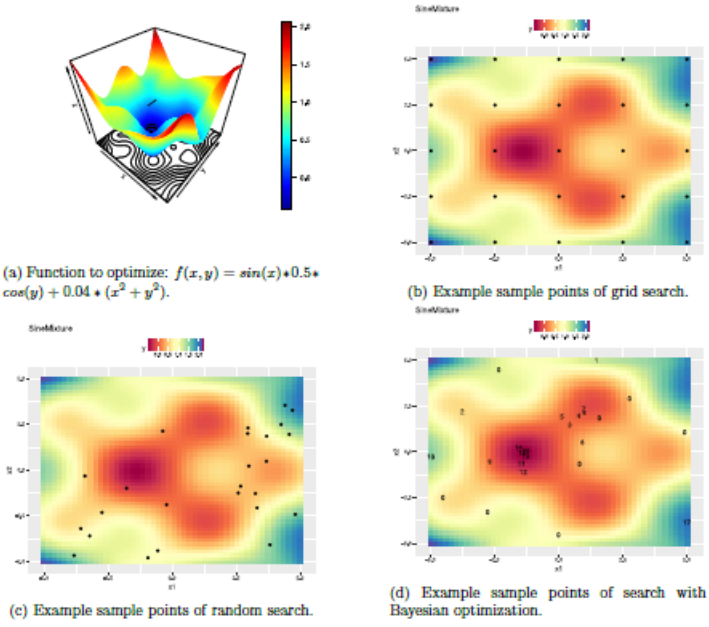
all algorithms.[29]

### 2.3.2 Hyperparameter Optimization

As discussed before, machine learning algorithms have hyperparameters. These type of parameters are very sensitive and can change the algorithm performance greatly, hence the hyper- prefix. These hyperparameters can be nonlinear and nonconvex which results in it being hard to find the optimal value. They are many different variable types of hyperparameters, which makes standardizing optimization hard. They can also be dependent on each other and therefore have useless combinations. At last they can change the computation time drastically for a minor change in parameter.[30]

There are several approaches to find the right values for the hyperparameters. Three approaches will be discussed. At first there is grid search, that checks all possible combinations for hyperparameters on a predefined interval. This approach does effectively check the complete area the optimal solution can be in, however it also takes much computational time due to the combinatorial explosion principle.[31] A second approach is the random search, when values are chosen for each hyperparameter at random. It is proven that this search is better in an empirical and theoretical way, due changing some hyperparameters hardly making any difference.[32]

A more advanced way of optimizing hyperparameters is Bayesian Optimisation. This hyperparameter optimization approach tries to use earlier results to find the best possible location for the hyperparameters to be optimized. At the start, with random sampling several sample points are measured. An acquisition function, with the input from earlier samples, which combination of hyperparameters should be tested next. It balances between trying to explore areas with good performance and trying to explore bigger areas for possible other good areas.[33] The three optimization techniques are shown together to explain the difference (Figure 2.3.2).



(a) Function to optimize: $f(x,y) = sin(x) * 0.5 * cos(y) + 0.04 * (x^2 + y^2)$.

(b) Example sample points of grid search.

(c) Example sample points of random search.

(d) Example sample points of search with Bayesian optimization.

[H!]

Figure 3: Examples of the three explained hyperparameter optimization techniques. The numbers in the bayesan optimization show the iteration of sampling.[8]

### 2.3.3 Preprocessing

Preprocessing is one of the three aspects of autoML. Whereas the goal is important when choosing a type of machine learning algorithm, preprocessing algorithms always need to be fine-tuned for the available data set, as every set is different. Data sets, and specially biomedical ones for this project, have specific challenges that can be solved using preprocessing (subsection 2.1). These challenges can be classified in two regions. The first one is about problems with the data, Tye second one is data preparation.[34]

From a preprocessing point of view, there can be three different types of problems with data (Figure 4). The first problem is that there is too much data. The data can be noisy, irrelevant, too big, different types of data can be present and feature extraction still has to be done. The data for this project is an example for noisy data (subsection 2.1), as 5% of the data is estimated to be wrong.

A second type is the opposite of having too much data. There can also be too little data. Values of an attribute or complete attributes can be absent from the data. The number of data points can also be very low. For the example data set, there were many cases where attributes were missing.

A third problem with the data can be that it is fractured. Multiple separate data sets can be incompatible, come from multiple sources or are on different processing levels. When taking the bariatric data as an example, it stems from two different data sets. A challenge is to combine these two as one data set, as one, while they are not specifically made for that.

To tackle those three data problems, again three types of techniques can be used. At first data can be transformed to become more usable. The most important transformation is noise removal. It can be removed with smoothing function[35, 36], or a more advanced machine learning technique to also detect it.[37]

A second preprocessing technique is gathering data if needed. Data selection is important, as it could be that not all data is as relevant as all the other data. Important techniques to be mentioned are principal component analysis, that checks the relevance for every feature in the data set.[38]

At last new information can be generated, if needed. This can be done by simulation or by adding new features. Data points can be fused to become a new point. This way more data is available for data analysis. Also when values are missing, several techniques can be used for value imputation. Extrapolation can be done, using regression to estimate its value. Other methods are based on nearest neighborhood frameworks.[39]

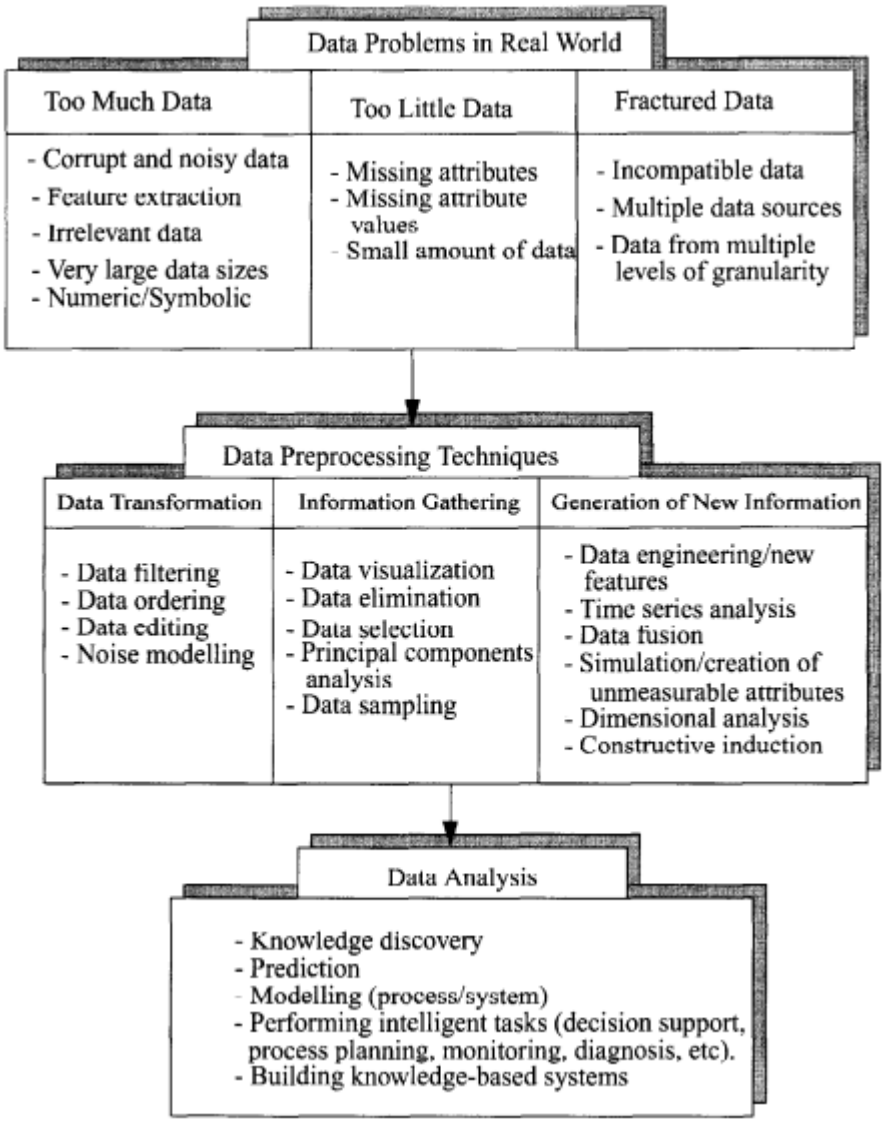



Figure 4: A schema that shows the process of preprocessing.[34]

Aside from data problems, another reason for preprocessing can be present.[34] Data can also be very raw. STo reach a certain goal, a data set has the necessary iformation, but only indirectly. An example would be hypertension. To know if someone has hypertension, its blood pressure must be measured and checked if that value is high enough. this preprocessing is highly data set specific, as computers do not know the specifications of blood pressure for someone having hypertension.

## 2.4 Tree-based Pipeline Optimization Tool

A tool that implements autoML is tree-based pipeline optimization tool (TPOT). It uses the machine learning pipelines and evolutionary optimization to find the best solution for every data set. This evolutionary optimization is done by genetic programming. Genetic programming evolves possible solutions to find a better solution. This evolution is done by first evaluating them and selecting the best ones to continue to the next generation. Then both crossovers between and mutations on possible solutions are performed. After that again evaluation and selection, followed by crossovers and mutations, take place a number of times until a certain quality is found, time has run out or another ending condition has been met.

TPOT makes use of this genetic programming with using the machine learning pipelines in a tree (Figure 5). TPOT consists of preprocessing and machine learning algorithms, that form the backbone of the pipelines. Their hyperparameters and the data set are the variables. TPOT makes mostly use of the machine learning and preprocessing algorithms of skikit-learn from Python in which it is also written.

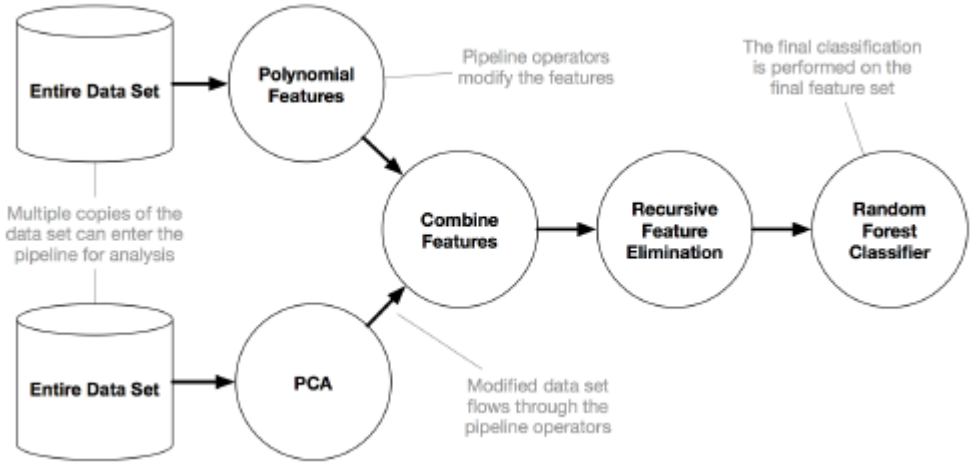



Figure 5: An example of a machine learning pipeline in TPOT. It only shows the primitive algorithms and not hyperparameter terminals. At the root is the machine learning algorithm.[8]

TPOT has three different types of mutations within one pipeline.The first one is insertion, inserting a primitive somewhere in the tree. An example would be the insertion of an additional preprocessing algorithm. The second one is replacement, which replaces a random terminal. It can for example change a binary hyper parameter from true to false. The third one is shrinking. A primitive is replaced by a terminal. For example a preprocessing step can be replaced by just raw data. This different mutations can all be seen visually (Figure 6)

TPOT also focuses on mutations between two pipelines through the means of crossovers. Between two pipelines, sub-trees and primitives can be changed, given that the both pipelines remain valid (Figure 7). Every time a crossover is performed, two separate pipelines are used and changed, creating two new ones.

Comparing the possibilities from TPOT and the challenges in biomedical data, it seems that TPOT has some implementations to tackle them. It has several different scalers (StandardScaler, RobustScaler, MinMaxScaler) to tackle feature heterogeneity between different data sets and errors. It also has some feature selection operators to tackle errors (VarianceThreshold, SelectKBest, SelectPercentile). For missing values, it seems no solution is present.

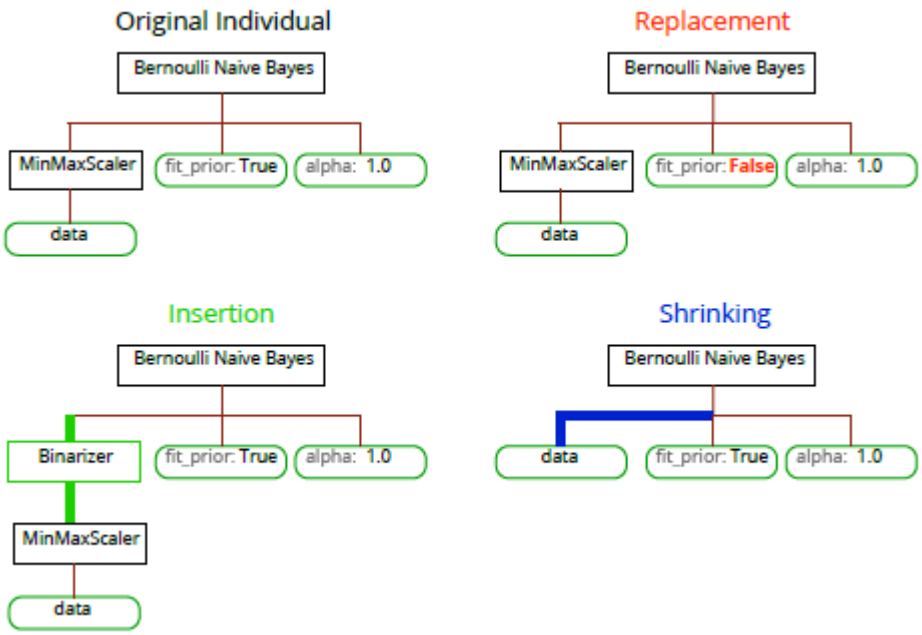



Figure 6: Examples of the three mutations in the TPOT algorithm: insertion, replacement and shrinking.[8]

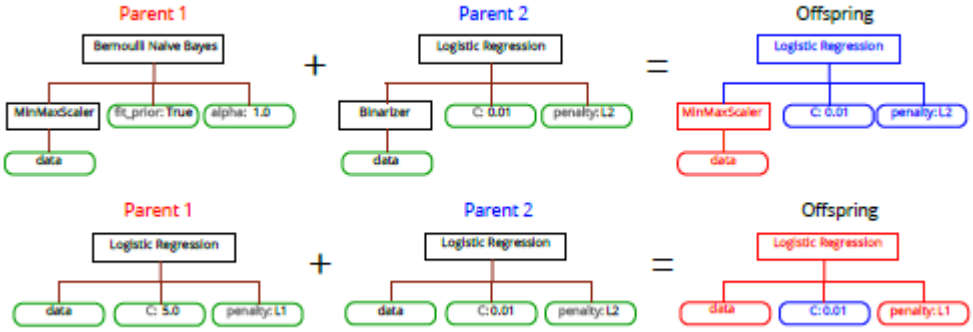



Figure 7: An example of a TPOT crossover[8]

# 3 Research Question

TPOT has shown to give promising results for several different data sets.[8] For this project the focus will be on biomedical data sets (subsection 2.1) and how it handles specific problems in these data sets. An example data set (subsection 2.2) is taken for analysis and to create suggestions for future extensions of TPOT. The research question for TPOT will be the following.

How does TPOT perform on specific biomedical data set problems and how can it be improved on them?

Knowing that some algorithms exist to tackle the problems with biomedical data sets, the first step should be to find out whether all problems can already be tackled by TPOT. This seems the case for fractured data and errors in the data, both challenges seem to have some way to be

handled. Missing values however do not seem to be targeted. There are no ways to have value imputation and the only way this seems to be tackled is by giving it the median value. It seems improvement on that part would be beneficial.

# References

[1] R. Deneer, "Scoring co-morbidity severity in bariatric patients based on biomarkers: a data mining approach," 2017.

[2] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, and F. Hutter, "Efficient and robust automated machine learning," in *Advances in Neural Information Processing Systems*, pp. 2962–2970, 2015.

[3] C. Thornton, F. Hutter, H. H. Hoos, and K. Leyton-Brown, "Auto-weka: Combined selection and hyperparameter optimization of classification algorithms," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 847–855, ACM, 2013.

[4] L. Kotthoff, C. Thornton, H. H. Hoos, F. Hutter, and K. Leyton-Brown, "Auto-weka 2.0: Automatic model selection and hyperparameter optimization in weka," *Journal of Machine Learning Research*, vol. 17, pp. 1–5, 2016.

[5] R. S. Olson and J. H. Moore, "Tpot: A tree-based pipeline optimization tool for automating machine learning," in *Workshop on Automatic Machine Learning*, pp. 66–74, 2016.

[6] W. Banzhaf, P. Nordin, R. E. Keller, and F. D. Francone, *Genetic programming: an introduction*, vol. 1. Morgan Kaufmann San Francisco, 1998.

[7] R. S. Olson, N. Bartley, R. J. Urbanowicz, and J. H. Moore, "Evaluation of a tree-based pipeline optimization tool for automating data science," in *Proceedings of the 2016 on Genetic and Evolutionary Computation Conference*, pp. 485–492, ACM, 2016.

[8] P. Gijsbers, "Automatic construction of machine learning pipelines," 2017.

[9] J. D. Bronzino and D. R. Peterson, *Biomedical engineering fundamentals*. CRC press, 2014.

[10] M. Bramer, *Principles of data mining*, vol. 180. Springer, 2007.

[11] H. Chen, S. S. Fuller, C. Friedman, and W. Hersh, *Medical informatics: knowledge management and data mining in biomedicine*, vol. 8. Springer Science & Business Media, 2006.

[12] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in Bioinformatics*, p. bbx044, 2017.

[13] D. Blythe, "Rise of the graphics processor," *Proceedings of the IEEE*, vol. 96, no. 5, pp. 761–778, 2008.

[14] C. Turkay, F. Jeanquartier, A. Holzinger, and H. Hauser, *On Computationally-Enhanced Visual Analysis of Heterogeneous Data and Its Application in Biomedical Informatics*, pp. 117–140. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014.

[15] A. Holzinger and I. Jurisica, *Knowledge Discovery and Data Mining in Biomedical Informatics: The Future Is in Integrative, Interactive Machine Learning Solutions*, pp. 1–18. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014.

[16] W. Dubitzky, M. Granzow, and D. P. Berrar, *Fundamentals of data mining in genomics and proteomics*. Springer Science & Business Media, 2007.

[17] Y. Peng, Z. Wu, and J. Jiang, "A novel feature selection approach for biomedical data classification," *Journal of Biomedical Informatics*, vol. 43, no. 1, pp. 15 – 23, 2010.

[18] I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J.-F. Chang, and L. Hua, "Data mining in healthcare and biomedicine: A survey of the literature," *Journal of Medical Systems*, vol. 36, pp. 2431–2448, Aug 2012.

[19] R. Bellazzi, M. Diomidous, I. N. Sarkar, K. Takabayashi, A. Ziegler, A. T. McCray, *et al.*, "Data analysis and data mining: current issues in biomedical informatics," *Methods of information in medicine*, vol. 50, no. 6, p. 536, 2011.

[20] D. Otasek, C. Pastrello, A. Holzinger, and I. Jurisica, *Visual Data Mining: Effective Exploration of the Biological Universe*, pp. 19–33. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014.

[21] K. J. Cios and G. W. Moore, "Uniqueness of medical data mining," *Artificial Intelligence in Medicine*, vol. 26, no. 1, pp. 1 – 24, 2002. Medical Data Mining and Knowledge Discovery.

[22] L. Marenco, T.-Y. Wang, G. Shepherd, P. L. Miller, and P. Nadkarni, "Qis: A framework for biomedical database federation," *Journal of the American Medical Informatics Association*, vol. 11, no. 6, pp. 523–534, 2004.

[23] V. Y. Bichutskiy, R. Colman, R. K. Brachmann, and R. H. Lathrop, "Heterogeneous biomedical database integration using a hybrid strategy: a p53 cantcer research database," *Cancer informatics*, vol. 2, p. 277, 2006.

[24] W. Sperzel, R. Abarbanel, S. Nelson, M. Erlbaum, D. Sherertz, M. Tuttle, N. Olson, and L. Fuller, "Biomedical database inter-connectivity: an experiment linking mim, genbank, and meta-1 via medline.," in *Proceedings of the Annual Symposium on Computer Application in Medical Care*, p. 190, American Medical Informatics Association, 1991.

[25] F. Aubry, S. Badaoui, H. Kaplan, and R. D. Paola, "Design and implementation of a biomedical image database (bdim)," *Medical Informatics*, vol. 13, no. 4, pp. 241–248, 1988.

[26] D. Windridge and M. Bober, *A Kernel-Based Framework for Medical Big-Data Analytics*, pp. 197–208. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014.

[27] P. Brazdil, J. Gama, and B. Henery, "Characterizing the applicability of classification algorithms using meta-level learning," in *European conference on machine learning*, pp. 83–102, Springer, 1994.

[28] R. Vilalta, C. G. Giraud-Carrier, P. Brazdil, and C. Soares, "Using meta-learning to support data mining.," *IJCSA*, vol. 1, no. 1, pp. 31–45, 2004.

[29] P. Brazdil, C. Giraud-Carrier, C. Soares, and R. Vilalta, "Development of metalearning systems for algorithm recommendation," *Metalearning: Applications to Data Mining*, pp. 31–59, 2009.

[30] M. Claesen and B. De Moor, "Hyperparameter search in machine learning," *arXiv preprint arXiv:1502.02127*, 2015.

[31] C.-W. Hsu, C.-C. Chang, C.-J. Lin, *et al.*, "A practical guide to support vector classification," 2003.

[32] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, no. Feb, pp. 281–305, 2012.

[33] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in *Advances in neural information processing systems*, pp. 2951–2959, 2012.

[34] A. Famili, W.-M. Shen, R. Weber, and E. Simoudis, "Data preprocessing and intelligent data analysis," *Intelligent data analysis*, vol. 1, no. 1-4, pp. 3–23, 1997.

[35] R. Somorjai, M. Alexander, R. Baumgartner, S. Booth, C. Bowman, A. Demko, B. Dolenko, M. Mandelzweig, A. Nikulin, N. Pizzi, *et al.*, "A data-driven, flexible machine learning strategy for the classification of biomedical data," *Artificial intelligence methods and tools for systems biology*, pp. 67–85, 2004.

[36] A. Karagiannis and P. Constantinou, "Noise-assisted data processing with empirical mode decomposition in biomedical signals," *IEEE Transactions on Information Technology in Biomedicine*, vol. 15, no. 1, pp. 11–18, 2011.

[37] D. Gamberger, N. Lavrac, and S. Dzeroski, "Noise detection and elimination in data preprocessing: experiments in medical domains," *Applied Artificial Intelligence*, vol. 14, no. 2, pp. 205–223, 2000.

[38] Z. Duszak and W. Koczkodaj, "Using principal component transformation in machine learning," in *Proceedings of International Conference on Systems Research, Informatics and Cybernetics, Baden-Baden Germany*, pp. 125–129, 1994.

[39] X. Zhu, S. Zhang, Z. Jin, Z. Zhang, and Z. Xu, "Missing value estimation for mixed-attribute data sets," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 1, pp. 110–121, 2011.