TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies

Parham Shams Ghahfarokhi 165642

ESTIMATION OF CENTRAL AORTIC PRESSURE WAVEFORM FROM RADIAL ARTERY ELECTRICAL BIO-IMPEDANCE USING CURVE FITTING AND NEURAL NETWORK APPROACHES

Master's thesis

Supervisor: Andrei Krivosei
Ph.D.

Tallinn 2020

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Parham Shams Ghahfarokhi 165642

Tsentraalse aordi rõhu lainekuju hindamine kasutades radiaalarteri elektrilise impedantsi kõvera sobitamist ja närvivõrgul põhinevat lähenemist

Magistritöö

Juhendaja: Andrei Krivosei
Ph.D.

Tallinn 2020

Author’s declaration of originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Parham Shams Ghahfarokhi

02.01.2020

Abstract

In the context of an increasingly connected world of medical science and Information Technology (IT), many projects are defined in a multidisciplinary fashion. There are reasons behind this connection such as accuracy, reliability, and speed of decision making. Scientists are trying to automate the decision which is made by computers. Central Aortic Pressure (CAP) and waveform convey invaluable information about the cardiovascular system state, but direct measurements include invasive methods. Peripheral pressure can be measured non-invasively, and although it often differs substantially from central pressures, it may be mathematically transformed to approximate the latter.

Therefore, this thesis focuses on the evaluation of the predicted model and the quality of the prediction, specifically concerning new methods and optimizing this method. The answers to some central questions are to be provided:

- How can the value of CAP be estimated, given the value of EBI is available?
- What is the prediction quality of the estimated CAP?
- How can the prediction quality of CAP be improved, given the original CAP is available?
- How can the prediction quality of CAP be improved, given the original CAP is not available?

The answer to these questions lies in using mathematical methods (including regression), machine learning algorithms, and evaluation metrics.

This thesis is written in English and is 98 pages long, including 5 chapters, 59 figures, and 21 tables.

Annotatsioon

Tsentraalse aordi rõhu lainekuju hindamine kasutades radiaalarteri elektrilise impedantsi kõvera sobitamist ja närvivõrgul põhinevat lähenemist

Meditsiiniteaduse ja infotehnoloogia üha enam ühendatud maailma kontekstis on paljud projektid defineeritud multidistsiplinaarsel viisil. On olemas kindlad seosed täpsuse, usaldusväärsuse ja otsuse kiiruse vahel. Teadlased üritavad automatiseerida arvutite poolt vastu võetavate otsustuste protsessi. Otsene arteriaalne rõhk (CAP) ja rõhulaine kuju annavad hindamatut teavet südame-veresoonkonna süsteemi seisundi kohta, kuid nende puhul on tegemist invasiivsete mõõtmistega. Perifeerset rõhku saab küll mõõta mitteinvasiivselt, kuid antud tulemused enamjaolt erinevad märgatavalt invasiivselt mõõdetud rõhkudest. Samas võib perifeerse rõhu tulemusi matemaatiliselt ümber kujundada ja läbi selle lähendada otsestele mõõtetulemustele. Seetõttu keskendub käesolev töö prognoosiva mudeli hindamisele ja prognoosi kvaliteedile, eelkõige uute meetodite ja nende optimeerimise vaatenurgast. Antud töö vastab järgnevatele kesksetele küsimustele:

- Kuidas saab otsese arteriaalse rõhu (CAP) väärtust tuletada, arvestades, et elektrilise bioimpedantsi (EBI) väärtus on olemas?
- Mis on otsese arteriaalse rõhu ennustuse kvaliteet?
- Kuidas saab parendada otsese arteriaalse rõhu ennustuse kvaliteeti, kui on olemas esialgne arteriaalse rõhu mõõtetulemus?
- Kuidas saab parendada otsese arteriaalse rõhu ennustuse kvaliteeti, kui esialgne arteriaalse rõhu mõõtetulemus puudub?

Antud küsimused saavad vastuse tänu matemaatiliste meetodite (sealhulgas regressioon), masinõppe algoritmide ja hindavate mõõdikute kasutamise.

Lõputöö on kirjutatud inglise keeles ning sisaldab teksti 98 leheküljel, 5 peatükki, 59 joonist, 21 tabelit.

List of abbreviations and terms

AI	*Artificial Intelligence*
ANN	*Artificial Neural Networks*
CAP	*Central Aortic pressure*
CF	*Curve Fitting*
CFTOOL	*Curve Fitting Tool*
CVD	*Cardiovascular disease*
DM	*Data Mining*
DV	*Dependent Variable*
EBI	*Electrical Bio-Impedance*
FF	*Feed Forward*
GD	*Gradient Descent*
GUI	*Graphical User Interface*
ICU	*Intensive Care Unit*
IT	*Information Technology*
IV	*Independent Variable*
ML	*Machine Learning*
NN	*Neural Networks*
RL	*Reinforcement Learning*
RMS	*Root Mean Square*
RMSE	*Root Mean Square Error*
SL	*Supervised Learning*
SML	*Supervised Machine Learning*
TUT	*Tallinn University of Technology*
UI	*User Interface*
USL	*Unsupervised Learning*

Table of contents

List of figures

List of tables

1 Introduction

1.1 Problem Statement

The cardiovascular system is one of the principal systems of the human body whose health can, directly and indirectly, lead to human well-being, and whose lack of health may cause serious direct or indirect threats to human's health or even life. Hence, it is of great importance to be able to understand what the indicators of cardiovascular health are, and how they are measured. Once it is known that measurement is key to the health of the cardiovascular system, other important questions arise, like:

- How can the precision of the measurement of Central Aortic Pressure (CAP) be increased?

- How can the CAP measurement be facilitated, so that more people, in different situations, can have access to the measurement services/equipment?

- How is this possible to decrease the risk of performing monitoring and measurement of cardiovascular system parameters?

If these questions are answered in a scientific way, with the correct knowledge, proper argumentation and scientifically-provable facts, a vast multitude of people can benefit from what comes after that. If a suitable answer is found, not only it directly helps people at risk of danger, but also it helps in other areas, like providing people with bigger motivation to track their health, and monitor their well-being long before they are at serious risk.

In this thesis, the main effort is on providing a method that can be used to measure one of the important cardiovascular health indicators (Central Aortic Pressure), in a risk-free procedure (answer to the third question above). This method can incorporate a low-trouble technique to gain input Electrical Bio-Impedance (EBI) data (answer to the second question). Furthermore, maximum effort is conducted in optimizing the precision of the measurement using mathematical and machine learning processing approaches (answer

to the first question). To answer these questions, we may think of a method which proves to be accurate enough for the estimation of the , and which can be implemented with little difficulty, in a risk-free procedure For developing such a method, four main questions come to mind, which are going to be the basis of the work in this thesis:

- What is the relation between CAP and EBI of a person when both CAP and EBI of the person are available?
- How to improve the goodness of the estimated relation between CAP and EBI?
- How to predict the value of CAP using the value of EBI when the CAP of the person is not available?
- How to improve the accuracy of estimated CAP using the value of EBI?

In the next chapters, the effort is on answering these questions.

1.2 Motivation

People with serious illnesses or injuries who require constant care from medical staff are posited in an Intensive Care Unit (ICU). Because patients in intensive care are often unconscious or, if not unconscious, are incapable of alerting medical staff during emergencies, they are monitored by electronic bedside equipment which ceaselessly records signals such as electrocardiogram and arterial blood pressure waveforms. Besides emergencies, Central Aortic Pressure (CAP) is a very important factor for physicians which convey various information related to the cardiovascular status, but direct CAP measurements are of invasive nature, and this can be counted as a limitation toward portability and ease of implementation of these methods [1].

The scientific association has acquired data using monitoring and acquisition equipment attached to the body of patients under medical supervision and has made available databases of medical information by a methodical recording of that data. These databases give us an indispensable chance to dig the history of medical information and realize patterns that can be used to make statements about the future development of patients. Being able to foretell a patient's future state would allow doctors to administer preventive treatment, thus saving lives, improving the use of financial resources and promoting human health [2].

One peculiar target that researchers have is to design algorithms and patterns using data from such databases that could forecast how the central aortic pressure of a patient in an ICU behaves. CAP prediction and the prediction of its waveform is of great interest to physicians and ICU medical personnel, because it makes them being able to recognize the patients that are at hazard of developing abnormally. CAP investigation and prediction would be useful in various ways:

- It would decrease ICU function costs and increase ICU function efficiency.

    1. Medical personnel could concentrate on monitoring sick people who are exclusively at risk if fewer cases of sudden problems are reported (Sudden medical problems which could have been solved before any treatment was needed).

2. Prevention is in most cases less costly than intervention. Once a person diagnosed with medical problems, has a critical state recorded, such as shock, an expensive life-saving intervention is going to be necessary. The cost depends on many factors, but in any case, this will be the cause for later expenses (time and money expenses) [1].

- Physicians may have extra data available to them when making decisions on how to remedy sick people.

Knowing this, we should notice that it is possible to tap predictive data within medical records databases, allowing this project to estimate how the central aortic pressure of a patient will change. Because central aortic pressure is influenced by many various factors, it is indispensable to create predictive models using a grand and diverse set of data so that the model can correctly recognize the status of a new person when predicting that person's central aortic pressure [3].

The development of practical supervision methods to assess patterns and trends correlated with major cardiovascular diseases improves the debarment and reconnaissance of cardiovascular risk factors, as well as the expansion of new tactics for better CAP analysis.

As mentioned above, the CAP waveform conveys significantly important information about cardiovascular status, but direct measurements are invasive. However, circumferential pressures may be recorded noninvasively, and later, with mathematical methods, it is possible to measure CAP data from those. But even by using this method, central pressures often vary substantially from measured peripheral pressures. Also, still, this method is regarded as fairly-difficult concerning the convenience of measurement [4].In the figure below, different methods for the measurement of Central Aortic Pressure has been demonstrated. In this work, we have used EBI information as a representative of CAP values.

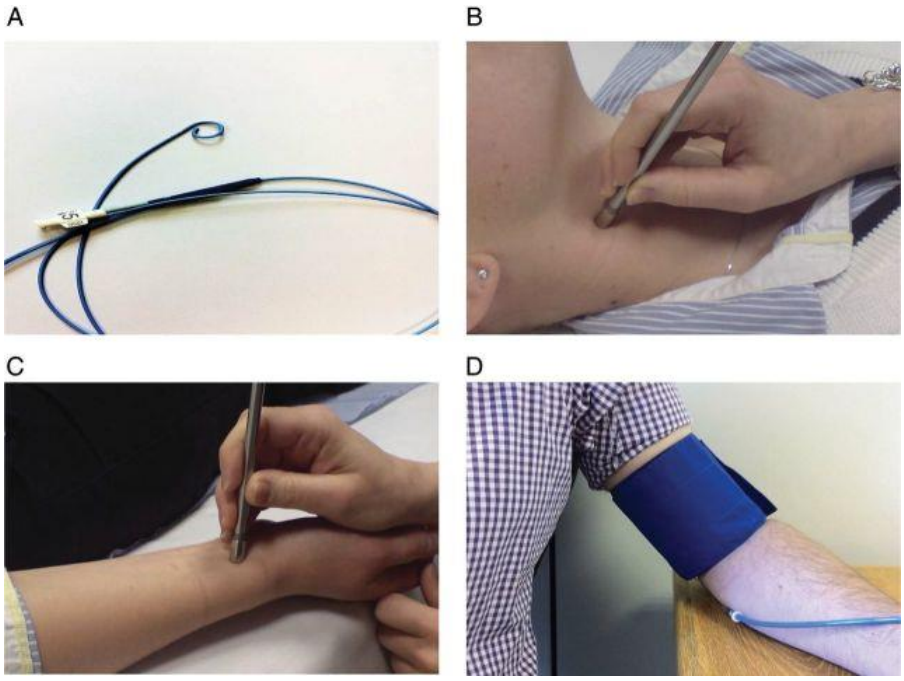

Figure 1 Different methods to measure the quantities under study [5]

Pressures of central aortic systolic and diastolic are determinants of cardiac loading and perfusion, and they vastly affect the cardiovascular system function. Knowledge about these pressures is mostly indispensable to meticulous monitoring and titration of interventions in disease states. As a further matter to the pressures themselves, the waveform of arterial pressure gives helpful information about systemic vascular rigidity, compliance, the reflection of waves, and other aspects of value in the bedside monitoring. At the same time, broad usage of such analysis has been prolonged by the requirement for invasive CAP measurements.

Cardiovascular diseases (CVDs) are the prominent mortality cause universally, and yearly deaths are associated with them, and we should consider that the number of these reported incidents are increasing on a year to year basis. Obviously, deeper comprehension of the human cardiovascular system would be useful for the remedy of such diseases.

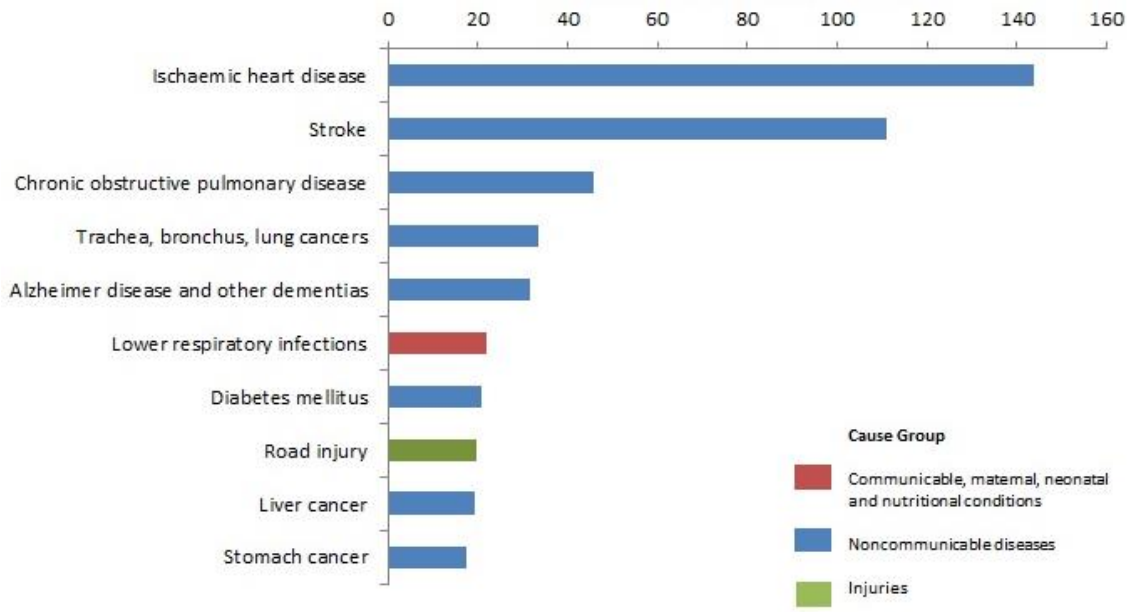

Figure 2 Top 10 causes of deaths in upper-middle-income countries in 2016 [6]

The changes in CAP and in blood pressure may provide vital clinical information about the exclusive patient's underlying physiologic regulation and the cardiovascular system. As a result, reliable algorithms, capable of forecasting CAP, may drastically help accurate finding of circulatory disease.

Therefore, the main focus and approach in this research is on finding methods (and models) that can make us capable of predicting CAP based on EBI information. If the EBI is the input, the CAP will be the output. There are many different approaches for finding a relation between a set of given inputs and their corresponding output. In this thesis, the usage of two popular methods has been considered, namely "regression" and "neural networks". Then, these methods are assessed and based on those asssessments, improvements are suggested and applied, until a suitable accuracy is achieved.

1.3 Expectations, Outputs, and Scope

According to the explanations in the previous sections, in this thesis, we expect to achieve a model that is capable of predicting the value of CAP, which is a desired cardiovascular health indicator, based on the value of EBI. Measuring the value of CAP includes difficulties and challenges, while measuring the value of EBI is an easier and less risky task. The effort is to come up with an initial model that is able to do the CAP prediction with minimum possible error. Hence, the output of this thesis will be methods that generate models that can predict CAP data based on EBI values measured through a medical instrument.

To sum up, the input is registered chronological EBI data using medical instrument by certified people, stored as datasets. This data is then loaded into Matlab and is preprocessed and then processed using the algorithms and methods of this thesis, and the output is predicted CAP data which is stored. **This whole process from the loading of data to saving the CAP output will be the subject of the effort in this thesis, and so is the scope of the thesis.** Finally the saved data should be registered in human-readable form and is passed to certified professionals for further study and investigation.

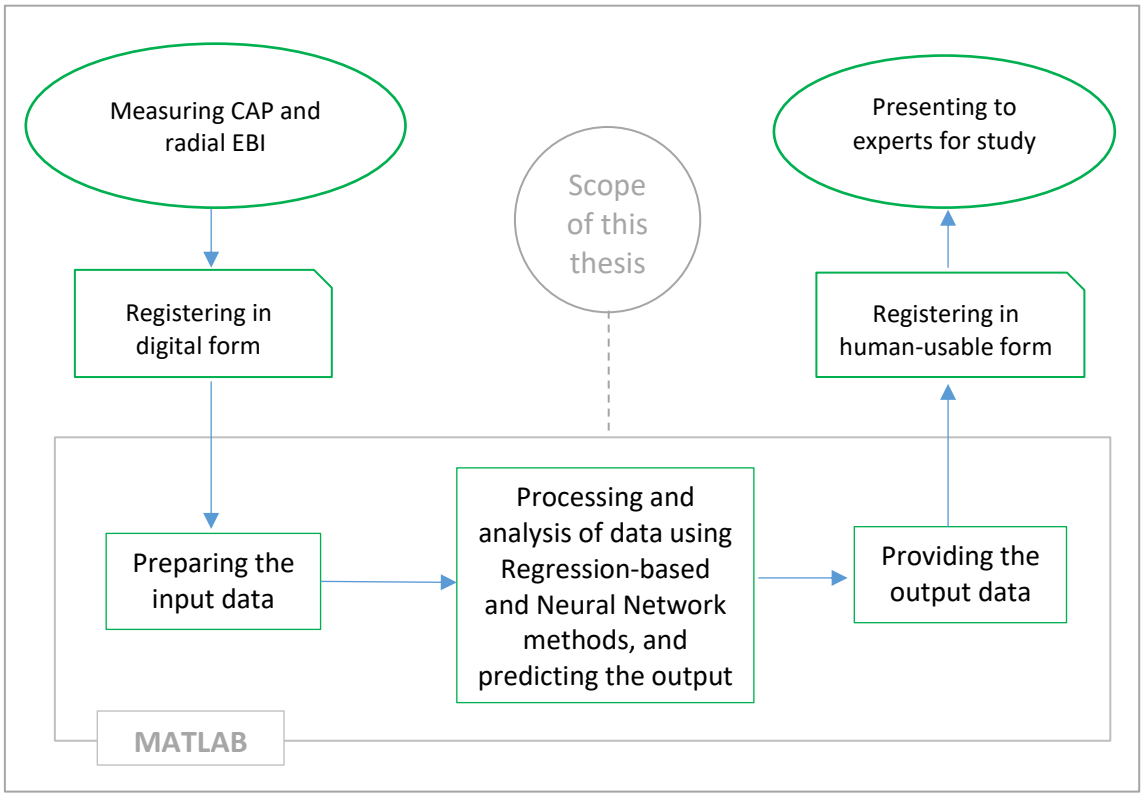

Figure 3 The overview of the flow of data before, during and after the preprocessing and processing

2 Theoretical Background

2.1 REGRESSION AND MODEL BUILDING

Regression analysis is a sub-part of statistics, which is used for investigating and modeling the relation between variables. Regression's applications are numerous and happen in almost every part of sciences such as engineering, chemical sciences, physical sciences, management, economics, life, and biological sciences. As a matter of fact, among the statistical technique, the regression technique is the most widely used technique.

The equation below is called a first-order polynomial equation. This display clearly suggests a relationship between *x* and *y*; in fact, the impression is that the data points generally, but not exactly, fall along a straight line [7]. If we let *y* represent delivery time and *x* represent delivery volume, then the equation of a straight line relating these two variables is

$$y = a1x + a0 \qquad (1)$$

where $a0$ is the intercept and $a1$ is the slope. Now the data points do not fall exactly on a straight line. Let the difference between the observed value of *y* and the straight line ($a0$+ $a1$.x) be an error ε. It is convenient to think of ε as a statistical error; that is, it is a random variable that accounts for the failure of the model to fit the data exactly. The error may be made up of the effects of other variables on delivery time, measurement errors, and so forth. Thus, a more plausible model for the delivery time data is:

$$y = a1x + a0 + e \qquad (2)$$

Equation (2) is called a linear regression model. Customarily x is called the independent variable and y is called the dependent variable. However, this often causes confusion with the concept of statistical independence, so we refer to x as the predictor or regressor variable and y as the response variable. Because Equation (2) involves only one regressor variable, it is called a simple linear regression model [8].

2.1.1 Gradient Descent

Frequently, in data science, we'll be trying to find the best model for a certain situation. And usually "best" will mean something like "minimizes the error of the model" or "maximizes the likelihood of the correctly predicted/explained data." In other words, it will represent the solution to some sort of optimization problem.
This means it is needed to solve several optimization problems. In particular, it is needed to solve them from scratch. Our approach will be a technique called gradient descent, which lends itself pretty well to a from-scratch treatment.

The gradient descent algorithm is used for minimizing a convex function and to analyze its convergence properties [9]. In order to minimize a convex function, it is needed to find a stationary point. One possible approach is to start at an arbitrary point and move along the gradient at that point towards the next point and repeat until (hopefully) converging to a stationary point. This point is claimed as the extreme point of the curve while found [10].

2.2 Machine Learning

With the huge amounts of data in digital form, the necessity for automatic procedures to carry out such data analysis emerges. One of the aims of machine learning (ML) is to find models, which are based on meaningful computable relation among data points, and then to utilize the generated models to anticipate the output of all future data. Machine learning is then closely linked to the field of statistics and data mining (DM) [11]. There are a variety of methods conducted by computers for learning, but only a limited number of them may be used for analysis in this thesis. Neural Network (NN) method is chosen as the best option and is further explained in the coming sections.

Machine learning is frequently divided into three key types, supervised learning (SL), unsupervised learning (USL) and reinforcement learning (RL). Also one specific type of learning is sometimes considered, which has characteristics similar to both SL and USL, it is called semi-supervised learning (Figure 5)

In the supervised learning method, the aim is to learn labeling from X as inputs to Y as outputs, given a set of input-output pairs D = $\{(x_i, y_i)\}_{i=1}^{N}$. Here D is called the training set, and N is the number of training examples [11].

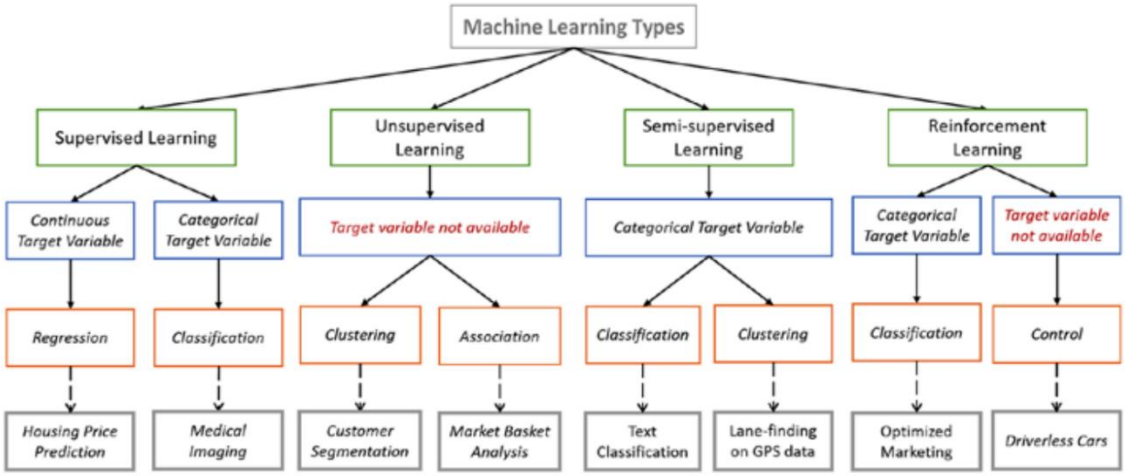

Figure 4 Map of different machine learning methods [12]

In the simplest setting, each training input x_i is a D-dimensional vector of numbers, representing, say, the height and weight of a person. These are called features, attributes or covariates. In general, however, x_i could be a complex structured object, such as an image, a sentence, an email message, a time series, a molecular shape, a graph [13].

Similarly, the form of the output or response variable can in principle be anything, but most methods assume that y_i is a categorical or nominal variable from some finite set, $y_i \in \{1 \dots, C\}$ (such as male or female), or that y_i is a real-valued scalar (such as income level). When y_i is categorical, the problem is known as classification or pattern recognition, and when y_i is real-valued, the problem is known as regression. Another variant, known as ordinal regression, occurs where label space Y has some natural ordering, such as grades A–F [11].

The second main type of machine learning is the unsupervised learning approach. Here we are only given inputs, $D = \{x_i\}_{i=1}^{N}$, and the goal is to find "interesting patterns" in the data. This is sometimes called knowledge discovery. This is a much less well-defined problem, since we are not told what kinds of patterns to look for, and there is no obvious error metric to use (unlike supervised learning, where we can compare our prediction of Y for a given X to the observed value).

2.3 Neural Network

Artificial Neural Network (ANN) or simply, Neural Network, is a structural systemic approach for linking a set (or sets) of inputs to a collection of (or a single) with the aim of finding a model to be able to predict future outputs of the system based on the input. ANNs work by mimicking the behavior of humans' neural network system, and inside them, the neurons and the synapses as part of the neural networks perform a wide range of mathematical calculations. Neural networks formation has origins in Artificial Intelligence (AI) and is being widely used in the world today to make humans capable of making relations between variables (or sets of variables) using the prior information they are fed, with proper accuracy [14]. This makes them a validated candidate for the purpose of our work since in this work a relation between CAP and EBI needs to be formulated, so that we can obtain values of CAP based on the values of EBI.

Regarding the approach, there are multiple approaches for the neural network to take, and the decision of choice can be depended on a number of factors. For instance, if there is no access to the dependent variable (i.e. outputs, labels, etc.), we can not make use of a supervised approach, since we have to feed the dependent variable to the system in this kind of approach and its sub-approaches. Other approaches include unsupervised learning (USL), Reinforcement Learning (RL), and some other approaches which work in between these methods. The method chosen to be used as the approach of this thesis is supervised learning (Figure 5).

Layers in the neural networks are sets of neurons. The data is provided through an input layer of the neural network. The hidden layer is where the computations take place. The output layer is where the finished computations of the neural network is presented for our further usage. The broad question on using suitable number of neurons and hidden layers are subjective to the complexity of the computation that is expected out of the neural network.

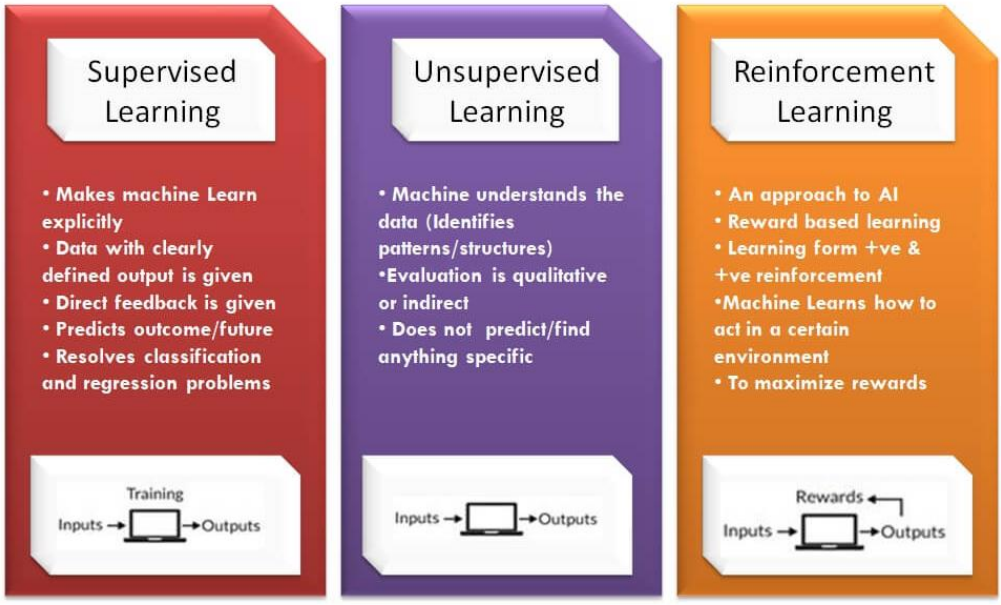

Figure 5 Specifications of different machine learning methods [15]

The complexity of the neural networks gives birth to a multitude of architectures of Neural Networks. Typically, each neuron is connected to every other neuron in the next layer, which is also known as Feed Forward (FF) neural networks. It is observed through various experiments that by connecting neurons to other neurons in certain patterns, better results can be obtained. Recurrent Neural Networks were created to address the flaws in neural networks that didn't make decisions based on provided knowledge. Typically, a Neural Network had learned to make decisions based on the context in the training model, but once it was making decisions for use, the decisions were made independent of each other [16]. Regardless of various implementations and approaches, the intention is always the same. In a Convolutional Neural Network, the connection between the layers appears to be random. The reduction of the number of parameters that are needed is achieved through the synapse setup. Convolutional Neural Networks are most commonly used in image processing. According to the complexity of our data, and our purpose, FF is used throughout the NN section of this thesis.

Regarding the accuracy of artificial neural networks, we need to take into account that as a general rule, the more the number of inputs, the more accurate the results are, but considerations about other factors affecting accuracy should also be respected and cared for.

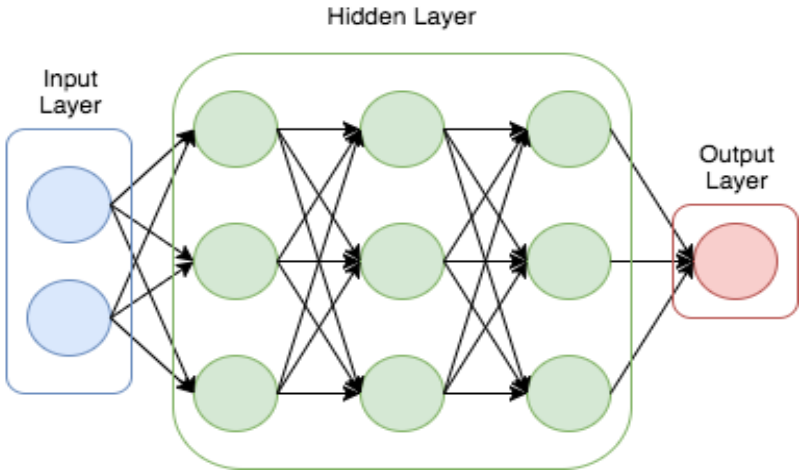

Figure 6 A Feed Forward (FF) neural network with 3 hidden layers, and 3 nodes in each, and 2 input nodes, and an output node [17]

2.4 Quantity under Process (Cardiovascular Health Indicator)

As mentioned in the previous chapter, Central Aortic Pressure (CAP) is a very important indicator of cardiovascular status which conveys various information related to cardiovascular health for physicians. It was also pointed out that direct CAP measurements are of invasive nature, and this can be counted as a limitation toward portability and ease of implementation of these methods. Therefore, we choose another indicator which is measured with more ease and less risk, but at the same time, can be used to predict CAP with good approximation. This indicator can be the blood pressure of the radial artery. As Figure 7 shows the relation between these two indicators (central aortic pressure, and radial artery pressure), these two indicators have a meaningful and close relation. A technique to derive a blood pressure-related waveform from the radial artery is bioimpedance, which is chosen as the method for recording the data that we use throughout this thesis.

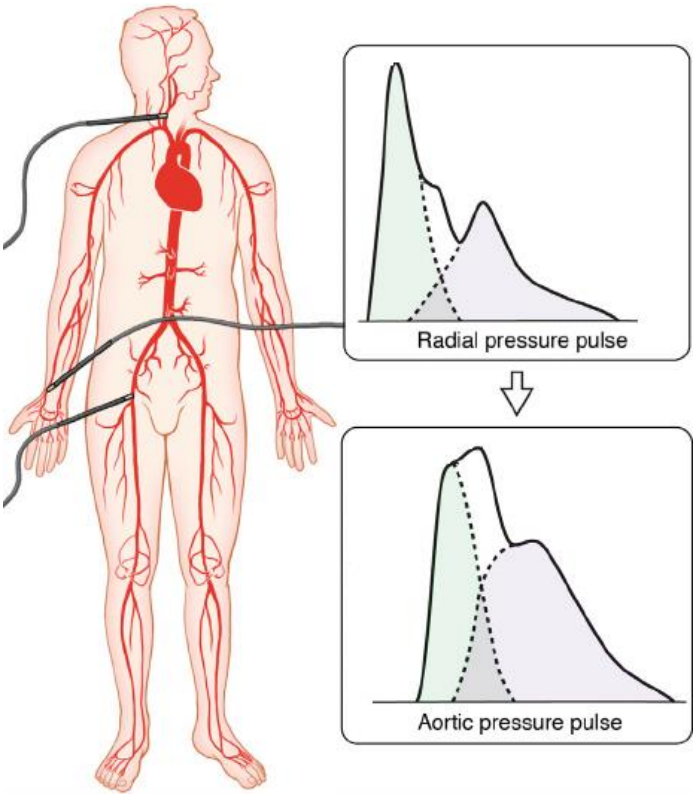

Figure 7 a schematic representation of the clinical gold-standard pulse-wave velocity (PWV) measurement in the carotid-femoral region [18]

For this method to work small current is applied to the site of interest through electrodes, and a voltage difference is measured (Figure 8). Bioimpedance is calculated from the exerted current and measured voltage, which gives us the change of the impedance during the cardiac cycle. With each heartbeat, the volume of the blood changes under the electrodes, and it reflects in the impedance curve which corresponds to blood pressure waveform. The device consists of four electrodes placed on the wrist, and it detects the blood volume fluctuation of the radial artery as the variations in Electrical Bioimpedance (EBI) to acquire the volume pulse wave. Figure 8 demonstrates useful information about the implementation of the hardware of EBI, the sectional view of the site of measurement on the human body, and the mechanism of the method [19].

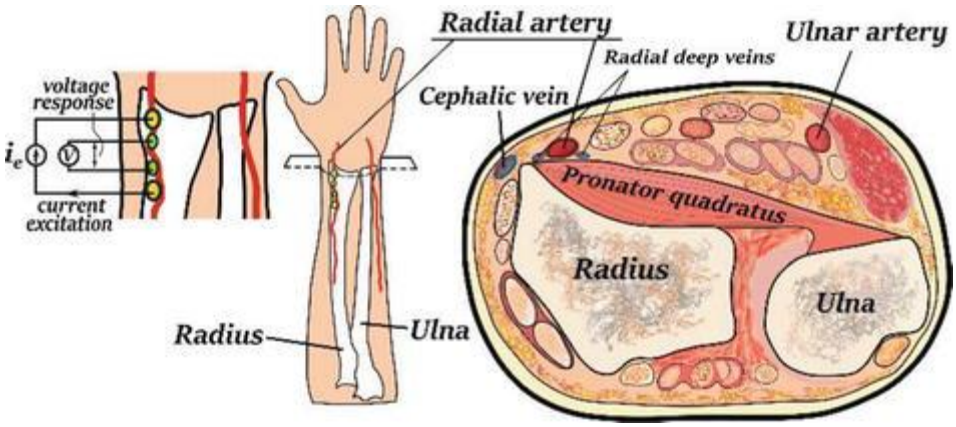

Figure 8 Electrode placement for measuring EBI in the wrist area[19]

Figure 9 shows the resemblance between EBI and CAP, and the meaningful relation they have is observable.

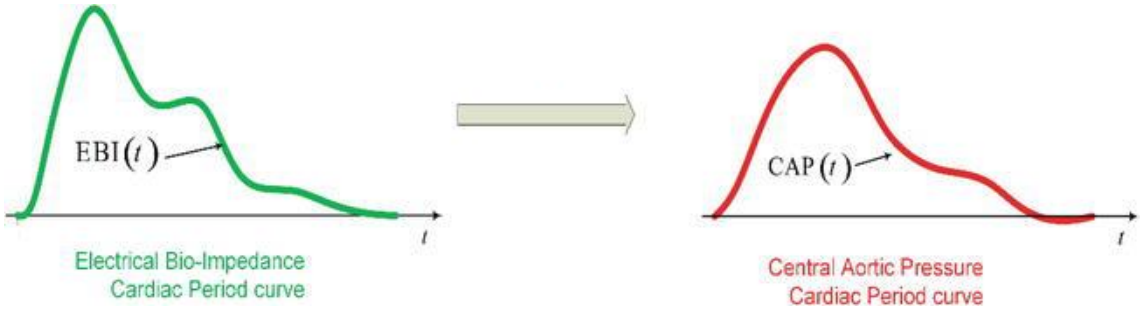

Figure 9 Demonstration of the CAP cycle reconstruction from the EBI cardiac cycle waveform [19]

Bioimpedance sensing does not need strong pressure on the artery, as it is with tonometry, only a permanent electrical contact is required. Therefore, the worry of affecting the blood circulation with the measurement procedure falls off in a large extent—the measurements become more passive. A pilot study presented in 2011 [20], where they provided first experimental evidence that electrical impedance tomography (EIT) is capable of measuring pressure pulses directly within the descending aorta. Their research measures the impedance on the thorax, not on the arm or wrist, but the study supports, nevertheless, the idea of central aortic pressure assessment with bioimpedance [19]. A number of researchers have had analogs thoughts and promising results, and a number of scholars have had practical results in the improvement of the EBI-based measurements of aortic pressure curve [19]. At the same time, the development of corresponding devices for clinical practice is still not significant. Nevertheless, the interest to get a blood pressure

measurement device that relies on bioimpedance is still very topical. Especially, when a big corporation such as Microsoft Technologies, got their patent published in 2018 for a wearable system that determines a pulse waveform based on bioimpedance measurement device together with pressure transducer [21].

3 Methodology and Results

3.1 Overview of procedures

3.1.1 Dataset description

The dataset used for calculations and analysis consists of recorded EBI and CAP data of 14 people which are registered over time domain, and the range of time is approximately equal to one complete period of CAP measurement. The time range can change according to different people and different states of each person since the heartbeat rate differs in the different people and conditions. The dataset is not completely ready for processing as it is, so we need to do some preprocessing on it to make it ready. Further description of the dataset can be found in section 3.2 (Overview of Dataset).

3.1.2 Data preprocessing

As mentioned in the previous section, data needs preprocessing before it can be used for the purpose of analysis and calculation. In many of the works which are engaged with finding a relation between two variables, or carrying out a prediction of some kind using the past data, preprocessing is a key step since interpretation and usage of data are different in different methods. Another reason which gives rise to the need for preprocessing is the fact that the data we acquire is not gathered correctly at some points and those points need to be corrected or eliminated from the dataset in order to be able to proceed with calculations. The following steps are taken to do the preprocessing as mentioned in section 3.3:

- Corrections related to outlier data points
- Dataset splitting
- Scaling
- Transformation (shifting)

It should be pointed out that for each of the methods which are used, a specific preprocessing procedure is needed which includes some or all of the above-expressed steps.

3.1.3 Data processing (making the model)

When the data is preprocessed and ready, a meaningful relation should be defined between the input and the output, so that we are able to construct a model based on that relation, which gives us the possibility of predicting the output of future input data.

There are many different approaches for finding a relation between a set of given inputs and their corresponding output which is also given. In this thesis, the usage of two popular methods has been considered, namely “regression” and “neural networks”.

Regression works by fitting a curve through all data points which is supposed to have the closest distance to all the data points combined. One of the best tools which has been developed and optimized for the purpose of regression is Matlab’s native Curve Fitting toolbox (CFtool). This toolbox gives the possibility of adding a certain input, and its corresponding output, and choosing the order of the curve to be fitted, and with all this data and configuration, the toolbox implements mathematical procedures to gain the best-fitted curve to user’s data. We can then use this curve to analyze and test how good of a fit we have achieved, and if the result is desirable, we can use that curve as the basis of a model for prediction of future datasets. In section 2.1 (Regression and Model Building), the theoretical background for regression (and consequently CFtool) is provided, and in section 3.5 (Curve Fitting using CFtool in Matlab) the usage of CFtool as a functional toolbox has been explained. It should be noted that CFtool has been used in this work to find the fitted polynomial curves of different orders (between 1 and 9).

For reaching a useful solution and having a clearer image of the steps taken, it is needed to make distinction between different methods that were used. Since data processing is the core of a method on which most of functionality is depended, we classify methods based on different data processing approaches that are taken in those methods. The next figure shows the major classification of methods (data processing methods), and the following paragraphs will describe how this classification has been reached.

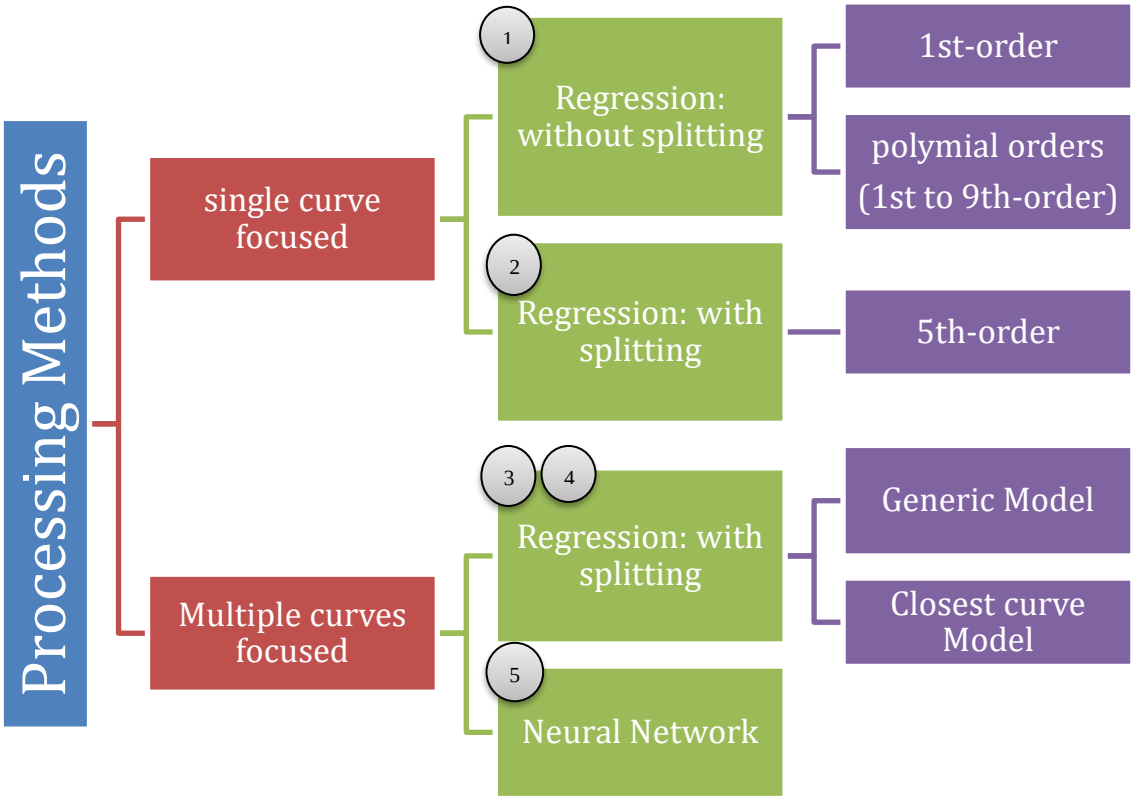

Figure 10 Overview of the 5 different methods used in this thesis

In the initial section of the work, the focus was only on one single patient's recorded data points as the subject of regression, and the curve was derived based on such data, and the result was an optimized form of such a curve. Methods 1 and 2 have single-curve focus. In the second section of the work, the data of all patients were considered and finally, the resulting model was derived by focusing on a multitude of curves. Methods 3, 4 and 5 have multiple-curve focus. So, our methods are split into two major categories (methods 1 and 2 are placed in one category, and methods 3, 4 and 5 lie in the other category).

Another point of distinction is whether our dataset is splitted (to increasing and decreasing sections) or is considered as a whole. Based on this criterion, methods with single-curve focus split into two methods (method 1 and 2). Method 1 uses regression with a dataset as a whole. Method 2 uses regression on increasing segment of data and decreasing segment data separately. Method 2 was developed because the results of method 1 were not satisfactory enough. The reason behind this was that data shows a large amount of dispersion at the peak (where increasing and decreasing sections of the curve meet). So, from this point on, all methods use splitted data of each patient.

In the end, methods of multiple-curve focus were divided into some groups based on the mathematical tools which were used. Methods 3 and 4 help to derive the model by the use of regression, and method 5 uses machine learning at its core to find a relation between X and Y.

Method 3 provides a solution to our problem by taking into account the data of all patients (all 14 patients) in one place. This is achieved by taking the average of 14 estimated curves of all patients (small intervals) and merging these averages to acquire one final curve.

Method 4, too, provides a model by taking into account the data of all patients' records in one place, but the mechanism of the method is totally different from method 3. In method 4, we take the mean value of the independent variable (X, A.K.A. EBI) of all data points of 14 estimated curves (each curve corresponds to one patient's records), and subsequently, we have 14 mean values. Then for every new set of X values of future patients, we calculate its mean value, and find the closest mean value of past data to the current data. Then we use this curve from past data (closest curve), to predict the current data's CAP values.

As pointed out, method 5 is based on machine learning and neural networks. Machine learning methods try to find a pattern between the dependent variable (Y) and independent variable(s) (here it is X). Since we have the expected output (dependent variable or CAP) of the data for the patients, we may go after methods of supervised learning, which consider dependent and independent variables simultaneously. One of the most successful methods of learning (supervised or unsupervised) for continuous data is neural networks (NN), and we are going to benefit from this method for the purpose of processing and model derivation.

3.1.4 Testing

After a model is achieved in any of the methods, we need to verify the prediction quality of the model through testing. For this to take effect, usually the process is: the real CAP data of one patient, which has not been used for generating/training of the model, is compared against the CAP data which is predicted by the model. The difference between the two mentioned values define the result of the test, and consequently defines the prediction quality or accuracy of the model.

3.2 Overview of dataset

The dataset mainly consists of recorded data of human's Electro BioImpedance and Central Aortic Pressure values through time. In Table 1, the characteristics of the dataset used in this study are presented. The dataset was provided by a group of researchers at Taltech whose focus of research is on signal processing and its characterization. The dataset was generated by logging instrumentation data of medical equipment during the monitoring of different people's central aortic pressure and electrical bioimpedance waveform [22].The dataset was provided in different log files (each log file corresponds to one individual's medical data). There are 3 variables X, Y and T that correspond to radial artery EBI waveform, Central Aortic Pressure (CAP) and time respectively. The variable labeled Y is the observation outcome with continuous values, and it is the variable that we would like to predict.

Table 1 The first 7 points of recorded EBI and CAP through time included in a dataset

EBI	**CAP**	**Time (ms)**
X	**Y**	
-10826	-17552	0,000
-10827	-17549	0,005
-10827	-17543	0,010
-10827	-17537	0,015
-10828	-17529	0,020
-10829	-17520	0,025
-10829	-17509	0,030

There are many regression models having strict requirements to scale the variables and to resolve the skewness in the variables before modeling. There are many ways to scale the amplitude which can be applied to manage these transformations, such as finding the maximum value of the variable, or such as dividing all samples by maximum value and shifting the minimum to the origin of coordinates which is (0,0). Before starting any process on the dataset, it is mandatory to scale the dataset, then the dataset is ready for the processing stage. Figures below demonstrate the original data (Figure 11 and Figure 12) and data after scaling (Figure 15 and Figure 16). As mentioned, the training set consists of the input to the regression algorithm that is the data gathered from radial artery

EBI or electrical bioimpedance waveform labeled as X. Each of these (EBI) values are assigned to one value of central aortic pressure which is labeled as Y. The aim is to predict the values of Y for a given X and to reconstruct the Y in the time domain. Dataset is visualized in different diagrams to provide a better understanding of the contained data.

For our implementation of the ordinary linear regression and non-linear regression of different order algorithms, the training set and testing set can be obtained by functional processing in Matlab, which gives us the capability to extract sub-samples and work with them flexibly. The number of samples for the below example in the training set is 139 samples. The outcome variables in the training set and testing sets are within the same range.

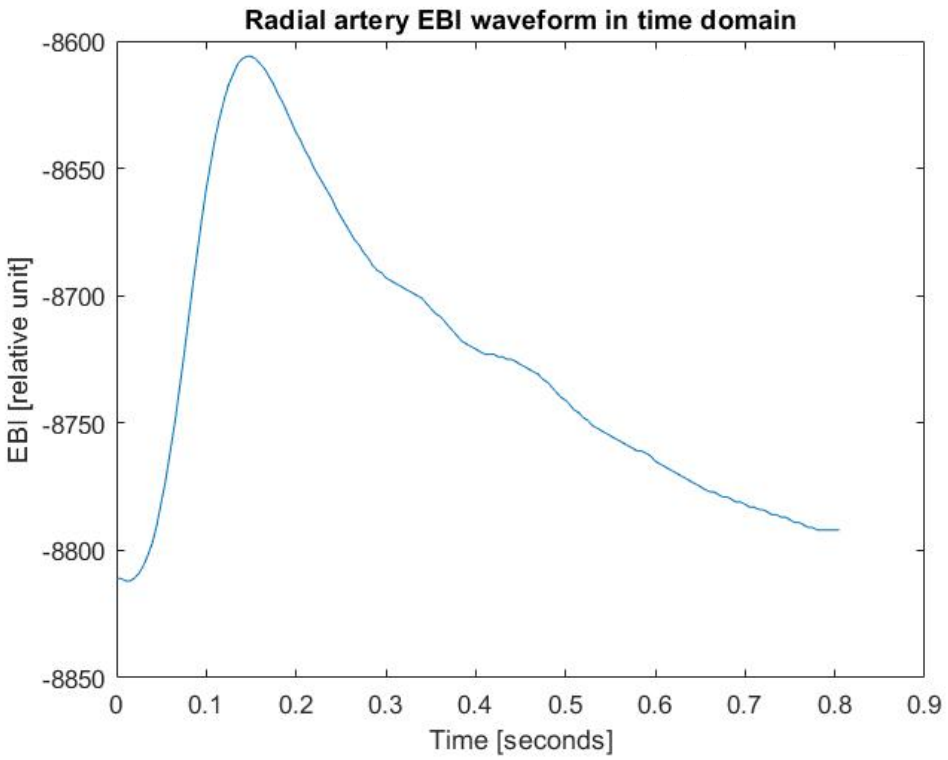

Figure 11 Radial artery EBI waveform in the time domain (unscaled)

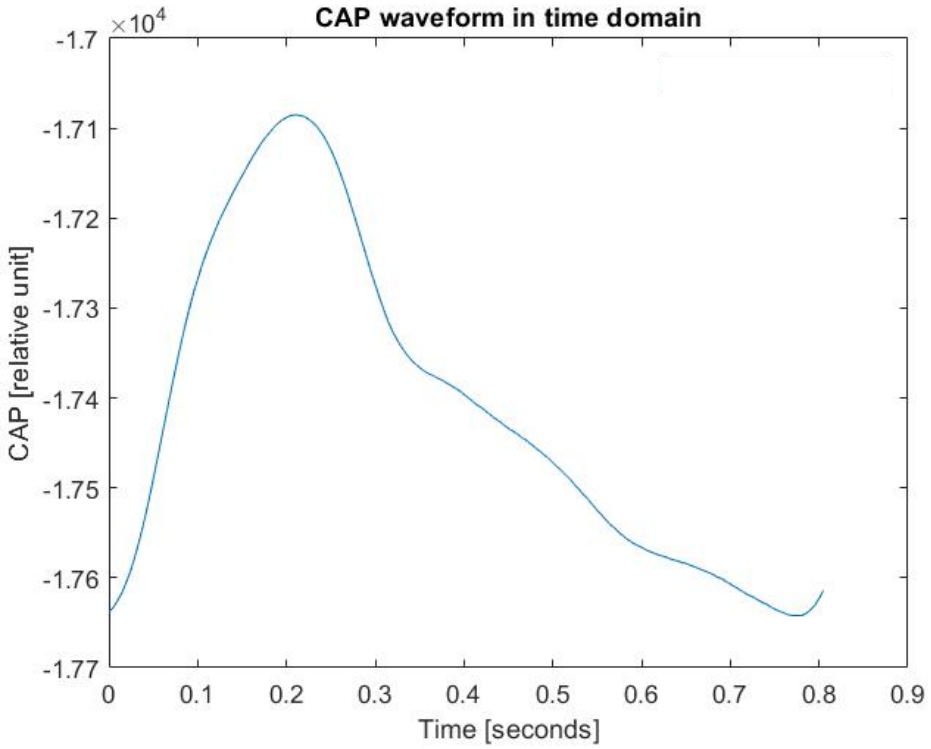

Figure 12 CAP waveform in the time domain (unscaled)

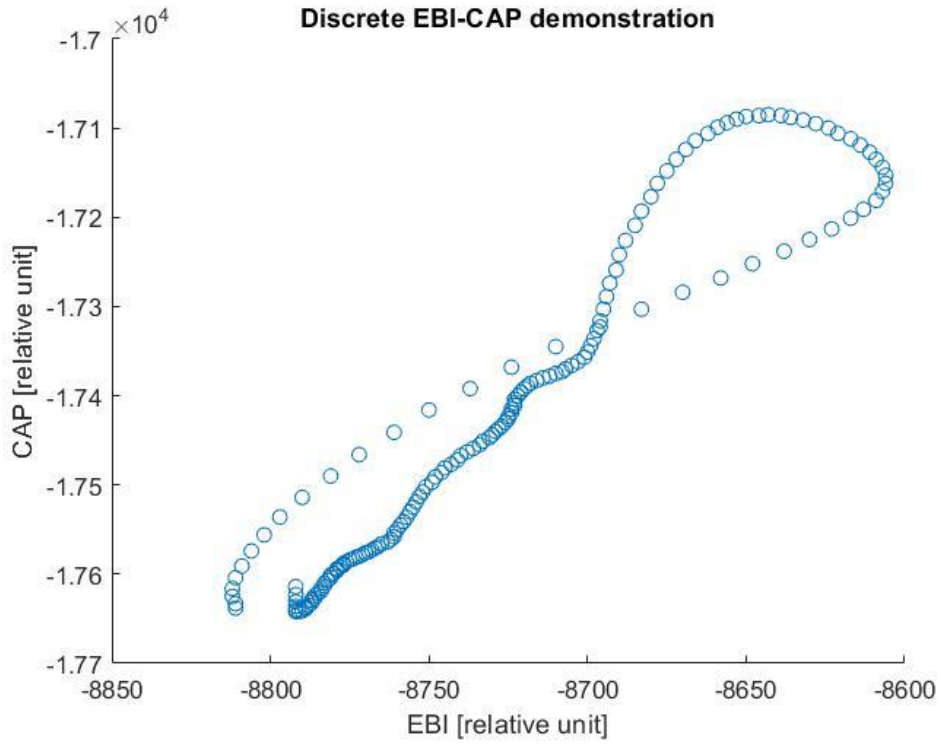

Figure 13 Discrete EBI-CAP demonstration (unscaled)

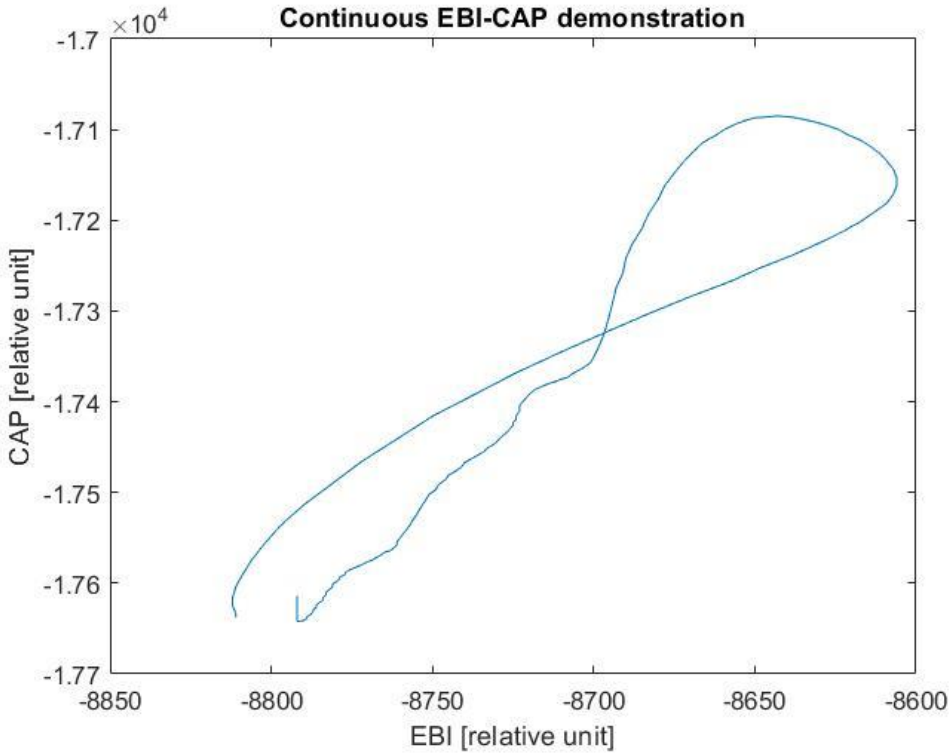

Figure 14 Continuous illustration of EBI-CAP

3.3 Data pre-processing

In this project, data pre-processing is of necessity since scaling, outlier correction, dataset splitting, and shifting are necessary for the regression to work correctly and come up with correct and suitable results that can be of benefit to the problem at hand. It needs to be mentioned that in most of the methods used, all of the named operations have been of necessity, but not in all of the methods. For instance, dataset splitting was not used at the start (in the first method), and the results of this method are calculated based on the non-splitted data. During the data pre-processing phase, the Matlab software was used for mathematical and statistical computations, as it is a flexible and powerful platform for such processes and has pre-made libraries and functions for the tasks we wanted to carry out in preprocessing phase.

In the following sub-sections, the use of each of the operations is discussed, and the reason why they are needed is mentioned. Moreover, a summary of how this has been applied in the case of the current project is provided.

3.3.1 Outlier correction

In our dataset, we have some irregularities at certain points. Irregular is used with the meaning of "not following the expected pattern". We expect EBI values to rise through time as the heartbeat period starts, and therefore the values of EBI should grow in time (from the start of one heartbeat to the maximum point). For various causes, like the recording device error, or the operator error, it might be the case that EBI values have a sudden fall after the start, as shown in the next Table 2 (in which EBI of data point No. 8 is clearly smaller than its former values). As we investigated different datasets of different patients, this problem was observed in the first 10 data points in some of the records. Therefore we have considered trimming the first few data points in cases that the EBI value suffers from a sudden fall. This helps in the analysis and calculation phase. It is notable that neither accuracy nor precision of the calculations is affected, because 10 points out of about 300 points is a small percentage of the whole.

Table 2 First EBI values of a dataset which do not chronologically increase, and may cause problem during the processing phase

No.	Time	EBI	CAP
1	0	-10826	-17552
2	0.005	-10827	-17549
3	0.010	-10827	-17543
4	0.015	-10827	-17537
5	0.020	-10828	-17529
6	0.025	-10829	-17520
7	0.030	-10829	-17509
8	0.035	-10830	-17496
9	0.040	-10829	-17483
10	0.045	-10828	-17468
11	0.050	-10825	-17452
12	0.055	-10821	-17435

3.3.2 Scaling

Scaling, in simple terms, can be defined as the process of adjusting values that are on different scales, to a commonly equal/equivalent scale. In data preprocessing, scaling is used to structure data in confined scales so that the difference in the value of different variables (in different datasets) does not cause misinterpretation and miscalculation errors. In other words, if we do not scale the values of different dependent and independent variables (CAP and EBI), since their range (maximum and minimum) vary noticeably across different datasets, the results are going to be affected and there is a high probability that we end up with incorrect final calculation results. For instance, in practice, we may have two datasets that have different maximum values, but when brought to scale, their maximum values are exactly equal (one of them is the scalar product of the other). If we use only these two datasets only to derive and calculate the CAP information with techniques involving averaging methods, we will end up with a curve which shows values between the two previously used dataset (non-scaled datasets). This result can not usually be scaled back to an applicable set of results by scalar multiplication and/or division, and this is why we need to scale datasets from the starts.

In the current section in the thesis, the focus is on how to scale the data, which data to scale, and what would be the results of scaling. As it is illustrated in Table 1 the data are not scaled. For the purpose of scaling in Matlab, the maximum value of dependent and independent variables (X and Y) are selected and all values of these variables are divided by maximum of the variables to scale the data between 0 and 1 (Figure 15 and Figure 16).

To scale the data, Matlab native function “range( )” has been used. It provides a convenient way to scale the data by passing in the variables you want to scale. The output of the function is the scaled output values in the range [0 1].

One debate would be, whether we have to scale both the dependent variable (DV) and the independent variable (IV), or only normalizing one of them will be sufficient. As there are different ranges of values for both EBI and CAP in different datasets, it is needed to scale both of the values according to their maximums. The range( ) function will apply this transformation, so all EBI values are divided by the maximum value of EBI, and all CAP values are divided by the maximum value of CAP. Using the range( ) function also shifts the data which might not start at zero (because of the outlier correction/elimination, mentioned in 3.3.1) to zero, and so the final result of this step will be the scaled data which start at zero.

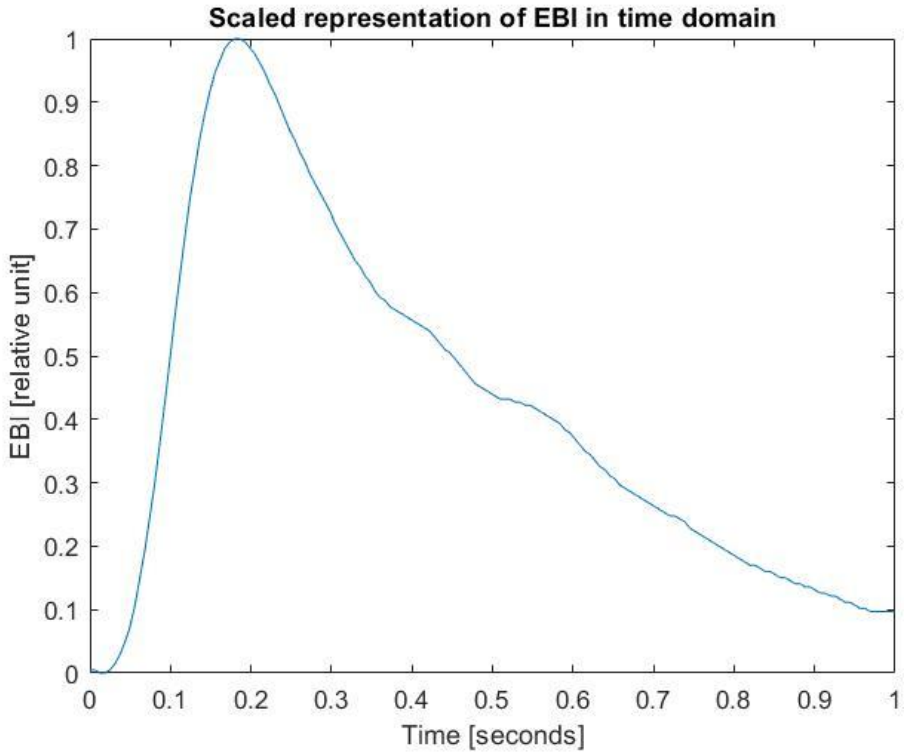

Figure 15 Scaled representation of EBI in time domain

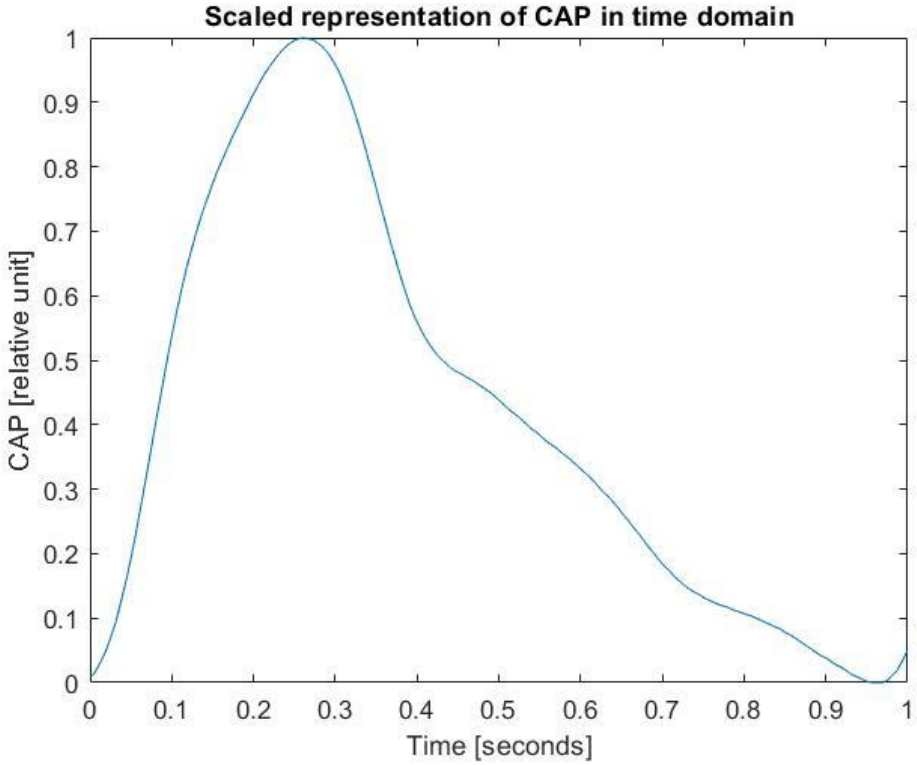

Figure 16 Scaled representation of CAP in time domain

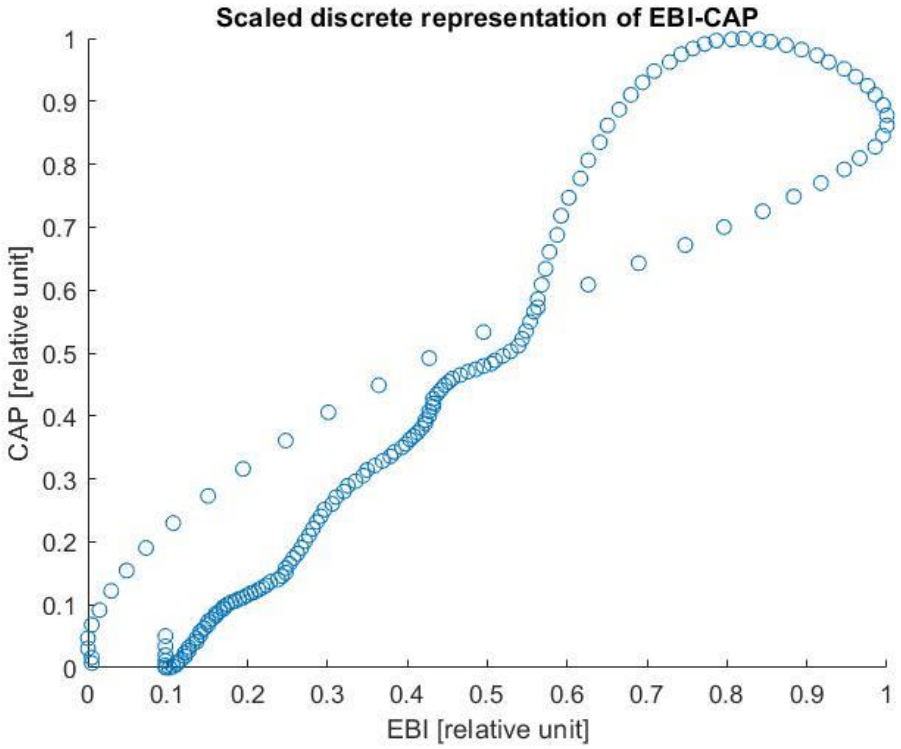

Figure 17 Scaled discrete representation of EBI-CAP in time domain

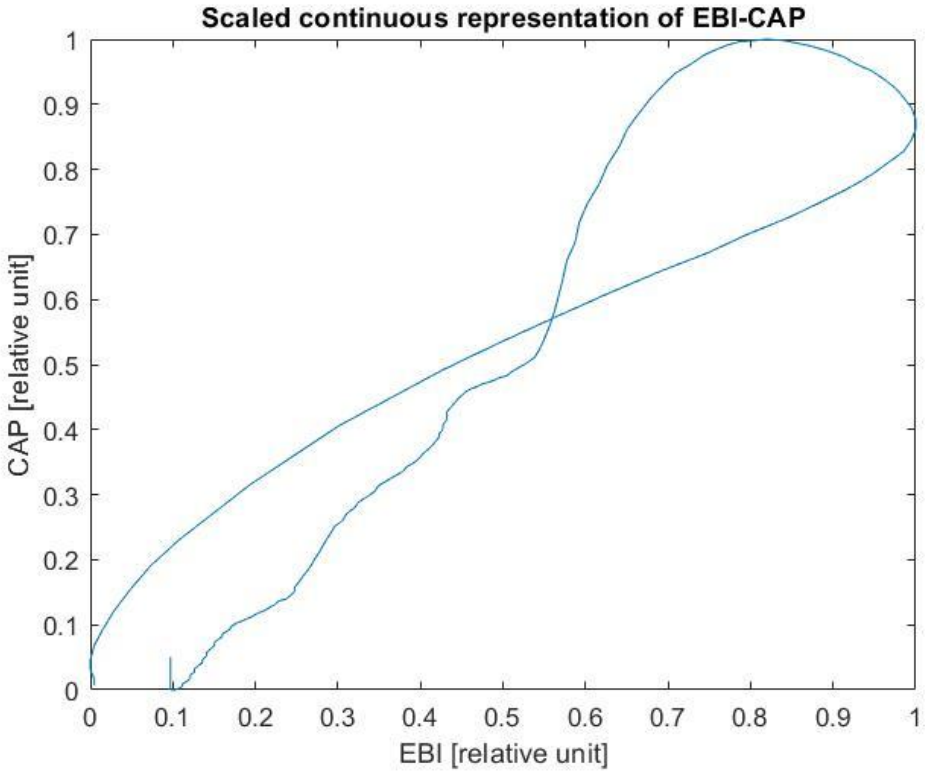

Figure 18 Scaled continuous representation of EBI-CAP in time domain

3.3.3 Dataset Splitting

When outliers are corrected and data is scaled, we have to take into consideration what we want to get out of curve fitting operation as the result, and how Matlab curve fitting toolbox is going to provide that result. It is known that the desired final result of the fitting is an EBI-CAP curve, which is not a function, and curve fitting result will be one certain curve - a function curve in a certain interval. Therefore, we need to define intervals in which the result of curve-fitting would be meaningful (the results are functions). If EBI-CAP curve is splitted into two curves, and the splitting point is considered to be the maximum value of all EBI's (which is 1, because of the scaling), then we end up with two separate curves which also fit in the definition of a function. These two curves also have good overlap with systolic and diastolic curves, respectively. One of the curves carries systolic data and the other curve carries the diastolic data.

With the above-mentioned statements and results, it can be deduced that for most of the methods used, the procedure is to split the data into two parts, and fit curves to those two parts, and finally when there is need to have all the data as one curve, merging of the two curves can happen, and the final result will introduce only one curve in the graph.

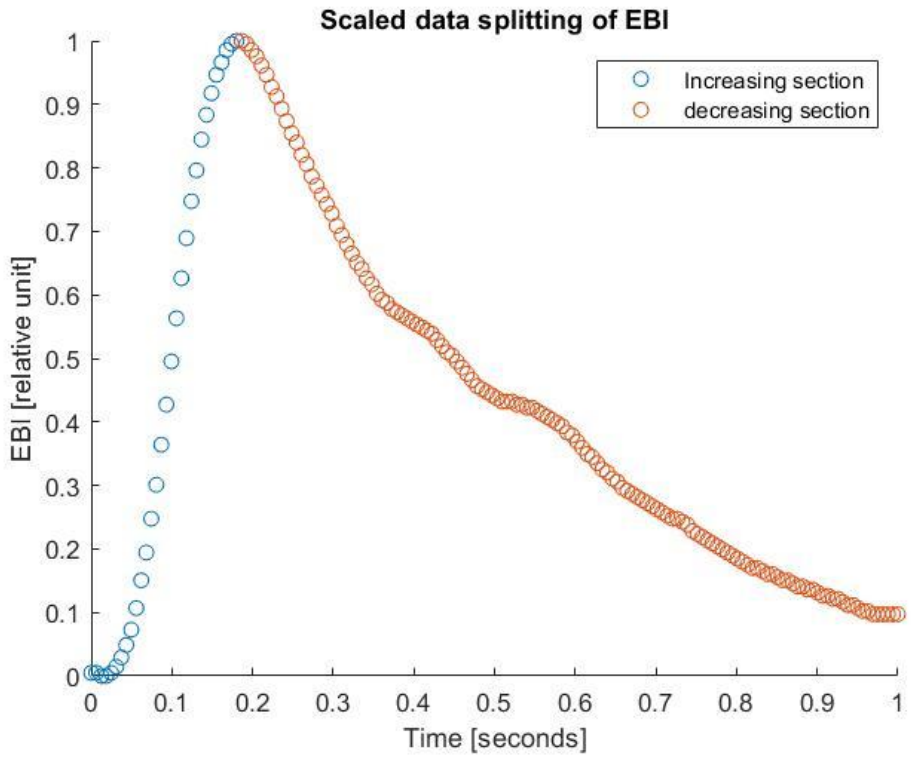

Figure 19 Splitting of scaled EBI dataset in time domain (maximum value of the EBI in time domain is the reference for splitting to increasing and decreasing sections)

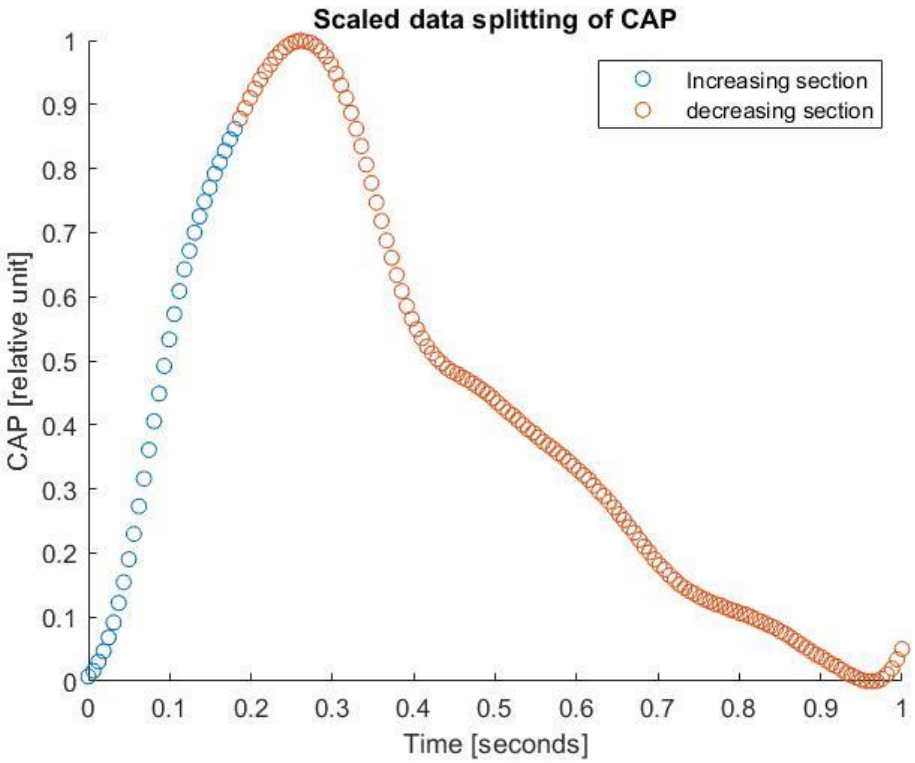

Figure 20 Splitting of scaled CAP dataset in time domain (maximum value of the EBI in time domain is the reference for splitting to increasing and decreasing sections)

After all the above-mentioned steps, the dataset is ready to pass to the related section of the program responsible for the main processing (calculation and analysis) and model generation.

3.4 Methods

At the point where everything is ready from the previous sections, we need to apply different methods of processing to the preprocessed data we have.

The following figure shares the same information demonstrated in Figure 10, plus it can be used as a navigation guide for the processing methods explained in this document.

.

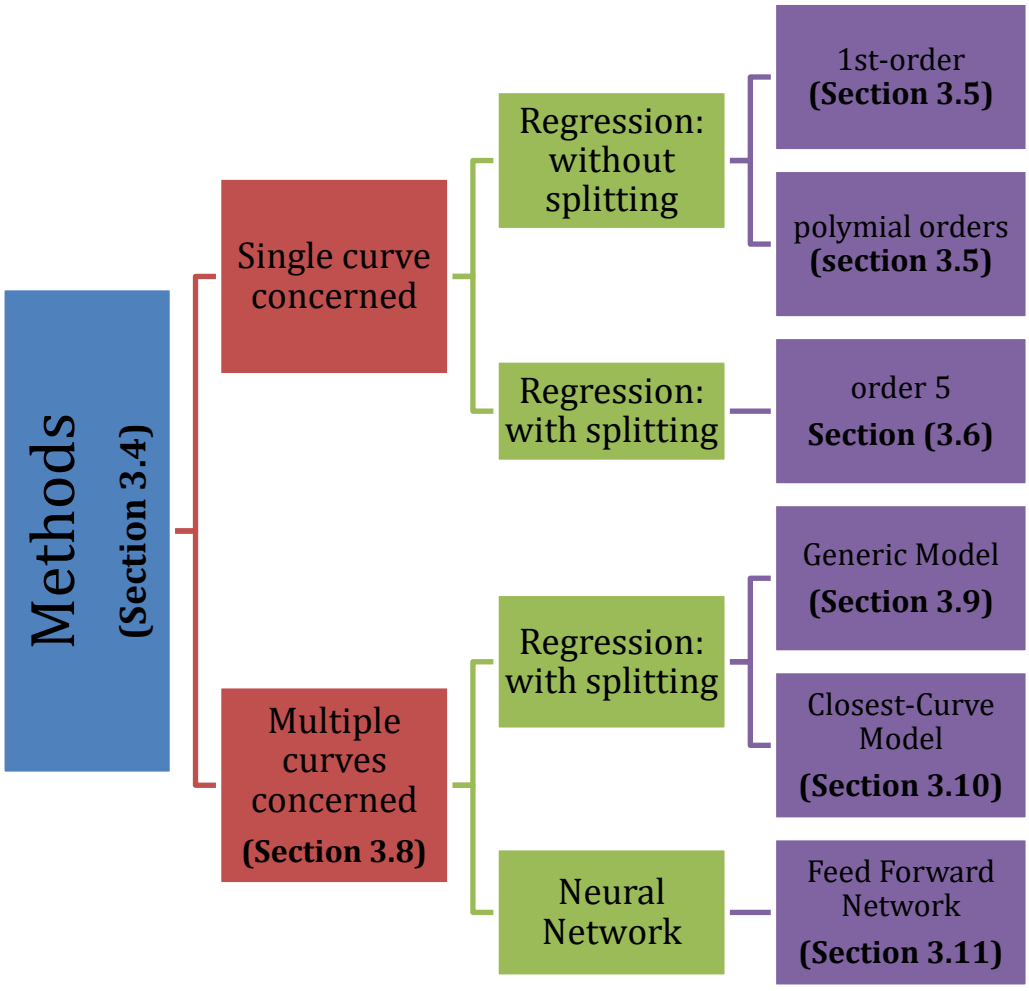

Figure 21 Overview of used method in this thesis. For better navigation, section numbers are included

3.5 Curve fitting

The first chosen approach for finding a solution to the thesis statement problem would be concerning a single curve and applying regression tools to achieve a final curve model capable of predicting future data.

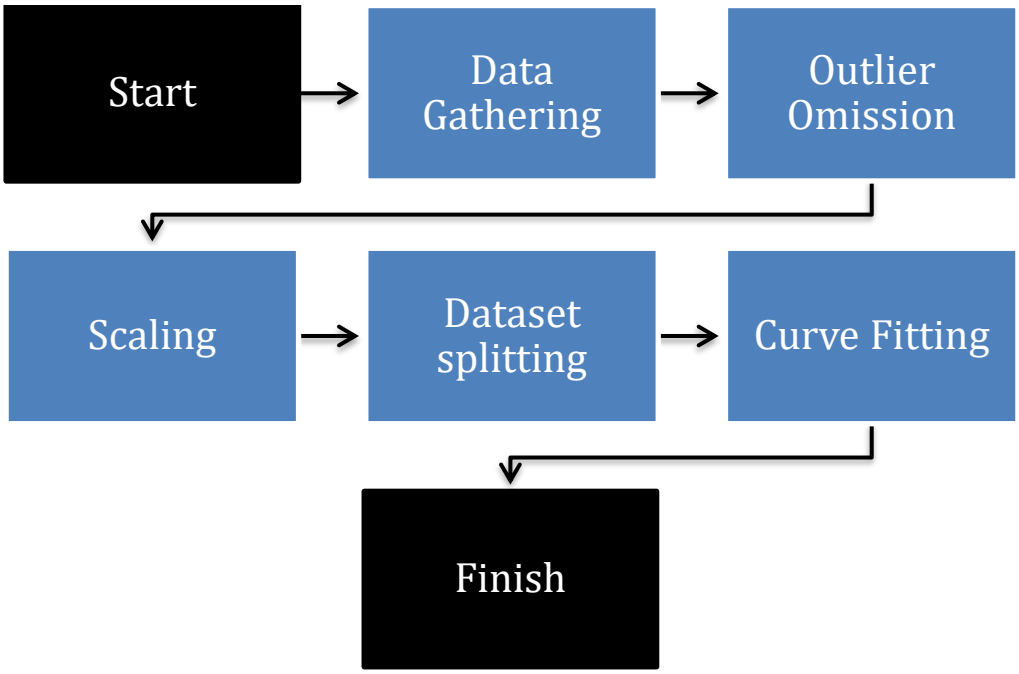

Figure 22 The operation flow of preprocessing and processing stages for curve-fitting methods (first two methods)

Matlab's CFtool toolbox is used as a tool that is specifically designed for curve-fitting operation. Although it is not a very difficult task to write a piece of code for doing the regression and estimating the resulting curve, using a Matlab toolbox might count as the better choice because:

- It is verified and is more accurate than personal code
- It has taken into account all the details and points of failure potential
- It saves a lot of time
- Writing the code of regression by the author does not help achieving the goals of this thesis in any possible way

As discussed before, CFtool works by receiving DV and IV, and type and order of the curve which is aimed. The graphical user interface of CFtool is observable in the next two

figures. Besides the graphical interface, CFtool also comes with a coding interface, which can be used inside the code. The coding interface is faster and provides the user with more options and more flexibility. For our purposed target, we used CFtool GUI in the first two methods for different orders of the regression method and each window of CFtool indicated in this section of the document. Next figure and the one after that (Figure 23 and Figure 24) are explained in the text.

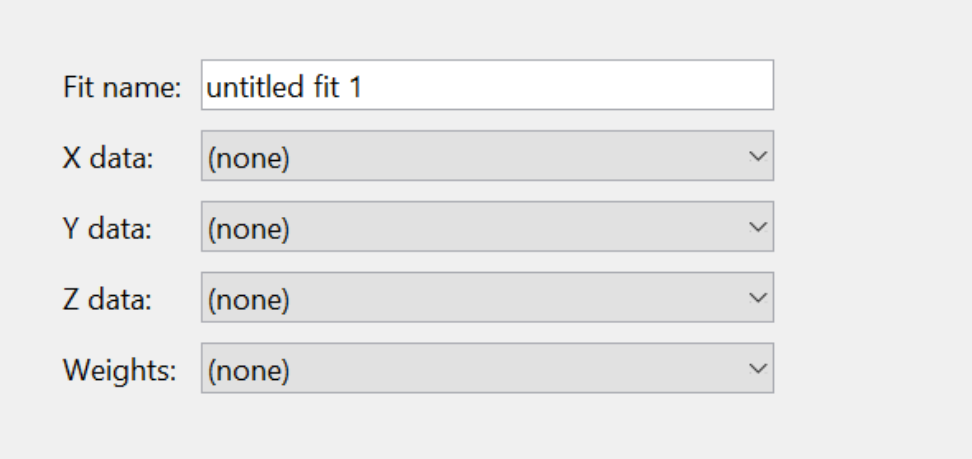

Figure 23 CFtool data selection panel

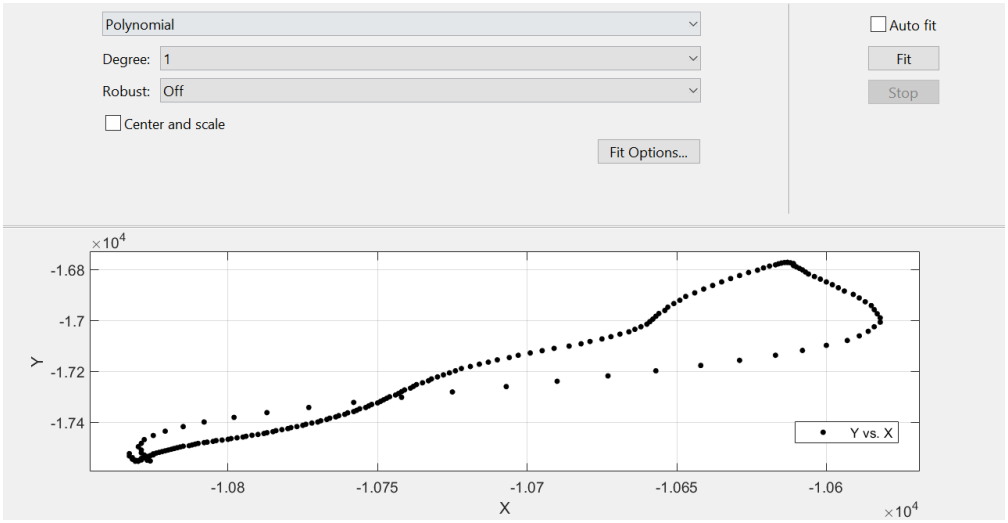

Figure 24 CFtool fitting options

3.5.1 Different orders of curve fitting

After the preprocessing stage, regression algorithms of different orders are used to find a curve with the best fit relative to CAP data. This curve may be a good representative of

the whole dataset if the fitting procedure takes place with sufficient precision and suitable configuration.

For each order of regression models, there exists a separate section below. These sections show that by increasing of regression's order the estimation curve getting better accuracy. Also, along each section, coefficients of different polynomials are indicated in tables such as Table 3.

To show the accuracy and goodness of estimation Table 4, four different characterization methods used, and they are explained as follows:

After fitting data with one or more models, the goodness of fitted models should be evaluated. A visual examination of the fitted curve displayed in figures should be the first step. Beyond that, the Matlab provides these methods to assess goodness of fit for both linear and nonlinear parametric fits [23].
After using graphical methods to evaluate the goodness of fit, the goodness-of-fit statistics should be examined. Matlab supports these goodness-of-fit statistics for parametric models:

- The sum of squares due to error (SSE):
  This statistic measures the total deviation of the response values from the fit to the response values. It is also called the summed square of residuals and is usually labeled as SSE [23].

$$SSE = \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2 \qquad (3)$$

  A value closer to 0 indicates that the model has a smaller random error component, and that the fit will be more useful for prediction [23].

- R-square:
  This statistic measures how successful the fit is in explaining the variation of the data. Put another way, R-square is the square of the correlation between the response values and the predicted response values. It is also called the square of the multiple correlation coefficient and the coefficient of multiple determination.

R-square is defined as the ratio of the sum of squares of the regression (SSR) and the total sum of squares (SST). SSR is defined as

SST is also called the sum of squares about the mean, and is defined as

$$SSE = \sum_{i=1}^{n} w_i(\hat{y}_i - \bar{y})^2 \qquad (4)$$

where *SST* = *SSR* + *SSE*. Given these definitions, R-square is expressed as

$$\text{R-square} = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \qquad (5)$$

R-square can take on any value between 0 and 1, with a value closer to 1 indicating that a greater proportion of variance is accounted for by the model. For example, an R-square value of 0.8234 means that the fit explains 82.34% of the total variation in the data about the average. If you increase the number of fitted coefficients in your model, R-square will increase although the fit may not improve in a practical sense. To avoid this situation, you should use the degrees of freedom adjusted R-square statistic described below. Note that it is possible to get a negative R-square for equations that do not contain a constant term. Because R-square is defined as the proportion of variance explained by the fit, if the fit is actually worse than just fitting a horizontal line then R-square is negative. In this case, R-square cannot be interpreted as the square of a correlation. Such situations indicate that a constant term should be added to the model [23].

- Adjusted R-square

  This statistic uses the R-square statistic defined above, and adjusts it based on the residual degrees of freedom. The residual degrees of freedom is defined as the number of response values *n* minus the number of fitted coefficients *m* estimated from the response values.

$$v = n - m \qquad (6)$$

  v indicates the number of independent pieces of information involving the *n* data points that are required to calculate the sum of squares. Note that if parameters are bounded and one or more of the estimates are at their bounds, then those estimates are regarded as fixed. The degrees of freedom is increased by the number of such parameters. The adjusted R-square statistic is generally the best indicator of the fit quality when you compare two models that are *nested* — that is, a series of models each of which adds additional coefficients to the previous model [23].

$$\text{adjusted R-square} = 1 - \frac{SSE(n-1)}{SSR(v)} \quad (7)$$

The adjusted R-square statistic can take on any value less than or equal to 1, with a value closer to 1 indicating a better fit. Negative values can occur when the model contains terms that do not help to predict the response [23].

- Root mean squared error (RMSE)

  This statistic is also known as the fit standard error and the standard error of the regression. It is an estimate of the standard deviation of the random component in the data, and is defined as

$$RMSE = s = \sqrt{MSE} \quad (8)$$

  where *MSE* is the mean square error or the residual mean square.

$$\text{MSE} = \frac{SSE}{v} \quad (9)$$

  Just as with *SSE*, an *MSE* value closer to 0 indicates a fit that is more useful for prediction [23].

3.5.2 First-order curve fitting model

Figure 25 indicates the first-order polynomial fitted to scaled data of EBI-CAP and the original CAP and its reconstruction in time domain illustrated in Figure 26. Also, coefficients are indicated in Table 3 and goodness of fit is shown in Table 4.

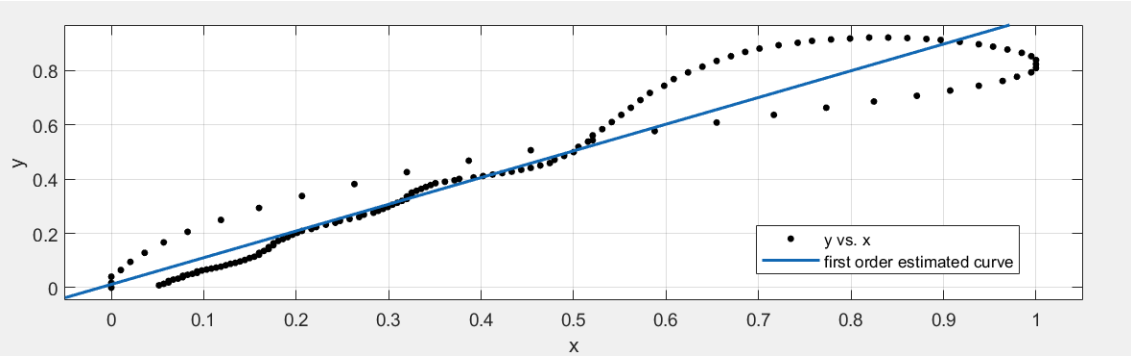

Figure 25 First-order curve fitted to EBI-CAP dataset

First-order polynomial equation:

$$f(x) = p1 * x + p2 \quad (10)$$

Table 3 Coefficients of first-order estimated curve

p1	p2
0.9844	0.01217

Table 4 Goodness of first-order fitted curve

SSE	R-square	Adjusted R-square	RMSE
0.9452	0.9223	0.9217	0.08306

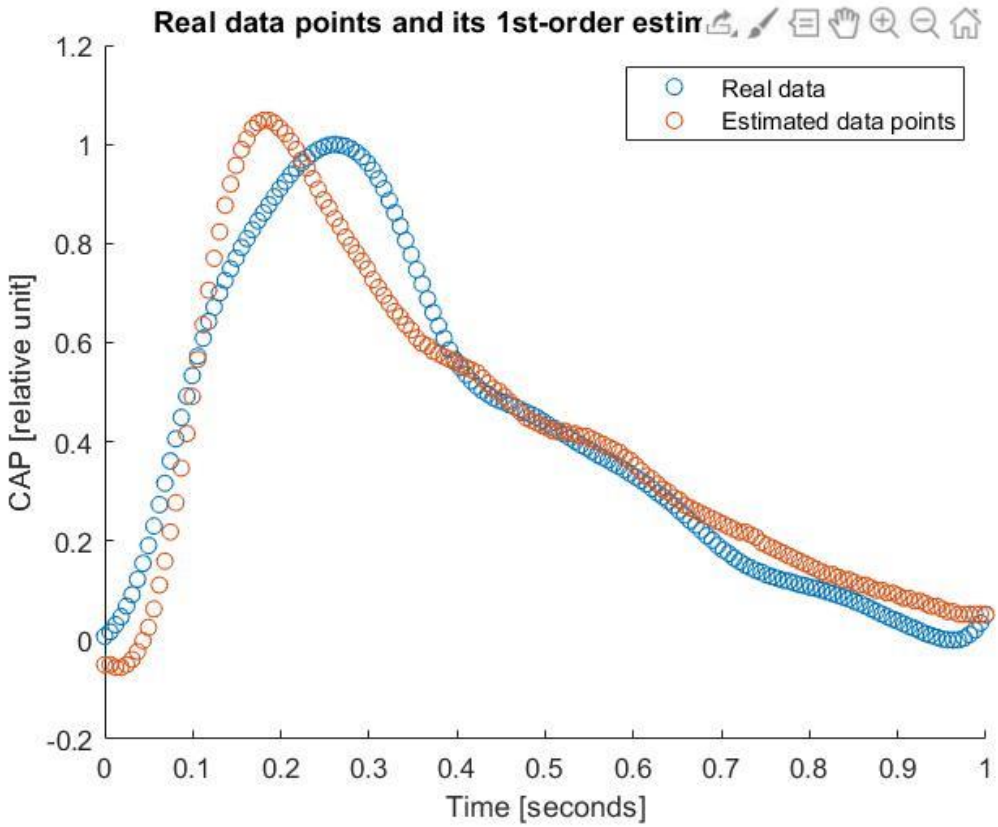

Figure 26 Reconstruction of CAP data in the time domain from first-order estimated curve

3.5.3 Second-order curve fitting model

Figure 27 indicates second-order polynomial fitted to scaled data of EBI-CAP and the original CAP and its reconstruction in time domain are illustrated in Figure 28. Derived coefficients are shown in Table 5 and goodness of fit is contained in Table 6.

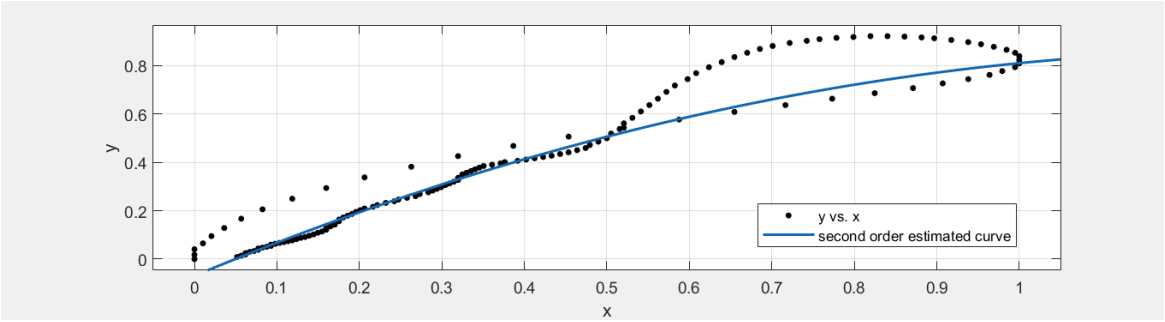

Figure 27 Second-order curve fitted to EBI-CAP dataset

Second-order polynomial equation:

$$f(x) = p1 * x^2 + p2 * x + p3 \qquad (11)$$

Table 5 Coefficients of second-order estimated curve

p1	p2	p3
-0.5417	1.421	-0.0688

Table 6 Goodness of second-order fitted curve

SSE	R-square	Adjusted R-square	RMSE
0.1638	0.9865	0.9863	0.0347

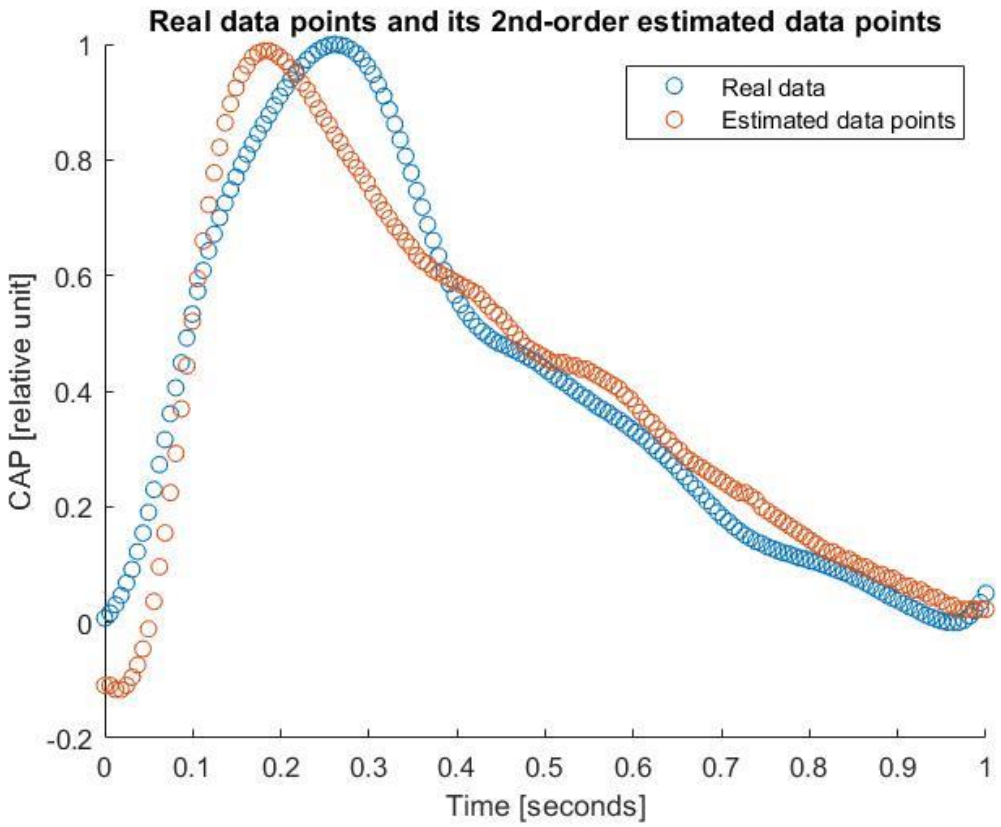

Figure 28 Reconstruction of CAP data from second-order estimated curve

3.5.4 Third-order curve fitting model

Figure 29 indicates third-order polynomial fitted to scaled data of EBI-CAP and the original CAP and its reconstruction in time domain are illustrated in Figure 30. Also, coefficients are indicated in Table 7 and goodness of fit is shown in Table 8.

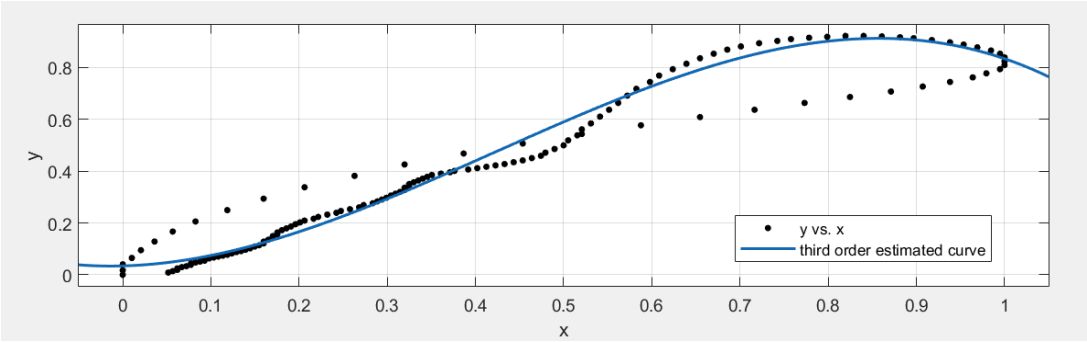

Figure 29 Third-order curve fitted to EBI-CAP dataset

Third-order polynomial equation:

$$f(x) = p1 * x^3 + p2 * x^2 + p3 * x + p4 \qquad (12)$$

Table 7 Coefficients of third-order estimated curve

p1	p2	p3	p4
-0.5417	1.421	-0.0688	0.0344

Table 8 Goodness of third-order fitted curve

SSE	R-square	Adjusted R-square	RMSE
0.2432	0.98	0.9796	0.0424

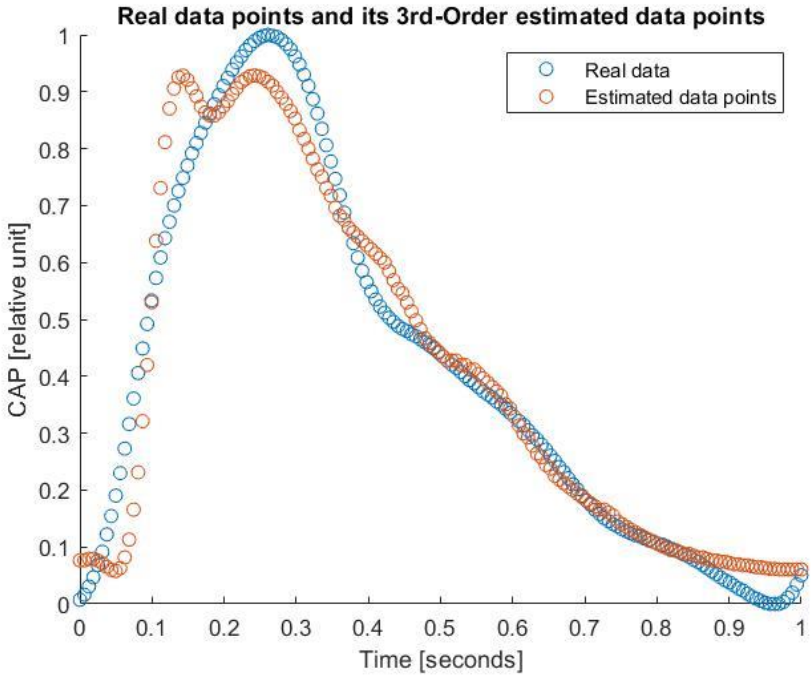

Figure 30 Reconstruction of CAP data from third-order estimated curve

3.5.5 Fourth-order curve fitting model

Figure 31 indicates fourth-order polynomial fitted to scaled data of EBI-CAP and the original CAP and its reconstruction in time domain illustrated in Figure 32. Also, coefficients are indicated in Table 9 and goodness of fit is shown in Table 10.

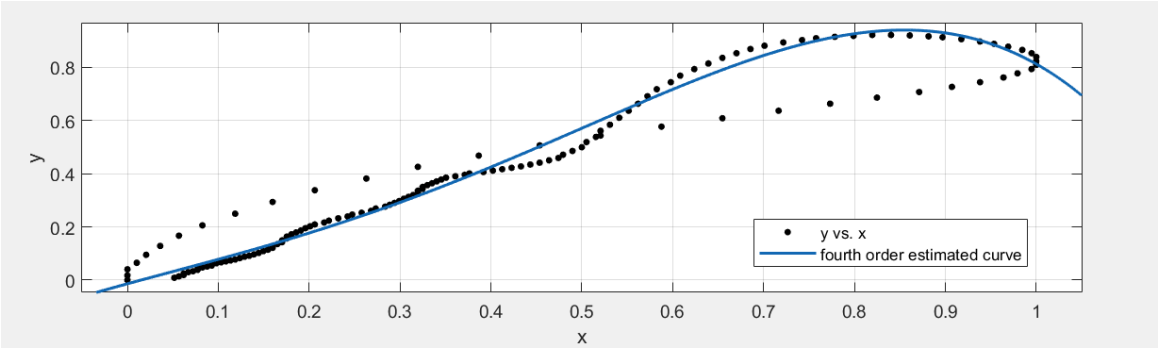

Figure 31 Fourth-order curve fitted to EBI-CAP dataset

Fourth-order polynomial equation:

$$f(x) = p1 * x^4 + p2 * x^3 + p3 * x^2 + p4 * x + p5 \quad (13)$$

Table 9 Coefficients of fourth-order estimated curve

p1	p2	p3	p4	p5
-2.96	3.279	-0.424	0.9323	-0.0137

Table 10 Goodness of fourth-order fitted curve

SSE	R-square	Adjusted R-square	RMSE
0.2167	0.9822	0.9816	0.0402

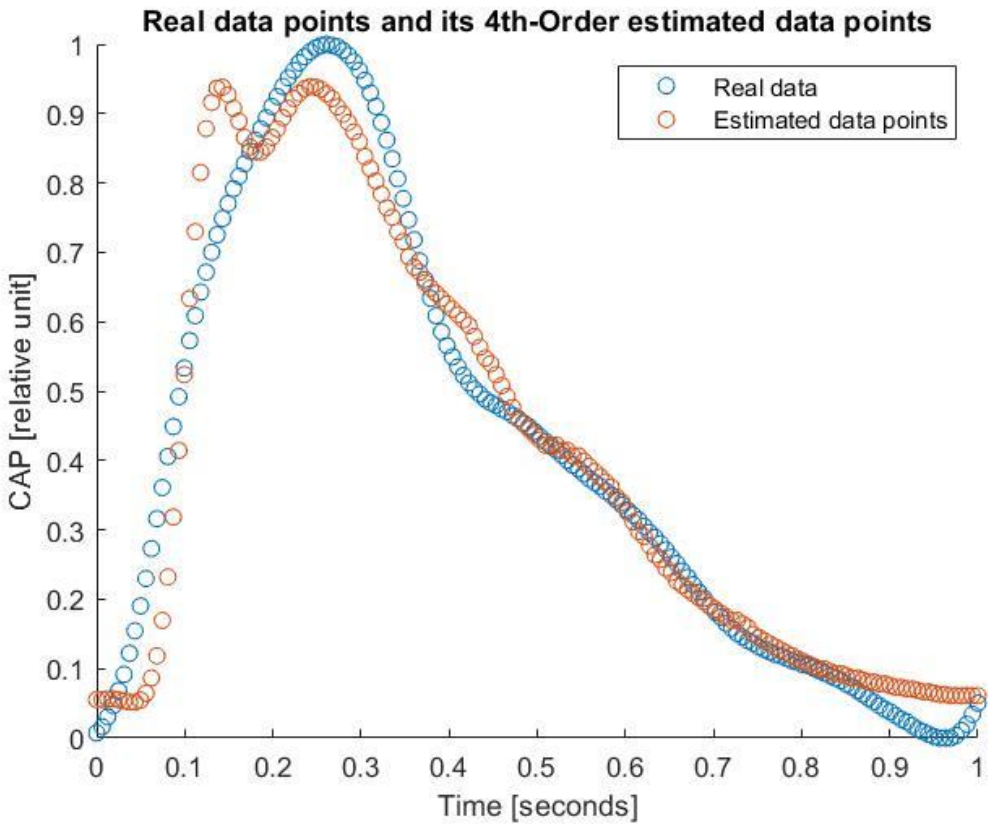

Figure 32 Reconstruction of CAP data from fourth-order estimated curve

3.5.6 Ninth-order curve fitting model

Figure 33 indicates the ninth-order polynomial fitted to scaled data of EBI-CAP. The original CAP and its reconstruction in the time domain are illustrated in Figure 34, and following that, coefficients are present in Table 11 and goodness of fit have been provided in Table 12.

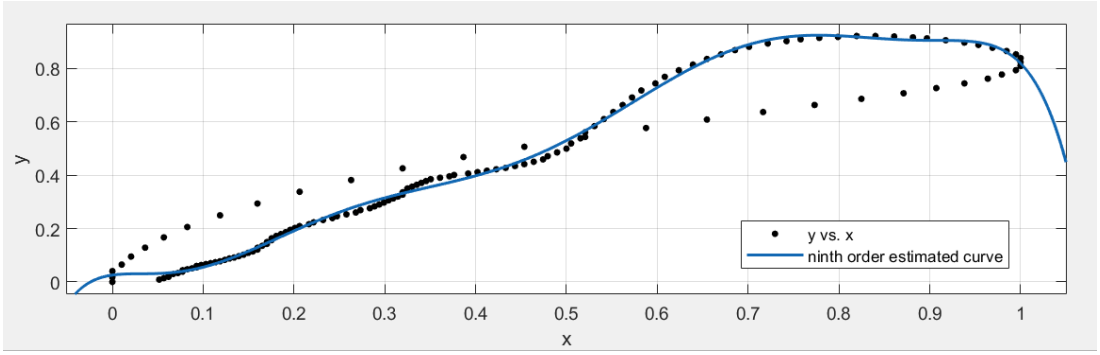

Figure 33 Ninth-order curve fitted to EBI-CAP dataset

Ninth-order polynomial equation:

$$f(x) = p1 * x^9 + p2 * x^8 + p3 * x^7 + p4 * x^6 + p5 * x^5 + p6 * x^4 + p7 * x^3 + p8 * x^2 + p9 * x + p10 \quad (14)$$

Table 11 Coefficients of ninth-order estimated curve

p1	**p2**	**p3**	**p4**	**p5**
0.0073	-0.146	0.2144	0.7323	-1.058
p6	**p7**	**p8**	**p9**	**p10**
-1.318	1.237	0.9827	0.9926	-0.031

Table 12 Goodness of ninth-order fitted curve

SSE	**R-square**	**Adjusted R-square**	**RMSE**
2.36	0.9829	0.9717	0.1353

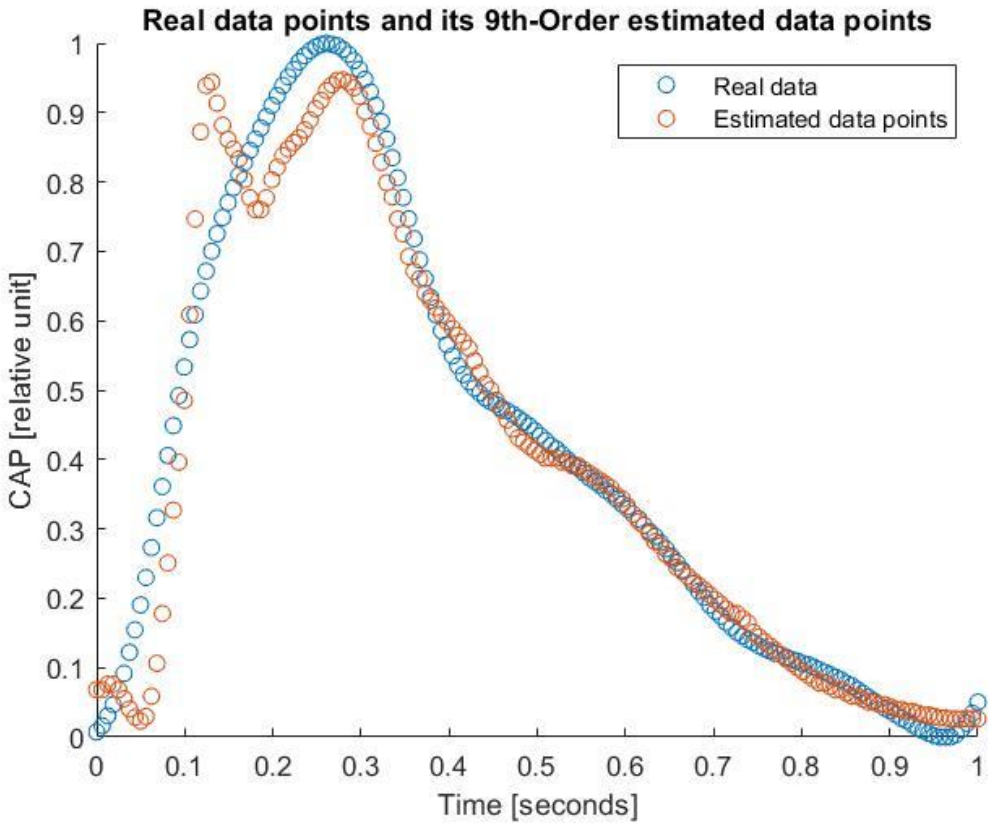

Figure 34 Reconstruction of CAP data from ninth-order estimated curve

The reason behind the choice of the ninth-order polynomial is that the curve of degree 9 fits data with the desired value of accuracy and small value of error when considering the EBI-CAP graph (Figure 34). When graphing the estimated order-nine CAP and the original CAP over time domain, some distortions in the diagram is obvious, which is not desired, and can cause problematic consequences. The reason for such a distortion might be the overfitting that happens because of choosing the n-9th-order curve. To solve the problem of distortion for the 9th--order curve, 5th-order curves polynomials are selected as a final choice in the next section.

According to the explanations given in the above paragraphs, 4 metrics are evaluated for every fitting of the curves. These metrics are different for each dataset and for each of the fittings, and depend on how much noise exists in the dataset. It can be said that the accuracy of a fitting is reflected in the value of these metrics, and if a good fit has been achieved, it can be understood by assessing these values. Detailed description of these metrics are provided at the start of section 3.5.

In this part, the aim was to find a suitable fit curve for a dataset, and regression has been chosen as the technique for such operation. One of the issues which is observed in this section is the low accuracy of the fit in the neighborhood of the maximum point of the curve (Figure 34). In the next section, the effort is on getting this issue solved.

It is necessary to mention as a concluding result that, up to this point, fitting curves with degrees 1 to 9 has been achieved using CFtool, and this can be used as a powerful engine for the calculations and processing in the following sections, and for other methods.

3.6 Optimization of the regression method

As it is illustrated above, estimation is done with curves ranging from first-order to ninth-order polynomial but still, it is not the best fit, especially close to the maximum value of the CAP. The algorithm has a relatively large error of 20% to 35% which seems to be because of the change of direction of the EBI-CAP curve around the maximum value.

To solve this issue, the dataset was splitted into two separate parts. The separation is from minimum to the maximum values (maximum value of EBI), and then from the maximum value, to the end of the dataset signal period. It is taken into account that each heartbeat consists of increasing section (Figure 35) and decreasing section (Figure 37) intervals. The idea behind the optimization stated in 3.6.2.

3.6.1 Fifth-order curve fitted to increasing section of CAP

Figure 35 indicates fifth-order polynomial fitted to scaled data of the increasing section of EBI-CAP. The original increasing section of CAP and its reconstruction in time domain are illustrated in Figure 36. Also, coefficients are expressed in Table 13 and goodness of fit is shown in

Table 14.

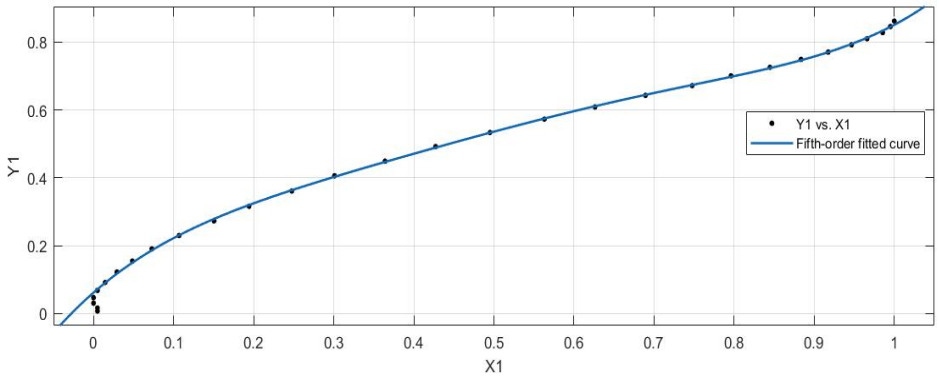

Figure 35 Fifth-order curve fitted to increasing section of EBI-CAP data

Table 13 Coefficients of fifth-order increasing section of estimated curve

p1	p2	p3	p4	p5	p6
0.08563	-0.1291	-0.1413	0.2506	0.09169	-0.2015

Table 14 Goodness of fifth-order increasing section of the fitted curve

SSE	R-square	Adjusted R-square	RMSE
0.008398	0.9996	0.9995	0.0216

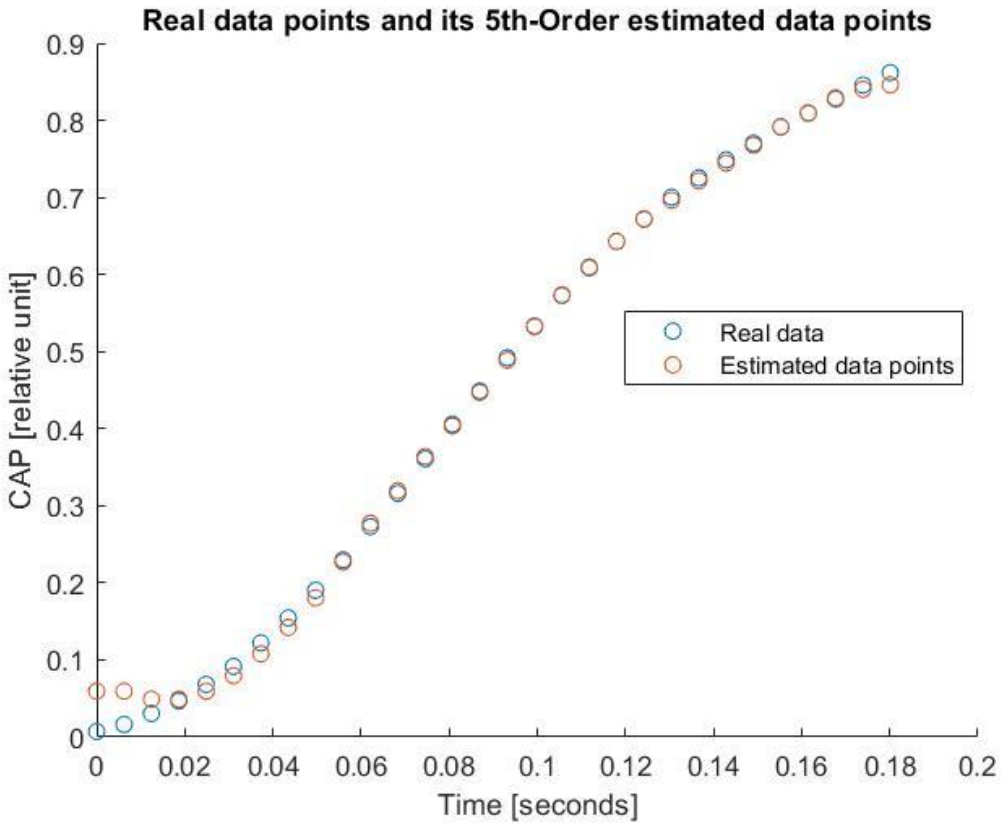

Figure 36 Reconstruction of CAP data from fifth-order estimated curve of increasing section

3.6.2 Fifth-order curve fitted to decreasing section of CAP

With the purpose of gaining the best possible fitted curve, on one hand, we should be careful not to follow the noises of the data (which happens when the order of the curve increases), and on the other hand, we should choose an order which can offer a good fit (which means a more complicated curve with higher order). So, the purpose of optimization is to find a balance between the two mentioned situations. Fifth-order curves seem to be the best possible fit for such a situation, and so the 5th-order is chosen in this section.

Figure 37 indicated fifth-order polynomial fitted to scaled data of decreasing section of EBI-CAP. The original increasing section of CAP and its reconstruction in time domain illustrated in Figure 38 also, coefficients indicated in Table 15 and goodness of fit is shown in Table 16.

Figure 39 shows the merging of increasing and decreasing sections of the curve to form one single continuous curve. It indicated that the problem related to error close to the maximum value of CAP Figure 34 is solved and indicated a better-fitted curve.

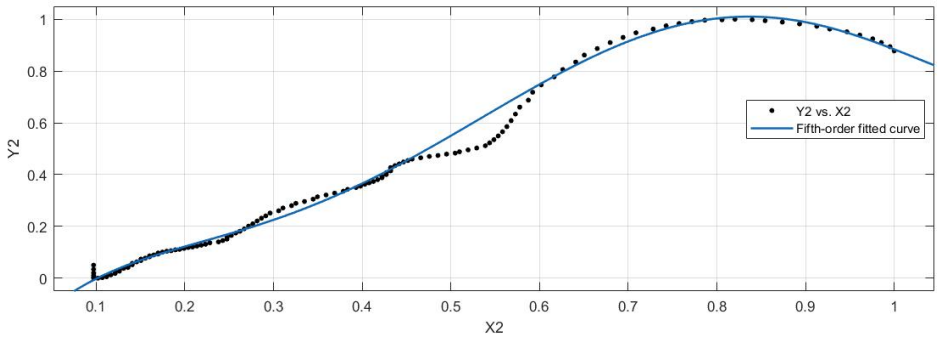

Figure 37 Fifth-order curve fitted to decreasing section of EBI-CAP data

Table 15 Coefficients of fifth-order decreasing section of estimated curve

p1	p2	p3	p4	p5	p6
0.04056	-0.2369	0.2186	0.843	-1.186	-1.174

Table 16 Goodness of fifth-order decreasing section of the fitted curve

SSE	R-square	Adjusted R-square	RMSE
0.2267	0.998	0.9978	0.04737

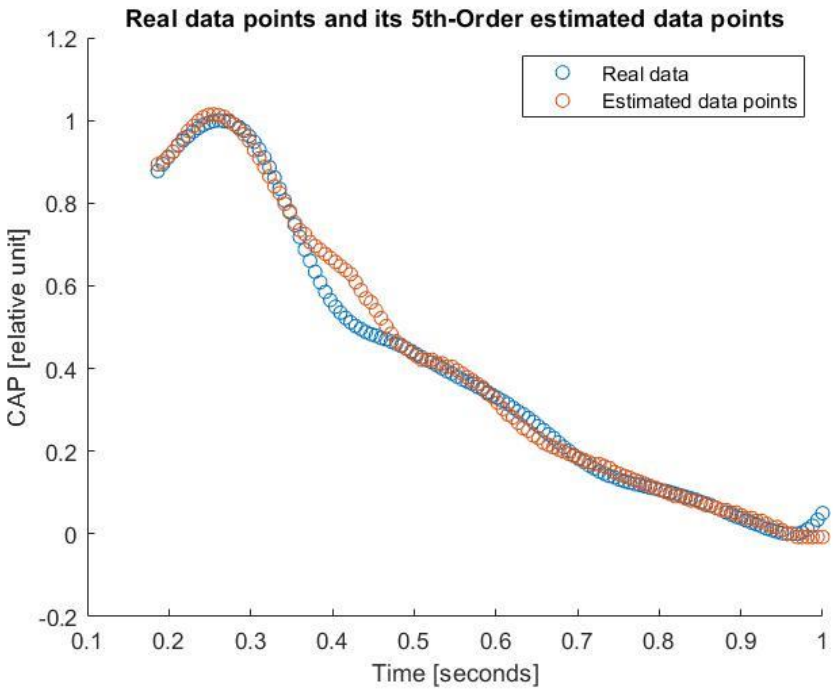

Figure 38 Reconstruction of CAP data from fifth-order decreasing section of estimated curve

3.6.3 Merging the two curves

The two resulted pieces of curves, obtained after splitting and regression, are merged in this step, to acquire a complete estimated CAP relation, over time (Y-T graph).

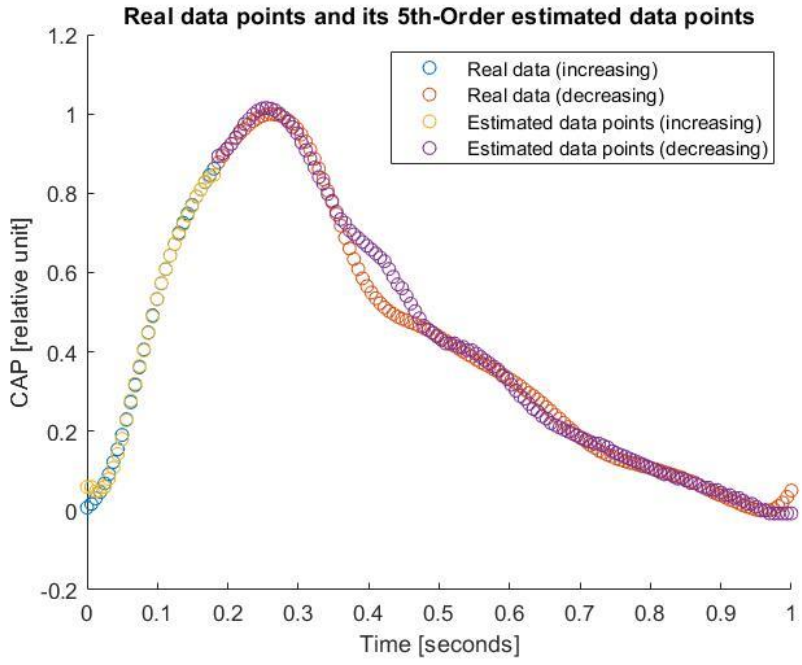

Figure 39 Merge of increasing and decreasing sections CAP on one figure over the time domain

3.7 Testing different datasets

In this part of the thesis, to illustrate that the method is globally working, it is tested on a different dataset.

3.7.1 Implementation of the method for the second person

In the figure below, EBI and CAP data illustrated for both not scaled and scaled waveform.

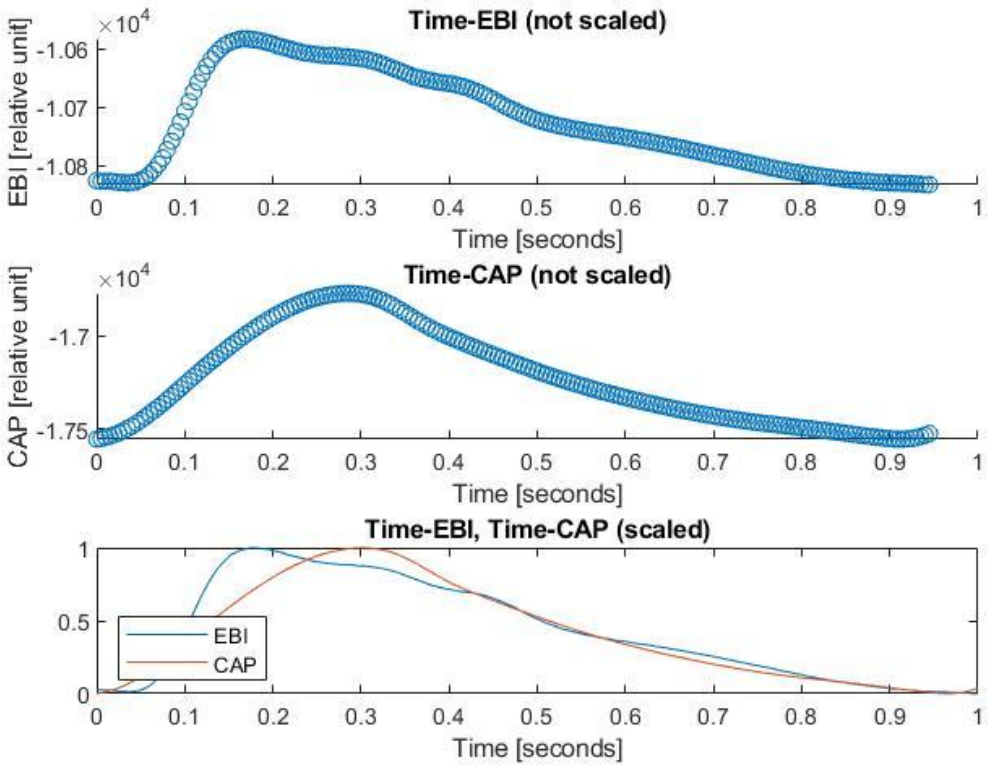

Figure 40 EBI & CAP illustration for the second person

In the figure below, the algorithm is tested for both the increasing and the decreasing CAP parts and the CAP data reconstruction is indicated on the same figure.

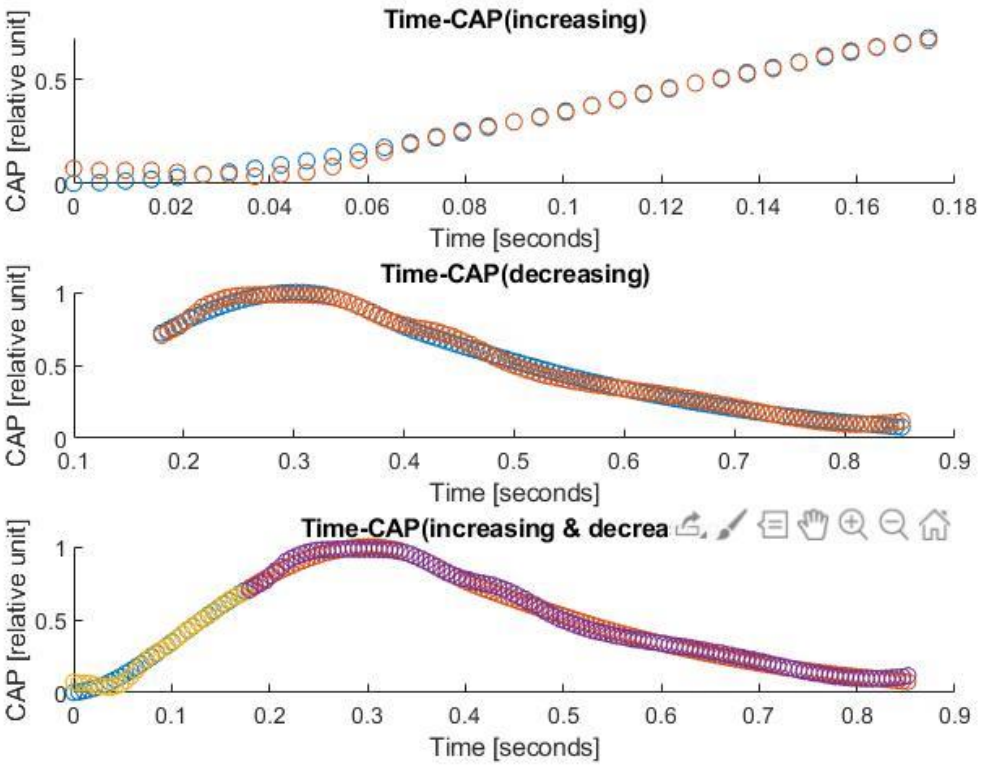

Figure 41 CAP prediction of increasing & decreasing for the second person

Table 17 Coefficients of fifth-order systolic estimated curve

p1	p2	p3	p4	p5	p6
0.01592	-0.0323	-0.0335	0.1021	0.02387	-0.1231

Table 18 Coefficients of fifth-order decreasing section of estimated curve

p1	p2	p3	p4	p5	p6
1.122	-5.525	8.01	-0.0039	-6.917	1.304

3.7.2 Implementation of the method for the third person

In the figure below, EBI and CAP data illustrated for both not scaled and scaled waveform.

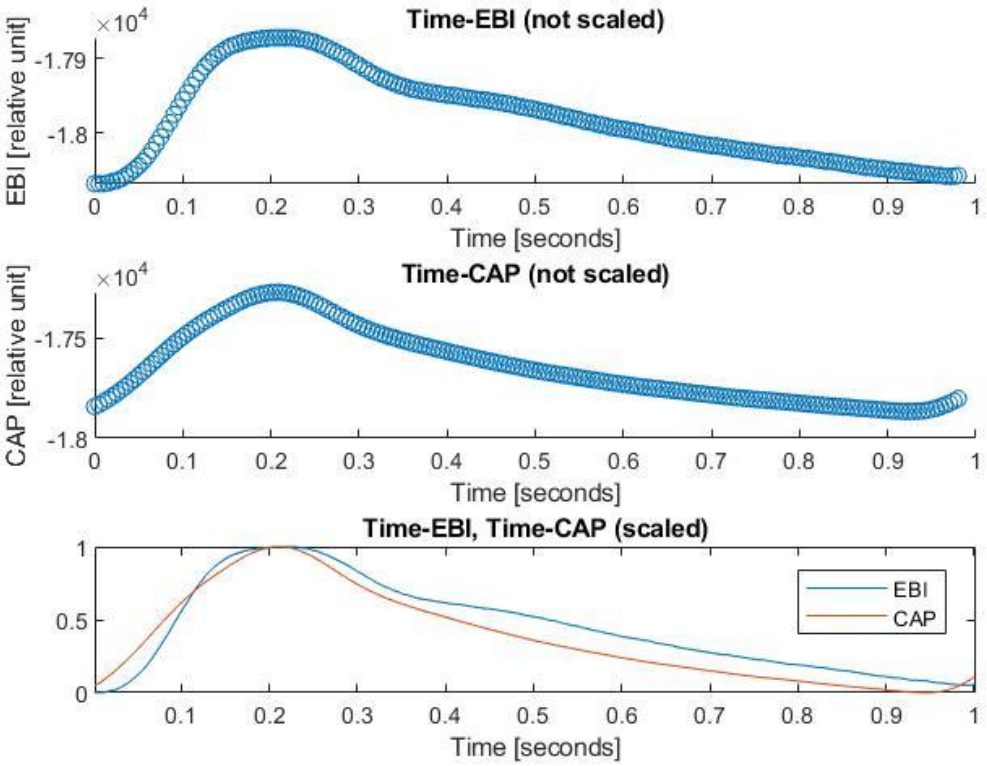

Figure 42 EBI & CAP illustration for the third person

In the figure below, the algorithm is tested for both increasing and decreasing CAP part and the CAP data reconstruction is indicated on the same figure.

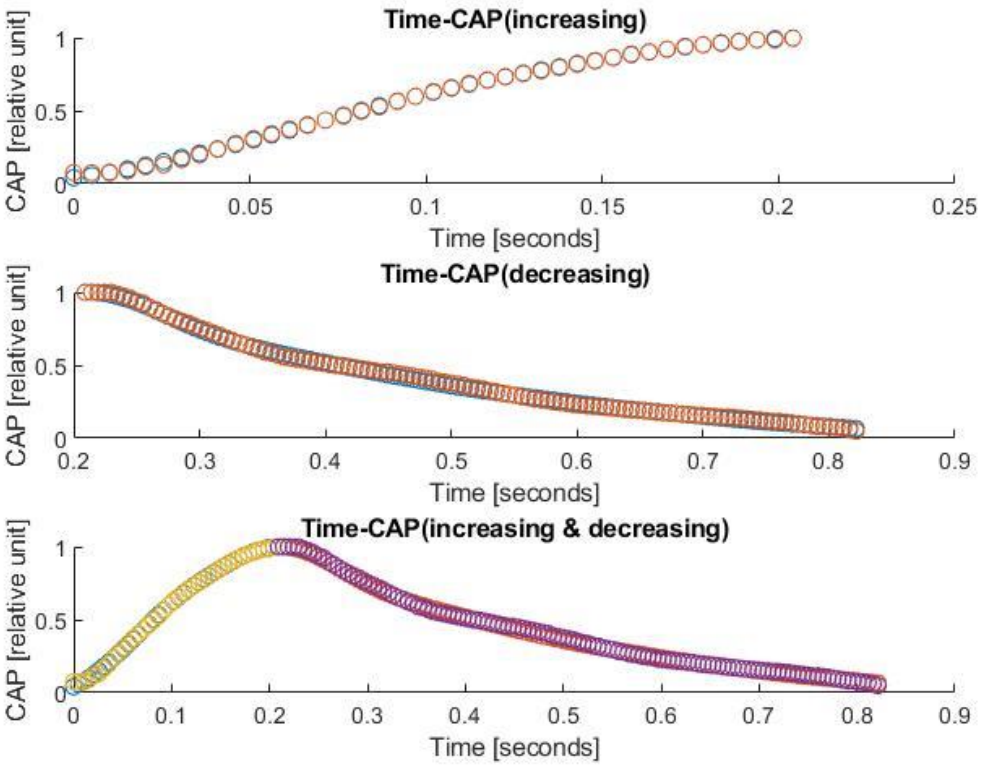

Figure 39 CAP prediction of increasing & decreasing sections for the third person

Table 19 Coefficients of fifth-order increasing sections of estimated curve

p1	p2	p3	p4	p5	p6
0.1472	-0.3519	-0.2076	0.8733	-0.02	-0.7042

Table 20 Coefficients of fifth-order decreasing section of estimated curve

p1	p2	p3	p4	p5	p6
-0.0318	0.2711	0.5974	-3.162	0.938	3.742

By splitting the dataset into increasing and decreasing sections, and estimation of 5th-order for each of these sections separately, it was possible to reduce the error in the neighboring points of CAP maximum point to about 5% for different individuals. Conclusively, the method including the splitting can be considered a more precise method than the method without split. This has been applied to different datasets and three instances of this application are presented in the former sections.

3.8 Development of updated improved methods (concerning multiple curves)

Facing the challenges and limitations of previous models, three new solutions were considered and some approaches were thought of, which are going to be explained in the current section. These improvements are wrapped and organized systematically and the results are three new models named “Generic Model” (GM), “Closest-Curve Model” (CCM) and “Neural Network Model” (NNM). In these three new methods (which will result in three new models) the drawbacks of previous methods are tried to be eliminated or reduced as much as possible, and new techniques have been introduced while developing. For the most part, the preprocessing stage is similar for these models, but the processing is different for each of the models. Post-processing procedures (like model reconstruction) is partly similar, and testing would consist of various non-similar steps as testing is affected by processing, and processing is different for each of these methods. Figure below demonstrates, in summary, the taken steps of these three methods.

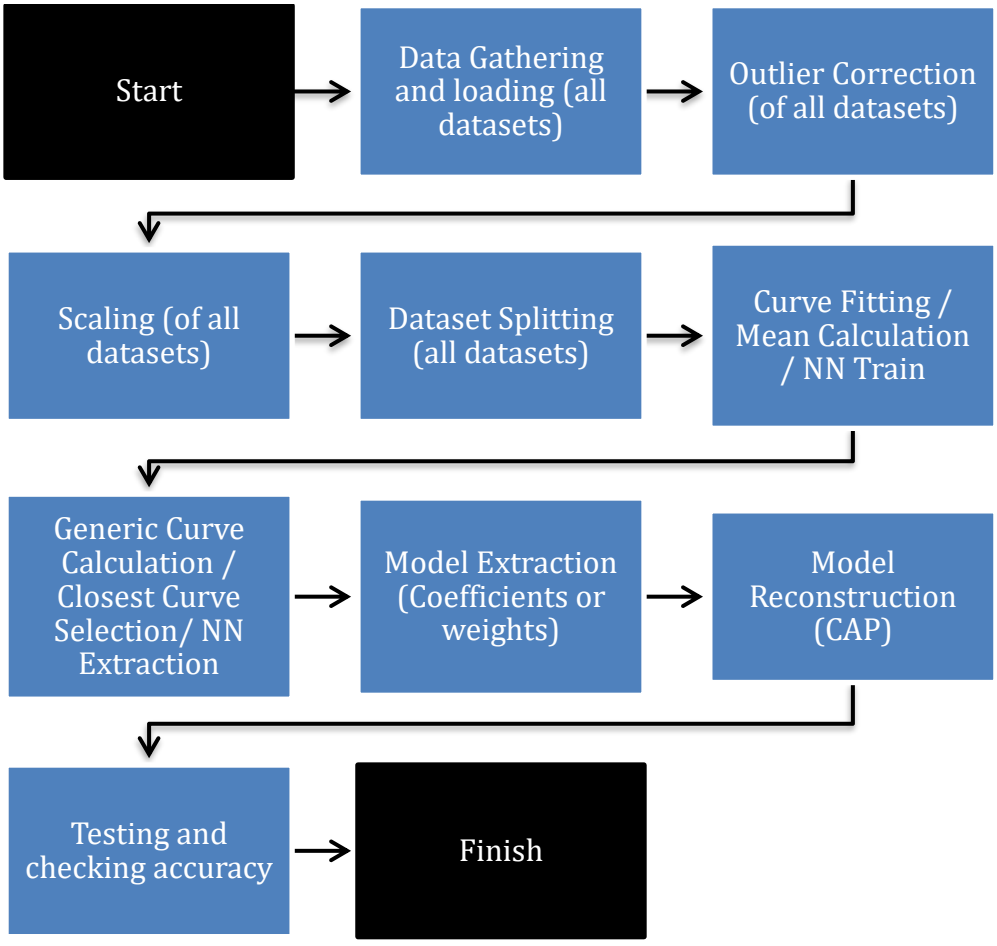

Figure 43 The operation flow of preprocessing and processing stages for last three methods

3.8.1 Drawbacks and limitations of the previous methods

As pointed out in the previous section (3.7), there were some limitations and drawbacks associated with the method used. There are a few items which can be expressed:

1. **Models with generic application and concern**: In the previous method, only one curve is fitted based on the data of one single dataset, and a model which considers "all patients' data" and which can be used with accuracy for a "multitude of inputs" is not provided. In the current method (as explained from this point and further), providing models that can be used in general cases are targeted - models which has a focus on all of the data we have (data for all the patients) as a whole to reach a better result.
2. **Automation**: Another concern was automation which was not given proper importance in the developed code for the former method. The processes did not take place in an automated way and there were hard-coded pieces of code that had to be changed every time a new dataset was introduced. Processes such as "loading the data from a folder", "using CFtool for curve fitting", etc. are carried out automatically in the current method. The following list of processes has been taken into account during the automation:

    a. Reading from and writing to files: The data is not considered on a file-by-file basis. All datasets are loaded from a folder including dataset files in ".log" format. The data is then saved in cell structure for further processing. Then, in the end, the test data is loaded in the same fashion (from a folder). Finally, the results are stored in different corresponding files.
    b. Use of functions: Throughout all the steps of writing the code, usage of functions has been a priority in the second section of the thesis. Functions make the code easier to read and provide ease for the repetition of a piece of code.
    c. Generalization of processes: For instance, it happens that while doing a calculation or applying a formula, the bounds are defined specific to a set of data, and for a new dataset these bounds should change respectively. To get rid of such limitations and to eliminate user interruption in the code, different sections of the code had to be re-written with generalization in mind.
    d. Using scripts to use toolboxes (not toolbox GUI): Usage of toolboxes like CFtool in GUI mode restricts the autonomous features of an algorithm/program because the power of scripting is missed and this causes the developer to have less control over the code. To solve this, I made use of this toolbox through its scripting interface and used arguments to pass values and set up the options.

e. Acquiring a general solution: Another level of automation has taken place in the way we have looked at the solution as a generalized solution which can be used for a variety of patients under study, and not just for people with very similar records (similar to the current data). In order to achieve this, the three updated models have been developed. Explanations of each of the models can be found in the following sections3.9,3.10 and 3.11.

3. **Tools other than regression (neural network):** One main tool which has been employed in this section, and has not been considered in the former section is Neural Networks. Although the use of the network can be improved by further tuning and increasing the amount of data in supervised learning, I have tried to implement a network to be able to compare and analyze the capabilities of neural networks for our dataset with our aimed functionality.
4. **Testing:** Model as a product of engineering/science processes (which in turn include the usage of the theories, calculations, etc.) is not valid until it can pass tests, and be validated with means of measurement, calculation, and analysis. Hence, in the second part of the thesis, for each generated model, there exists a testing phase, through which the resulting model is tested with real datasets. This test is different from the test in the first two methods on different levels. For instance, in the first two methods, we only fitted the curves and each model was created for one person, and so a real test could not be performed. (section 3.7, Testing different datasets)

3.8.2 Faced Challenges

One of the challenges which were encountered while working with numerous datasets was that in some of the data there existed some distortions whose form was out of range in a non-conventional way, which in turn caused the curve to be unnaturally distorted. This suggested the faulty record of data at the time of measurement. It was not possible to correct or eliminate these distortions because of the large interval that they occupied, and on the other hand, the error they caused could not be neglected. Therefore, the choice was to eliminate those datasets whose data are distorted. In Figure 44, the distorted data points of one of the datasets are visualized.

Another group of challenges which were faced included the corrections/eliminations needed for the algorithm to work properly and in the expected style. These have been mostly explained in data preprocessing section (3.3)

The last group of challenges consisted of the decision making about the configuration of an algorithm in regression and in neural networks.

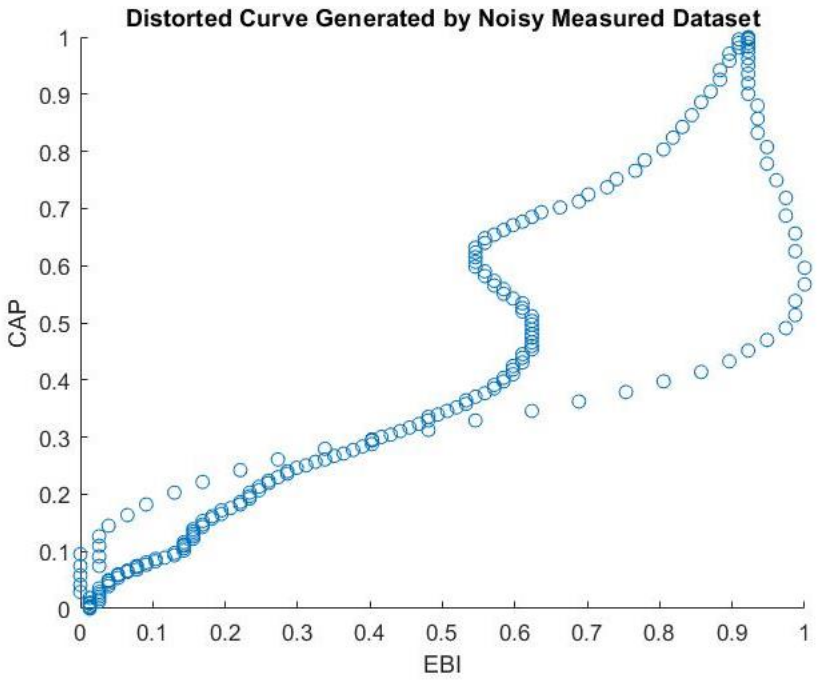

Figure 44 Distorted curve, which is the result of noisy dataset

3.8.3 Acquired Output

Three procedures were included, each of which incorporated a specific technique to achieve models capable of prediction, and hence three models were acquired. All of these models (equation, curve, etc.) has the ability to predict future CAP based on future EBI data, each in its own specific fashion.

The results mostly consist of:

- Extracted model: in the form of coefficients or weights (NN)
- Reconstructed curve: EBI-CAP curve based on the extracted model
- Comparison and testing: comparing the data calculated by acquired models against real true data that has been recorded in the measurement process

3.9 Generic Model

In Generic Model (GM) approach, EBI and CAP data of multiple people is the matter of focus and process. In brief, this method works by calculating a curve (EBI-CAP curve) whose CAP value is the mean value of CAP of all curves in EBI domain. Mean is a central tendency measure and so is a suitable choice for making this model. For this model to take effect and be functional, a few steps need to be taken, which are going to be described in more detail in the sections to come.

3.9.1 Selection, Loading, and Preprocessing

Because of the distorted datasets (3.8.2, Faced Challenges), the first step is to choose a number of datasets and eliminate distorted ones. The loading of different datasets into Matlab is the next step. The files related to different patients are placed in a folder and the code starts to load the files in that folder into Matlab cell data structure. In the current work, files related to 14 patients were chosen to work with. The preprocessing is also conducted on all datasets. A detailed explanation of preprocessing can be found in 3.1.2.

3.9.2 Fitting curves on current data of multiple individuals

When the data is ready, 14 curves are fitted to the data of the 14 people, and the result is saved in Matlab cells. The process of fitting the 14 curves is not different from the previous fitting procedure in previous methods. The result of fitted curves for increasing and decreasing sections are presented in Figure 45.

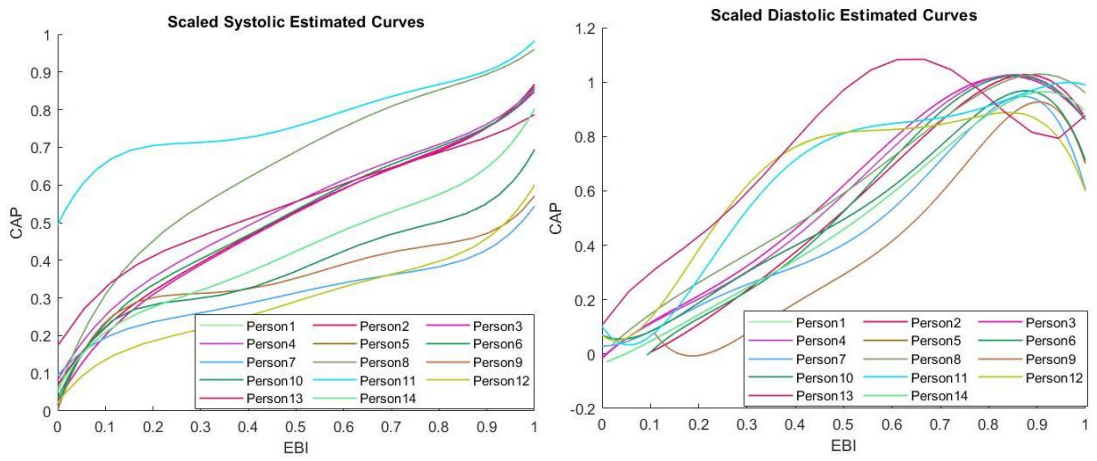

Figure 45 Curves fitted to EBI-CAP data of 14 persons. On the left, the curves related to increasing EBI (through time) is observed, and on the right the decreasing section of the EBI (through time)

The reason that curves are fitted, and the current data points are not directly used for averaging (next step) is that data points for the exact value of a certain EBI (e.g. EBI = 0.1) do not exist in all datasets, and by fitting a curve, any arbitrary EBI can be accessed and processed.

3.9.3 Calculating the mean value of all fitted curves (step size)

When the curves are fitted, it is time to generate points at specific EBI's on each curve. As an example, for person #4, 1000 points are created by choosing step size of 0.001, and fetching points at EBI = 0.000, 0.001, 0.002, 0.003, …, 0.999, 1.000. This process is repeated for all curves and as a consequence, $\frac{EBI\ range}{step\ size}$ points are created for each curve, and $14 \times \frac{1}{step\ size}$ points are created in total. EBI range is 1 because of the previous scaling.

Then at each specific value of EBI, the average value of CAP of all curves (corresponding to that EBI) is calculated and stored. The result will be a set of $\frac{1}{step\ size}$ ordered pairs of the form $(EBI_i, average\ CAP_i)$, where i denotes the number of a specific interval. The outcome of this process will be set as follow:

$$Z_{avg} = \{\ (EBI_1\ , avgCAP_1)\ , (EBI_2, avgCAP_2)\ , \dots\ , (EBI_n, avgCAP_n)\ \} \qquad (15)$$

Where n is used to indicate the number of the specific interval on which averaging has been applied. n is an integer value between $1\ and\ \frac{1}{step\ size}$.

3.9.4 Fitting a curve through average values (Z_{avg})

After acquiring the average points, the average curve is fitted through all these points (points of Z_{avg}). This fitted curve is the eventual model which can be representative of all curves in a single curve, and can be used to predict future values of CAP based on any arbitrary EBI.

It is noteworthy that the average curves of increasing and decreasing sections are separately calculated (because of the reasons mentioned in 3.6), and then they are merged at the time of prediction. Therefore, the model can work for all the values whether they are in decreasing or increasing scope.

Since the model is able to take into account all EBI and CAP values of all the patients at the same time and is able to make predictions based on the described processing procedure, it deserves the name “Generic Model”, and hence it has been named so.

3.9.5 Testing

For the purpose of testing and checking the accuracy, predicted data by the acquired model is compared with the real true data which has been acquired through medical measurements of patients. It should be noted that the test data is different from the data used for computing the model. The comparison made between the two pointed out curves in EBI-CAP diagram is observable in the following diagram.

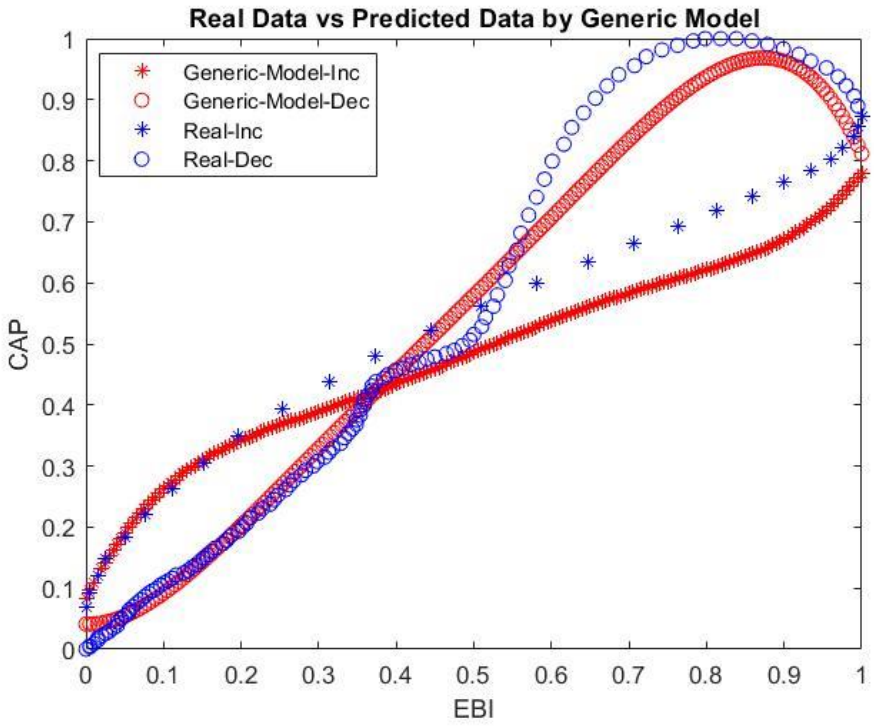

Figure 46 Real data vs Predicted data of EBI-CAP using generic model

Also, after reconstructing the CAP predicted data in the time domain, the diagram of predicted CAP data vs real true CAP data is graphed. The result is shown in the next figure.

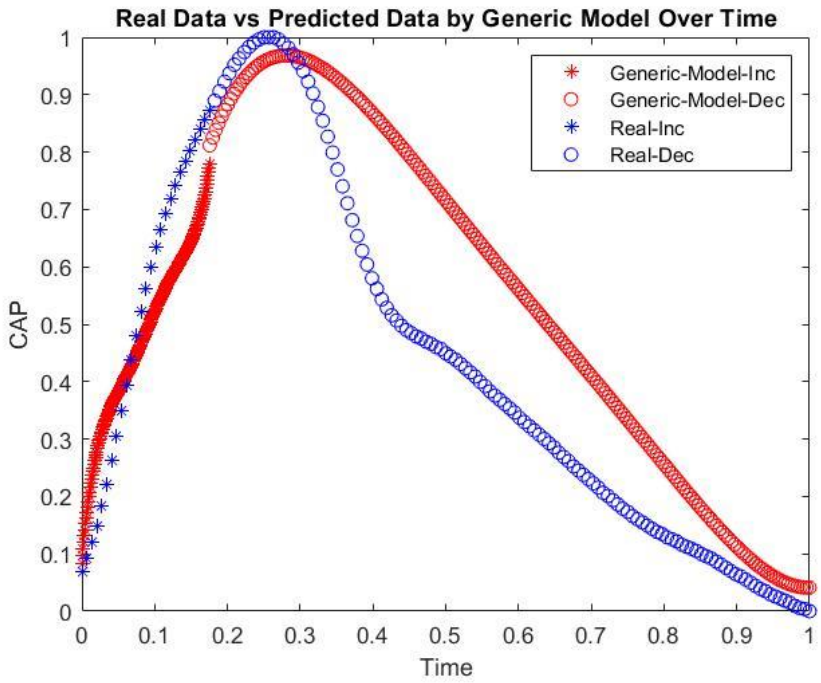

Figure 47 Real data vs Predicted data of CAP in time domain using generic model

3.10 Closest-Curve Model Method

In Closest-Curve Model (CCM) method, similar to the Generic Model method, the focus of the process and analysis are multiple curves (datasets). This method works by selection of one of the current datasets (curves) which has the closest EBI mean value to the dataset under prediction. The mean value of EBI is selected as a measure because of the useful properties it reflects from the dataset as a whole.

This method is different from all other methods used in this thesis in the sense that the output is not immediately/directly resulted from some mathematical computation, but it is the result of some comparison between some statistical measure (EBI mean value) of current curves and the new curve which is about to be predicted. In other words, the final model is decided and suggested based on the EBI data of the person for whom we want to predict CAP.

In this method, we should use the data of the patients in one particular body state, for predicting the CAP of individuals in the same specific body state, because EBI values may differ significantly in different states of the body. Hence, it might lead to confusion and complications if EBI values of two persons in different body states (rest, sleep, doing sports, etc.) is compared, and it should be taken into consideration.

3.10.1 Selection, Loading, and Preprocessing

Because of the distorted datasets (3.8.2, Faced Challenges), the first step is to choose a number of datasets and eliminate distorted ones. The loading of different datasets into Matlab is the next step. The files related to different patients are placed in a folder and the code starts to load the files in that folder into Matlab cell data structure. In the current work, files related to 14 patients were chosen to work with. The preprocessing is also conducted on all datasets. A detailed explanation of preprocessing can be found in 3.3.

3.10.2 Fitting curves on current data of multiple individuals

Although there is no need to curve fitting for finding the closest curve, there needs to exist fitted curves to the datasets for the purpose of model representation in the final step. Hence, we proceed similar to the previous method for fitting the curve, as mentioned in the following lines. When the data is ready from the previous step, 14 curves are fitted to the data of the 14 people, and the result is saved in Matlab cells. The process of fitting the

14 curves is not different from the previous fitting procedure in previous methods. The result of fitted curves for increasing and decreasing sections are presented in Figure 48.

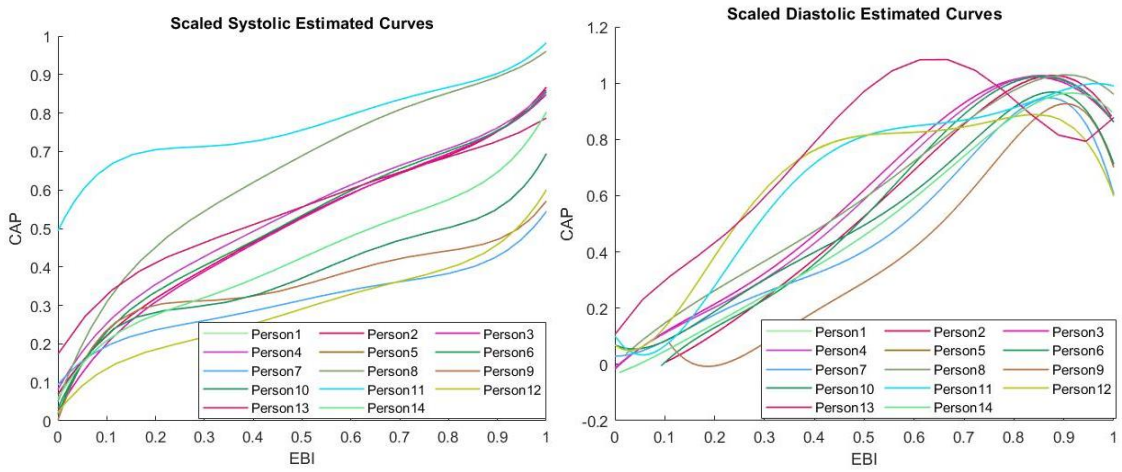

Figure 48 Curves fitted to EBI-CAP data of 14 persons. On the left, the curves related to increasing EBI (through time) is observed, and on the right the decreasing section of the EBI (through time)

3.10.3 EBI Mean value calculation

The mean values of the EBI of all 14 current curves are calculated at this stage. The result of the operation will be a row vector of dimension 1×14 :

$$\boldsymbol{b} = [\, avgEBI_{p1}, avgEBI_{p2}, \dots, avgEBI_{p14} \,] \qquad (16)$$

All 14 mean values are inserted in a Matlab cell structure.

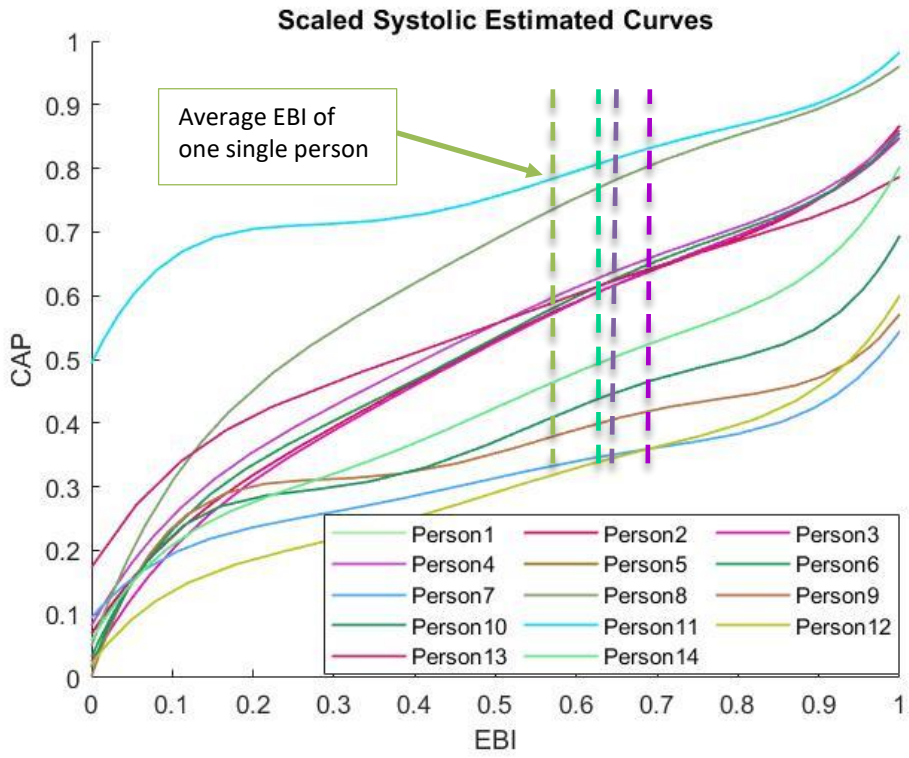

Figure 49 Demonstration of different EBI's average values for different people for increasing section of estimated curves

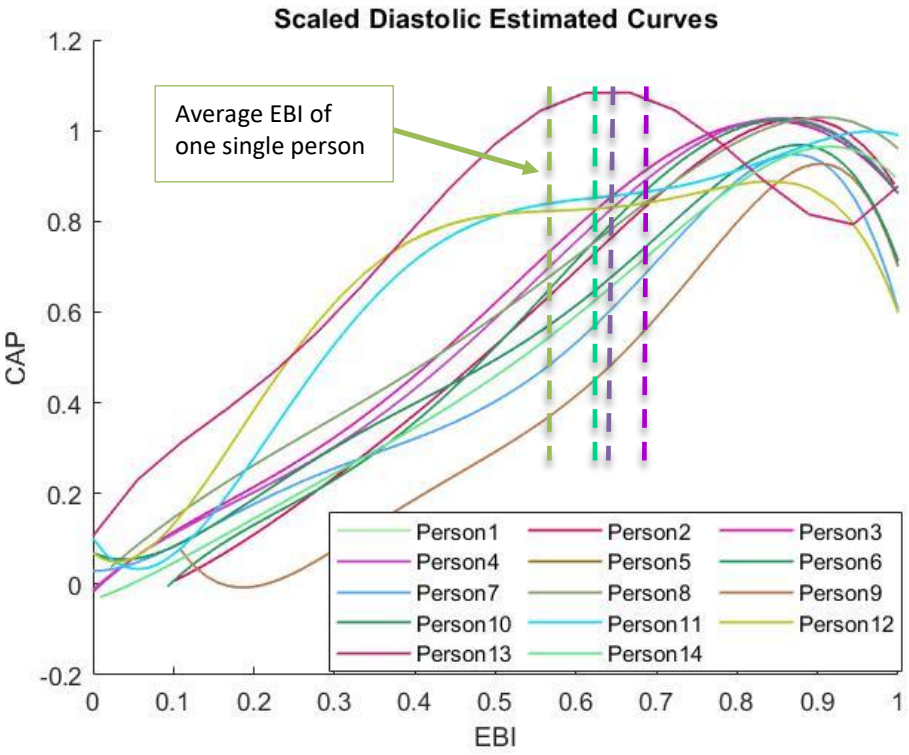

Figure 50 Demonstration of different EBI's average values for different people for decreaing section of estimated curves

At this point, the fundamental steps to develop the method is passed, but the model is not identified until the input dataset (the dataset for which we want to make predictions) is provided.

3.10.4 Inserting the input EBI data and acquiring its mean value

It is worthy to note that in this method, unlike other methods in this thesis, the model is not suggested (created) until some input data is provided. The mechanism of this method necessitates providing the input EBI (for which we want to predict the CAP) in order to get the proper model. As a result, in this step, the EBI information of an individual for whom we want to make CAP predictions is read, and is preprocessed. Following this, the mean of all EBI values of that individual (which are just read) is calculated. In the next step (3.10.5 this mean value is used to choose the best model the individual in question. We name this average $avgEBI_i$.

3.10.5 Comparison of mean values seeking the closest curve

For the matter of reaching the best possible model, in this method, the criterion is the proximity of mean value of input EBI, and the mean value of EBI of all previous individuals for whom we know the value of CAP. In other words, we should compare the value of $\boldsymbol{avgEBI_i}$ with every element of vector $\boldsymbol{b}$ which we previously calculated, and choose that element of vector $\boldsymbol{b}$ which is closest in value to $\boldsymbol{avgEBI_i}$. The index of the closest curve (in terms of mean EBI) to the input curve would be:

$$k \mid k \in \{1, 2, \dots, 14\}, |avgEBI_i - avgEBI_k| \; is \; min \qquad (17)$$

So, it follows that the model is closest-curve:

Closest Curve selected model for **person** $\boldsymbol{i}$ $with\ avgEBI_i$ is $\boldsymbol{curve_k}$

3.10.6 Selecting the found curve and choosing it as the model

When we find the closest curve, it is chosen as the final model, and all future CAP prediction needed for person i (the person for whom we read the input EBI) is made based on this model.

3.10.7 Testing

As for testing, a new dataset of a patient (one which has not been used for making the model) is loaded, and preprocessing operations are applied on its data in the same manner that they were applied on training data (correcting the outlier, splitting the dataset into increasing and decreasing sections and scaling the data). After this, the mean value of all input test data (EBI data) is computed. Then, as described in previous sections (3.10) the computed mean value is compared to the mean value of all processed datasets, and the dataset whose mean value is closest to the test input mean value is chosen as a good estimate. Now that we have the predicted model, we can compare the predicted model output (predicted CAP), with the test output (real CAP).

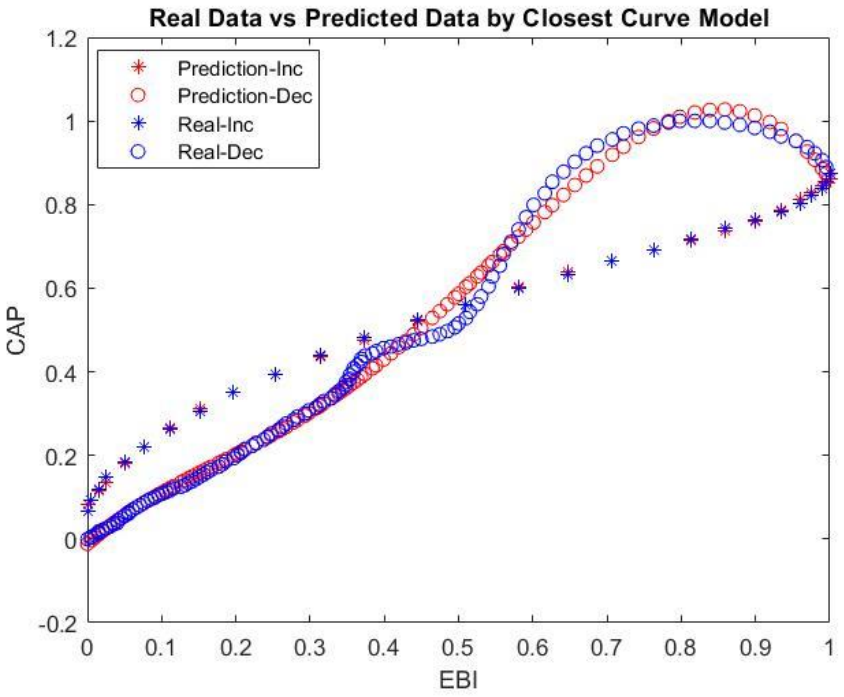

Figure 51 Real data vs Predicted data of EBI-CAP using closest-curve model

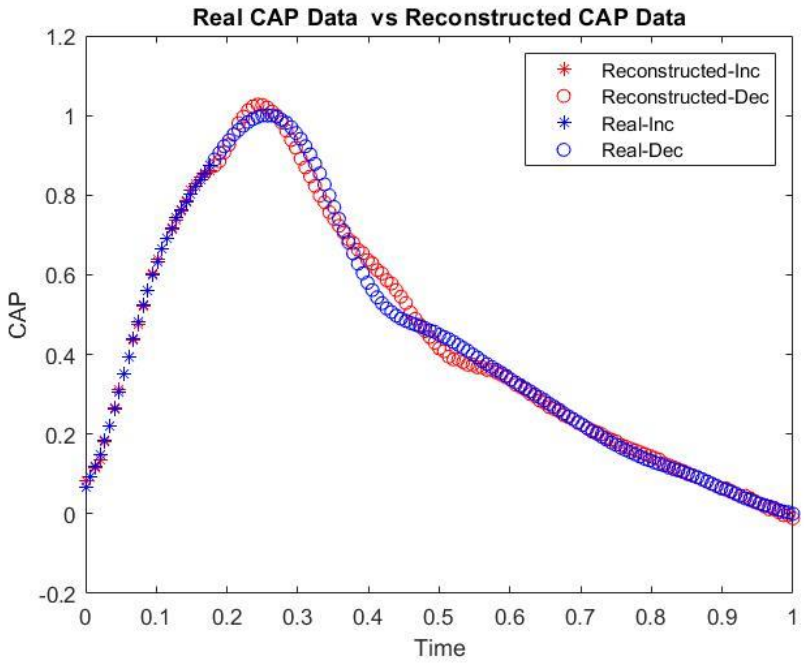

Figure 52 Real data vs Predicted data of CAP in time domain using closest-curve model

3.11 Neural Network Method

As discussed briefly in the method overview in 2.3, neural networks were chosen as the 5th method to provide a model able to predict the values of CAP based on EBI data.

3.11.1 Network type, configuration and parameters

The approach that was opted was to train the network based on current available input and output data (dependent and independent variables), and so supervised learning (SL) was put into effect. Regarding the choice of the mechanism of dataflow in the network, feed-forward networks seemed to function suitably taking the present datasets into account [24]. Feed-forward networks are widely used for variables with linear and non-linear dependency and relation, and in the case of this thesis's scope of data, it was expected that FF networks will provide results that are quite close to the predictions of networks trained with more complicated paradigms.

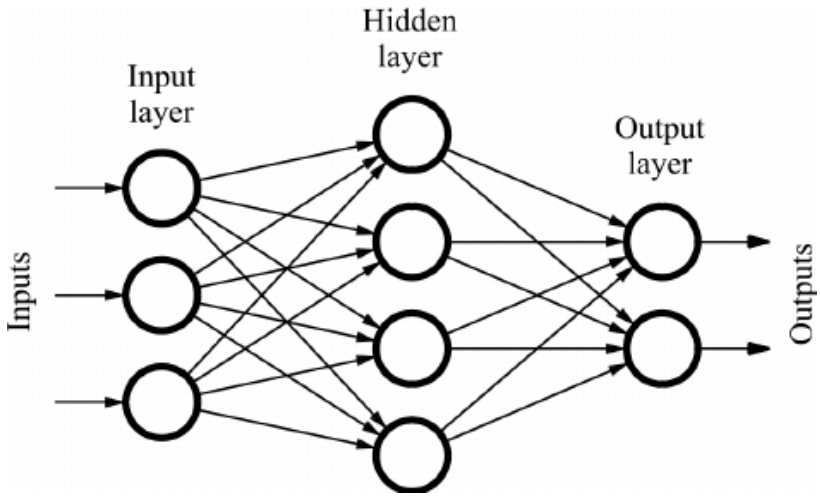

Figure 53 A three-layer, fully interconnected feedforward neural network [25].

As for configuration of the neural network, since the network does not have abundant quantity of data points and features, and since the relation for each curve does not show very complicated behavior, and also since the total run time of the algorithm is fairly short, it was not necessary to dig deeply into previous works to find out a proper configuration for the network. One thing which was kept in mind during the process was the problem of overfitting which can occur in situations like setting a more-than-enough number of epochs for the network. Solutions like early stop implementation were considered to avoid running into mentioned traps.

As for other parameters of the network, including hidden layers, and activation functions, values were selected with a balance of trial and error results, and similar network study results. The final choice showed similarities to common values that are chosen for conditions alike.

Table 21 Configuration for training of the neural network

Neural Network Configuration		
	Increase	**Decrease**
Epochs to Stop	1000	1000
Time to Stop	10000 seconds	10000 seconds
Accuracy to Stop	1e-11	1e-11
Training Accuracy	1.79e-3	1.79e-3
Hidden Layers	2	2
Hidden Layer Nodes	10	10
Activation Function	tansig	tansig
Training Algorithm	Levenberg-Marquardt (LM)	Levenberg-Marquardt (LM)

3.11.2 Organizing the data to feed into the network

The data to feed in the network should be organized in a data structure that stores independent and dependent variables of all data points. Moreover, the training toolbox requires a specific organization of data to work properly. As a result, Matlab matrices were chosen as the data structure for this purpose.

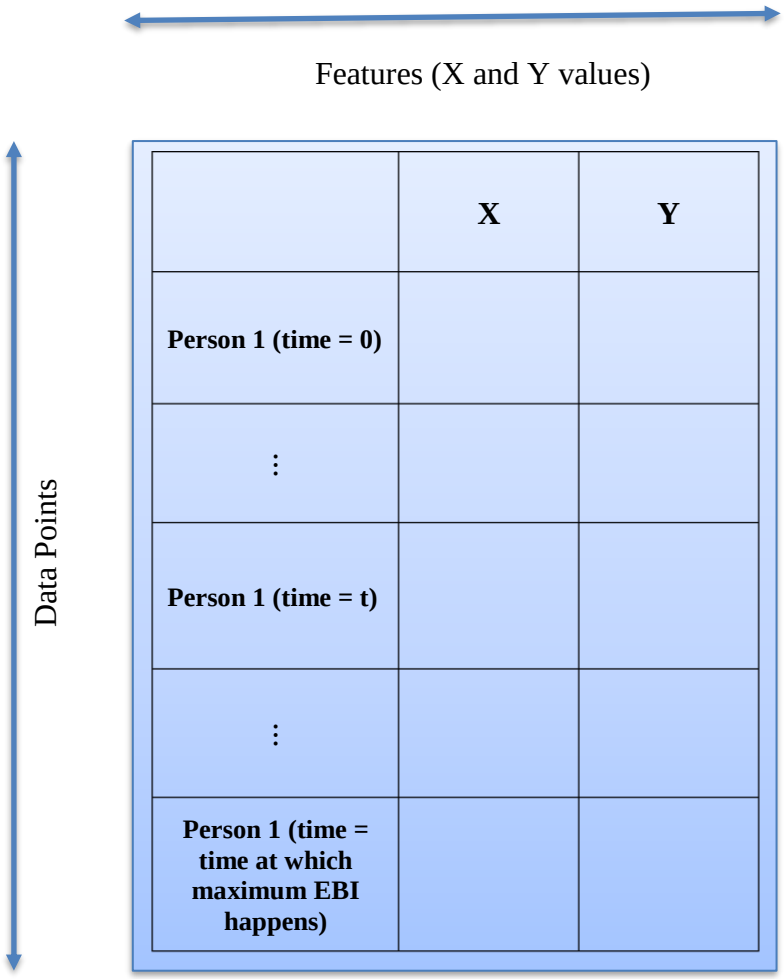

Figure 54 The arrangement of data for insersion in neural network input

3.11.3 Training the two networks (increasing and decreasing sections)

As mentioned in previous sections, for a better result, we need to make distinction between the increasing and decreasing portions of data when using regression, and the same consideration should be pondered and respected here. We train two networks for the two previously-splitted segments of data. One network is trained for systolic points, which shape the increasing section of EBI-CAP curve, and one network is trained for diastolic data points, which shape the decreasing section of the EBI-CAP function. For the prediction of data in later steps, the input should be splitted based on its place on the curve (increasing or decreasing section) and the trained model of the according section is used for prediction. In the next Figure, the schematic of the mechanism is observable.

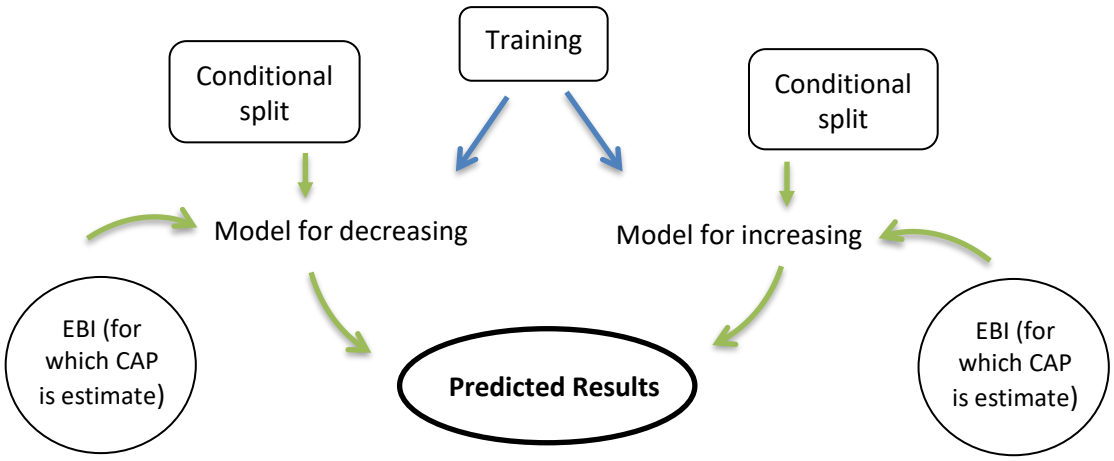

Figure 55 Schematic presentation of model mechanism

The training process took place in Matlab software, and the status of the training can be identified by the next figures. One of the figures (Figure 57) belongs to the training of decreasing section, and one of the figures (Figure 56) belongs to the training of the increasing section of the curve.

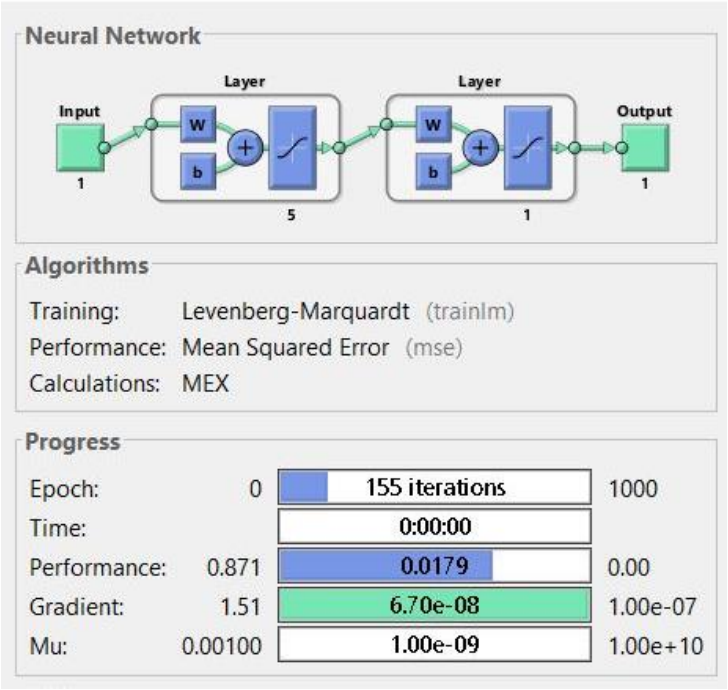

Figure 56 Trained neural network propertice for increasing section of dataset

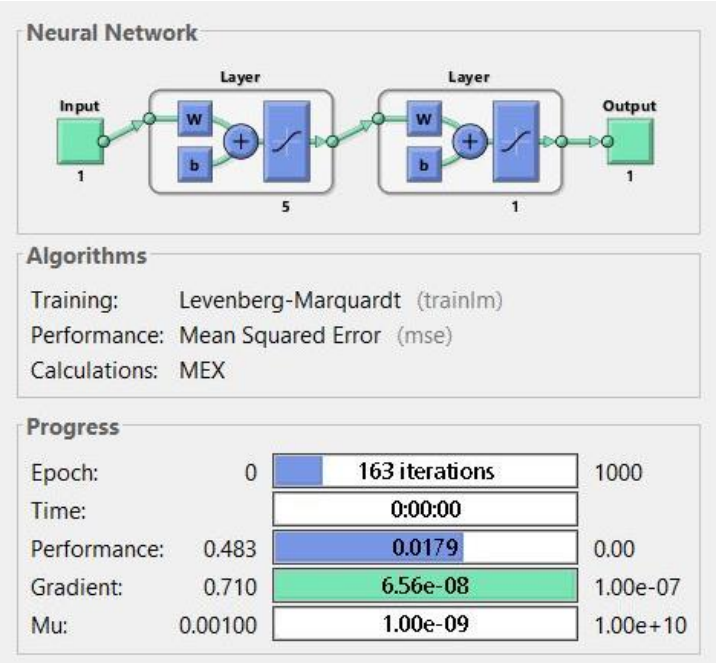

Figure 57 Trained neural network propertice for decreasing section of dataset

3.11.4 Testing and checking the accuracy

For the purpose of testing the network, we need to compare the real values of CAP of a test dataset, with predicted values of the CAP of the same dataset, and evaluate the result of the difference of these two values in a measurable way. The calculated difference using MSE for the tested datasets is roughly between 6% and 15%. Then, the EBI-CAP graph of the two sets of data is viewed, as observable in the next figure.

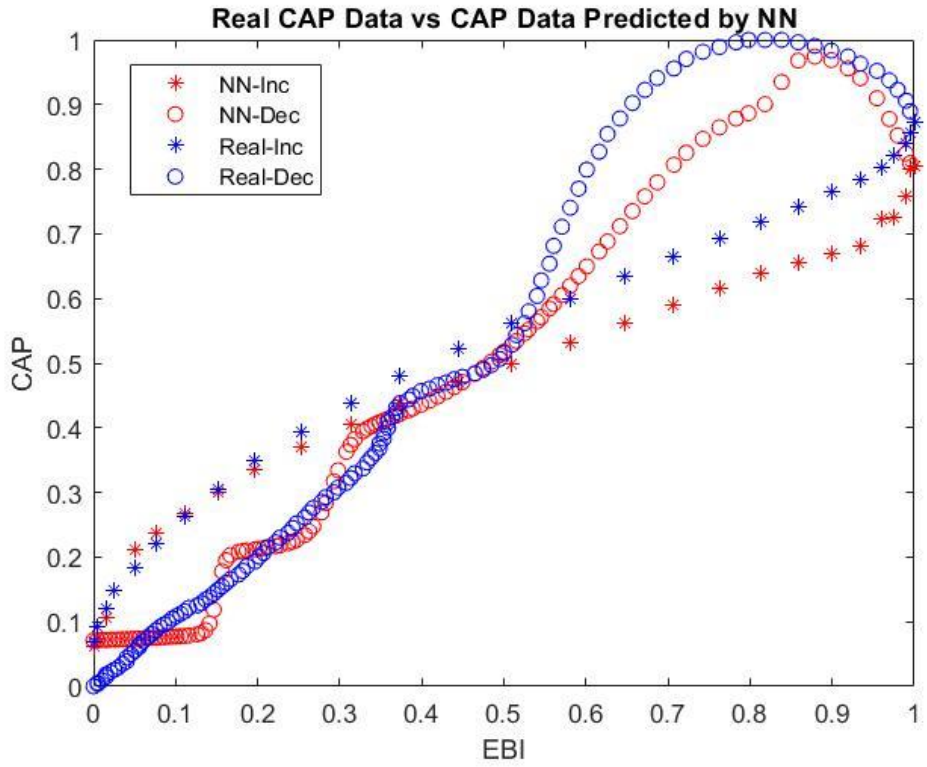

Figure 58 Real data vs Predicted data of EBI-CAP using Neural network model

Then the time-CAP data related to the real cap and predicted CAP is graphed (Figure 59). As suggested by curves and the calculated data, the method of neural network does not seem to be very promising for the amount of the training data that we have at hand, and probably, with a larger amount of data, better results are achievable.

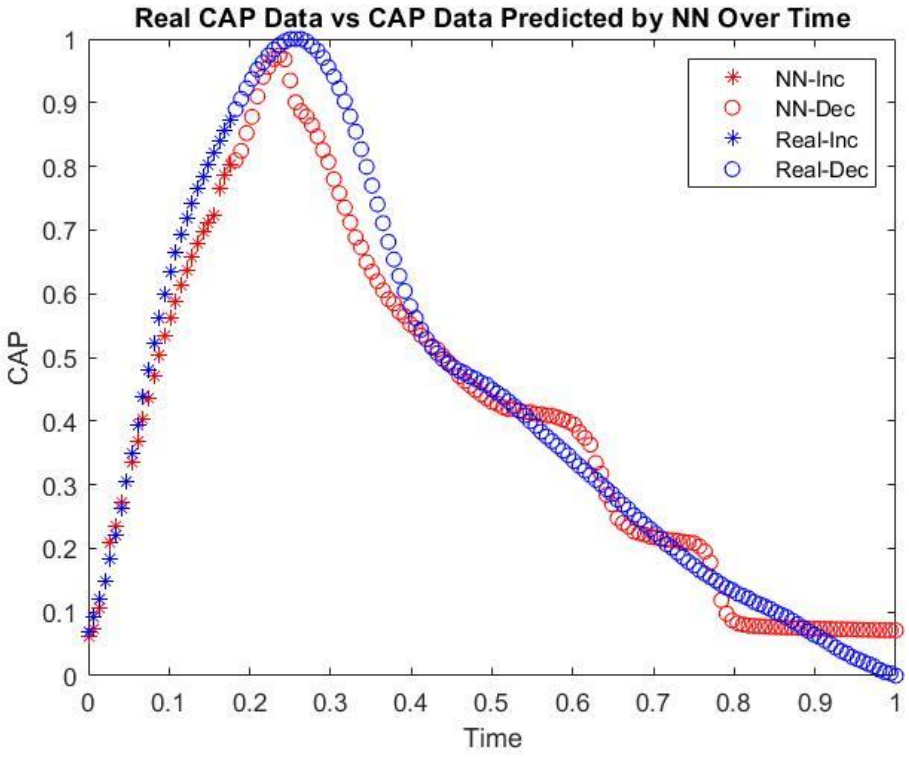

Figure 59 Real data vs Predicted data of CAP in time domain using neural network model

3.12 Comparison, discussion and analysis of the results

Throughout chapter 3 the aim was to come up with different methods to address the different challenges mentioned in the initial research problems, and to improve those methods. These methods try to present a way to determine CAP values based on input data. In total, 5 methods were presented, and two of them were used as supporting engines for the next three. Each of these methods offers specific advantages and positive points or may include negative points and limitations and we try to go through these in a summarized way.

3.12.1 Overview of the results

The first two methods of the thesis which focused on a single dataset (methods 1 and 2) and their results were investigated in previous sections. In the following paragraph, the latter methods (methods 3, 4 and 5) are under the focus.

- **Generic Model:**

  - Positive points / advantages:

    1. Because of the splitting of the dataset to two increasing and decreasing sections, an improvement of about 15% to 25% was observed in the fitting of the curve, specifically in the neighborhood of the maximum point, which in turn caused the final model to work more accurately.

    2. The whole procedure, from loading to prediction, is done automatically in the code, and no manual modification of the code is needed.

    3. The possibility of saving the models may be advantageous in the future for generating CAP data based on EBI. This is also the case for fitted curves of the patients, and they might prove beneficial for research purposes in the future.

    4. This method gives us the possibility of using fits of different orders for increasing and for decreasing sections.

  - Negative points / limitations / challenging issues:

    1. When CAP data of different datasets (persons) undergo high variations at a specific EBI, the mean value of CAP will be subject to noticeable

variations at that EBI, and this may have undesirable effects on the resulting model.

2. As the final result, we receive one single final model, which in the current thesis, does not count for different states and conditions of the body at the time EBI measurement (physical activity, age, sexuality, physical wellness, etc.). Since these changing conditions may have considerable results on EBI-CAP relation, the mentioned point might stand as a limitation of this model.

3. The accuracy of this model was observed to be less than the accuracy of closest-curve model. The possible reason is explained in the following lines, in closest-curve model advantages.

- Closest-Curve Model

- Positive points / advantages:

1. Because in this model we choose the CAP of one person with the closest EBI to the EBI of the person under prediction, the state and conditions of the person under prediction can change the resulting model used for this perosn, and therefore can change the CAP value that we get for this person. So, since this method accounts for changes between different people in different states (age of people, physical activity, physical wellbeing of people, etc), the estimation of CAP is more accurate, specifically in cases where changing conditions is a concern.

2. The first two positive points mentioned for generic model, can be stated as positive points of this method too.

- Negative points / limitations / challenging issues:

1. We assume that individuals with similar EBI behavior have similar CAP behavior (CAP and EBI are not two variables independent of each other).

- Neural Network

  - Positive points / advantages:

    1. This method shows better accuracy than the generic model.

    2. The first two positive points of the generic model are offered by this method too.

  - Negative points / limitations / challenging issues:

    1. Neural networks need a fairly large amount of data for training to function in an accurate manner, and produce better results.

    2. The accuracy of this model was observed to be less than the accuracy of the closest-curve model.

4 Summary

4.1 Conclusion

This thesis answered the following research questions:

- What is the relation between CAP and EBI of a person when both CAP and EBI of the person are available?

  In section 3.3, after preprocessing (outlier correction, and scaling), noiseless and clean datasets for the processing stage were acquired. By applying regression methods ("cftool" in Matlab), a polynomial function was presented to demonstrate the relation between the CAP and EBI values of the dataset.

- How to improve the goodness (accuracy) of the estimated relation between CAP and EBI?

  Two major actions were performed to improve the goodness of the fitted curve. These actions are splitting of the datasets, and selecting an optimized value for the order of the fit of the curves. Splitting the datasets helps in avoiding the error in extremums, and finding the optimized value for the order of the fit helps in avoiding the overfit and underfit issues.

- How to predict the value of CAP using the value of EBI when the CAP of the person is not available?

  Three main different methods (generic model, closest-curve model, and neural network), were provided to estimate the CAP value of a new person given the value of EBI of him/her.

- How to improve the accuracy of estimated CAP using the value of EBI?

  Three main different methods were presented to estimate the CAP value, generic model, closest curve, and neural network:

To evaluate the performance of these methods, an error was defined as the difference between estimated and actual values of CAP in the dataset.

For these methods, one action which significantly improved the prediction results (prediction quality) was using a variety of datasets (curves) to do proper training/fit, which in turn helped in considering data for different people at the same time.

Secondly, the splitting of datasets helped in decreasing the error. The results showed the error was around 40% for all the methods, without splitting the dataset. Then, by splitting the datasets to increasing and decreasing sections, the quality of prediction improved significantly (smaller error), and the error dropped to around 5% for the closest-curve model.

One other thing which made the improvement happen was the nature of each of the methods proposed. By the term nature, the characteristics of each method is meant. That is to say that each method proposes to solve the problem in a certain way and each of these ways will use a certain kind of training/fitting and mapping algorithm. Therefore, each of the methods comes up with different results that are characterized by the nature of the method. We tried to choose/improve/write these algorithms in a manner that the final prediction can be closer to the expected value. For instance, for the method of NN, we tried to choose one proper mechanism, configuration, architecture, etc., with the purpose of minimizing the error. This all helped to characterize the method and we could then hope to get a better result using that method tailored for our purposes, inputs, and for our needs.

Finally, the methods were ranked in terms of being accurate as follow:

(The best) 1. Closest-Curve Model

2. Neural Network

3. Generic Model

This was done using a knowledge discovery process including different phases: understanding the CAP and EBI data, data preprocessing, choosing an appropriate way to find a suitable model (suitable curve) to CAP data, and testing that model. The dataset contains Time, EBI and CAP data of 14 people, and is tested with the data of one or more people containing the same quantities.

4.2 Achievements

Achieving the main goals of this thesis relies on finding the answers to the proposed questions mentioned in the abstract, looked at from a different perspective in the problems statement section (1.1), and pointed out in the previous section (4.1). For the purpose and extent of this thesis work, the provided results showed proper accuracy and conformance with what was expected as the outcome of the research. **Therefore, it can be claimed that the initial goals of the research are achieved**, and even some extra steps were also taken toward the improvement of the results, although, like any other research work, there might exist some errors and challenging issues, and there is always place for improvement. In the next section, some recommended solutions are provided to address some of the limitations.

4.3 Recommendations

The following points can be recommended to further improve the results, to solve some of the issues faced during the research, and to address and overcome some of the limitations present during the work:

1. Improving the dataset by collecting more accurate data
2. Improving the dataset by collecting data of more patients
3. Improving the dataset by collecting more data of the same patient (also maybe data in different states, like doing sports, etc.)
4. Improving prediction methods and algorithms
5. Adding user interface (UI) to the code

6. Improving the interactive capabilities of the user interface of the model building section of the program, letting the user choose different processing methods, different curve fits (type and order), and different dataset paths.

5 References

[1] A. Waldin and K. Veeramachaneni, “Learning Blood Pressure Behavior From Large Blood Pressure Waveform Repositories and Building Predictive Models,” 2013.

[2] X. Chen and and R. M. D Xu, G Zhang, “Forecasting acute hypotensive episodes in intensive care patients based on a peripheral arterial blood pressure waveform,” in *n Computers in Cardiology*, 2009, pp. 545–548.

[3] A. P. Avolio, M. Butlin, and A. Walsh, “Arterial blood pressure measurement and pulse wave analysis-their role in enhancing cardiovascular assessment,” *Physiological Measurement*, vol. 31, no. 1. 2010.

[4] V. G. Almeida *et al.*, “Machine learning techniques for arterial pressure waveform analysis,” *J. Pers. Med.*, vol. 3, no. 2, pp. 82–101, 2013.

[5] C. M. McEniery, J. R. Cockcroft, M. J. Roman, S. S. Franklin, and I. B. Wilkinson, “Central blood pressure: Current evidence and clinical importance,” *European Heart Journal*, vol. 35, no. 26. 07-Jul-2014.

[6] “WHO | Disease burden and mortality estimates,” *WHO*, 2019.

[7] C. M. Douglas, A. P. Elizabeth, and V. G. Geoffrey, *Introduction to Linear Regression Analysis, 5th Edition*. 2012.

[8] S. Chatterjee and A. S. Hadi, *Regression Analysis by Example*. .

[9] S. Boyd and L. Vandenberghe, *Convex Optimization*. 2004.

[10] Y. Singer, “Advanced Optimization,” 2016.

[11] I. Maglogiannis, *Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in EHealth, HCI, Information Retrieval and Pervasive Technologies*. 2007.

[12] “Types of Machine Learning Algorithms You Should Know.” [Online]. Available: https://towardsdatascience.com/types-of-machine-learning-algorithms-

you-should-know-953a08248861. [Accessed: 02-Jan-2020].

[13] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. 2012.

[14] A. Cochocki and R. Unbehauen, *Neural Networks for Optimization and Signal Processing | Guide books*. 1993.

[15] "Types of Machine Learning | Now Enlighten Me." [Online]. Available: https://nowenlightenme.com/2018/03/18/types-of-machine-learning/. [Accessed: 02-Jan-2020].

[16] A. D. Almási, S. Woźniak, V. Cristea, Y. Leblebici, and T. Engbersen, "Review of advances in neural networks: Neural design technology stack," *Neurocomputing*, vol. 174, pp. 31–41, Jan. 2016.

[17] T. D. Sanger, "Optimal unsupervised learning in a single-layer linear feedforward neural network," *Neural Networks*, vol. 2, no. 6, pp. 459–473, 1989.

[18] T. Coutinho, "Arterial stiffness and its clinical implications in women," *Canadian Journal of Cardiology*, vol. 30, no. 7. Pulsus Group Inc., pp. 756–764, 2014.

[19] M. Min, H. Kõiv, E. Priidel, K. Pesti, and P. Annus, "Noninvasive Acquisition of the Aortic Blood Pressure Waveform," in *Wearable Devices [Working Title]*, IntechOpen, 2019.

[20] J. Solà, A. Adler, A. Santos, G. Tusman, F. S. Sipmann, and S. H. Bohm, "Non-invasive monitoring of central blood pressure by electrical impedance tomography: First experimental evidence," *Med. Biol. Eng. Comput.*, vol. 49, no. 4, pp. 409–415, Apr. 2011.

[21] G. A. COHN and R. KUSCHE, "BIOIMPEDANCE BASED PULSE WAVEFORM SENSING," US 2018 / 0078148 A1, 2018.

[22] A. Krivoshei, H. Uuetoa, M. Min, J. Lamp, P. Annus, and T. Uuetoa, "Estimating the Transfer Function between the CAP and Radial EBI Cardiac Periods: Use of PCA for Dominating Spectral Features Analysis," 2015.

[23] “Evaluating Goodness of Fit - MATLAB & Simulink - MathWorks Switzerland.” [Online]. Available: https://ch.mathworks.com/help/curvefit/evaluating-goodness-of-fit.html. [Accessed: 02-Jan-2020].

[24] D. Brezak, T. Bacek, D. Majetic, J. Kasac, and B. Novakovic, “A comparison of feed-forward and recurrent neural networks in time series forecasting,” in *2012 IEEE Conference on Computational Intelligence for Financial Engineering and Economics, CIFEr 2012 - Proceedings*, 2012, pp. 206–211.

[25] J. M. Brotzer, E. R. Mosqueda, and K. Gorro, “Predicting emotion in music through audio pattern analysis,” in *IOP Conference Series: Materials Science and Engineering*, 2019, vol. 482, no. 1.