Aalborg University

Biomedical Engineering and Informatics

Master's thesis

Forecasting depth of anesthesia in laboratory pigs using mixed-data neural network

Written by:

Tobias Brinkmann Duncker

Main supervisor:

Winnie Jensen

Co-supervisor:

Thomas Gomes Nørgaard dos Santos Nielsen

AALBORG UNIVERSITY
DENMARK

01/06/2021

AALBORG UNIVERSITY
Student Report

The Faculty of Medicine
Biomedical Engineering and Informatics
Niels Jernes Vej 10
9220 Aalborg Øst
www.hst.aau.dk

Title:

Forecasting depth of anesthesia in laboratory pigs using mixed-data neural network

Project:

Master's thesis

Project group:

21gr10405

Members:

Tobias Brinkmann Duncker

Project period:

02/02-2021 to 01/06-2021

Main supervisor:

Winnie Jensen

Co-supervisor:

Thomas Gomes Nørgaard
dos Santos Nielsen

No. of pages: 65
Appendix: 5
Finished: 01/06-2021

Synopsis

Hvis dybden af anæstesi i forsøgsdyr bliver for lav, kan forsøgsudføreren risikere, at forsøgsdyret vågner op. Dette forringer kvaliteten af forsøgsdata, da dyrets bevægelser og aktivering af kroppens systemer, f.eks. ved forøget puls, kan uhensigtsmæssigt påvirke forsøgsdata. Hvis dybden af anæstesi bliver for dyb, kan forsøgsudføreren risikere, at forsøgsdyret enten dør eller at dybden påvirker optagelsen af forsøgsdata ved at dæmpe kroppens systemer for meget. Ydermere vil begge tilfælde etisk set være en unødvendig situation, som kan skabe unødvendig smerte eller død, hvilket kræves formindsket mest muligt af europæisk lov.

For bedre at kunne kontrollere dybden af anæstesi er det relevant at undersøge muligheden for at forudsige ændringer i dybden af anæstesi. I dette projekt forudsiges ændringer i dybden af anæstesi 12 minutter ud i fremtiden. Dette gøres ved at skabe et neuralt netværk, som både kan analysere udtrukne features fra et elektroencefalogram, samt ved også at analysere elektroencefalogrammet som tidsseriedata ved at benytte lag i det neurale netværk til analyse.

Resultaterne viser, at et neuralt netværk som kombinerer analyse af simple features og analyse af det fulde elektroencefalogram bedre kan forudsige et skift i dybden af anæstesi, end hvad neurale netværk som udførte de analyser separat kunne forudsige. Ydermere viser resultaterne også, at der i nogle samples er en stor sikkerhed i et fremtidigt skift, hvilket er relevant at undersøge til fremtidig forskning.

Preface

This project was created by Tobias Brinkmann Duncker in project group 21gr10405 as part of a Master's thesis in Biomedical Engineering and Informatics at Aalborg University between the 2nd of february and 1st of june 2021.

Special thanks goes to Winnie Jensen and Thomas Gomes Nørgaard dos Santos Nielsen at Aalborg University for project supervision and guidance. Additional thanks goes to Taha Janjua for preparation of data and constructive discussions.

Reference instructions

This project used the Harvard refrence method. The bibliography is sorted according to surname of the first author. Both active and passive references were used. For passive references, the position in the text indicates how much of the paragraph is supported by the reference. If the reference is placed before a period, it supports all previous text until any period. If the reference is placed after a period, it supports all text before the reference in the same paragraph even past other periods, except sentences supported by references before a period. Thus, all text in a paragraph with a reference at the paragraph is supported by said reference, except any sentences inside the paragraph that have their own reference attached before a period.

Contents

1 Introduction

Scientific research is often required to use non-human models in some types of experiments for ethical reasons. Animal models are used both in basic research but also testing of medical drugs or equipment. As an example for basic research, a study investigated pain in pigs by following them at the age of 2 to 6 months old and showed developmental changes in mammalian pain response pathways, which is difficult to investigate in human models, because deliberately causing such pain in humans of similar age is not considered ethical. In regards to testing, animal models are used as a first safety check before testing on human models. [Gieling et al., 2011] Therefore, animal models may be used as a substitute to human models for early research whose use is expected to later be translated into clinical studies with humans. [EUR-Lex, 2010; Kobeissy et al., 2016; Speaking of research; Tannenbaum and Bennett, 2015]

General anesthesia in relation to animal models is important to research for several ethical reasons. By European law, it is mandatory to apply general- or local anesthesia as needed to keep pain, suffering and distress to a minimum, unless methodologically inappropriate. [EUR-Lex, 2010]

Furthermore, research using physical animal models depend on the ethical principles of the 3R's as mandated by European law, of which general anesthesia may conflict with two of the principles: Reduction and Refinement, if the depth of anesthesia (DoA) is not controlled properly. The principle of Reduction considers the experimenter ethically obligated to reduce the amount of required animals for experimentation and to not use an unneeded amount of animal lives, while the principle of Refinement requires the experimenter to ease any unnecessary pain and discomfort. [EUR-Lex, 2010; Tannenbaum and Bennett, 2015] As such, if the DoA is too light, subjects might experience intraoperative awareness with or without pain, or explicit recall postoperatively, which go against the principle of Refinement [Montupil et al., 2019]. Although difficult to verify in animals, intraoperative awareness is commonly described as the worst experience of their lives of human patients. [Silva and Antunes, 2012] Jahanseir et al. [2018] also notes how light anesthesia can cause heart arrhytmia in human patients, if they have a history of heart disease, which theoretically could apply to experiments where pigs are given heart diseases, although not verified. Too deep DoA can cause the unintended intraoperative death of subjects which go against the principle of Reduction by possibly causing a need for new subjects. [Tannenbaum and Bennett, 2015] Additionally, prolonged intraoperative periods of too low values of electroencephalogram (EEG) indexes have shown correlation to long-term mortality in humans, while proper monitoring of anesthesia also reduces post-operative side-effects significantly [MacKenzie et al., 2018; Montupil et al., 2019].

Besides ethical concerns, it is also important to consider the methodological concerns of having a properly controlled DoA.

Too light DoA can cause unintended somatic-, physiological or electroencephalographic events. Somatic events can be the occurence of movement, which can interfere with the laboratory setting [Montupil et al., 2019]. Such movement might also elicit an unintentionally measured signal response due to muscle movement or neurological activation, depending on how measuring of the wanted signal is done. Physiological events can be related to hemodynamic responses like bleeding, changes in blood pressure or other similar physiological changes that might present itself as systemic [Montupil et al., 2019]. Furthermore, the quality of EEG data might worsen, if the subject becomes too conscious and this consciousness affects for example EEG directly or if the consciousness might cause other effects, that could muddle the measured data, e.g. unintended pain, which would show in physiological parameters or directly affect the purpose of an experiment [Haberham et al., 1999].

Too deep DoA for prolonged intraoperative periods is associated with increased long-term mortality as previously mentioned, which might cause problems if experiments require surgery and postoperative long-term surveillance. Simultaneously, it may cause increased long-term morbidity in humans, which further might interfere with correlation and causation of experiments if similar happens in pigs [MacKenzie et al., 2018; Montupil et al., 2019]. During conventional anesthetic practice, the tendency is to overdose to avoid intraoperative awareness, which might make the problems of too deep DoA a more frequent problem than too light DoA [Montupil et al., 2019].

Another problem of improperly controlled DoA is the stability of maintenance. For any system assessing a physiological value or a state like DoA, it is important to avoid changing it too much. When a value, e.g. a number representing DoA, increases or decreases too much, the problem can be alleviated by trying to influence it to change back, e.g. by increasing or decreasing titration of the anesthetics agents. Though, this can be done too fast, which would risk the change to shift the value from one side of the spectrum to the another, while the intended result is the middle of the spectrum. This situation can repeat every time an experimenter tries to solve the problem, e.g. unintended DoA, which in turn could cause an oscillating effect of the DoA. Such an oscillating effect of the body's physiology could be harmful to the patient's health. [Pauldine et al., 2008] Additionally, it could cause confusing information in regards to data collection of an experiment, if the oscillating effect were to have an effect of what is measured for the purpose of the experiment.

Overall, both the ethical and methodological concerns make it important to investigate methods to better assess and control the DoA.

Problem analysis 2

2.1 Use of animal models

This section will elaborate on what type of research animal models are used for and summarize relevant practical and biological considerations between two popular animal models: Pigs and rodents.

2.1.1 Scientific purpose of and ethical consideration of animal models

The nervous system is frequently used in research involving animal models, and this system is also one which often requires general anesthesia. [European Commission, 2019]

Animals used for scientific purposes are primarily being used for basic research (45%), translational and applied research (23%) and regulatory use (23%). Of the basic research category in descending order of size, 22% of animals were used for multisystemtic research, 15% for the nervous system and 13% for oncology, thus making the nervous system one of the most widely researched areas in terms of number of animals used. [European Commission, 2019]

Researchers are also required to report the severity of use experienced by an animal under an approved procedure using the categories in table 2.1. If general anesthesia is used, the severity of use is reported as at least moderate, if the procedure involves any kind of surgery under general anesthesia. General or local anesthesia is to be used unless methodologically inappropriate in concern to the scientific purpose. [EUR-Lex, 2010] Thus, any procedure requiring general anesthesia is considered important in regards to the principles of the 3Rs, since the default severity is at least moderate.

Category of severity of use	Frequency	Definition
Mild	51%	Short-term mild pain, suffering or distress, and no significant impairment of the well-being of the animal
Moderate	32%	Short-term moderate- or long-lasting mild pain, suffering or distress, as well as moderate impairment of the well-being of the animal
Severe	11%	Short-term severe- or long-lasting moderate pain, suffering or distress, as well as severe impairment of the well-being of the animal
Non-recovery	6%	Procedures performed entirely under general anesthesia from which the animal does not recover consciousness

Table 2.1. Explains the severity categories of the mandatory reporting as mandated by European legislation [EUR-Lex, 2010]. Frequency of reported categories is from 2017[European Commission, 2019]

The category of basic research is responsible for most cases of severe use is the nervous system. Although the frequency between severity of use pr. category of basic research use could not be found, basic research into the nervous system may be expected to have a different frequency of severities of use, with less cases being mild and more being non-recovery in regards to encephalographic research and because pain research is related to the nervous system, where it might be needed to expose and excite nerves. This further underlines the importance of proper control of general anesthesia, especially for research of the nervous system, which has the biggest use of animal models[European Commission, 2019].

2.1.2 Choice of animal model

Several factors play into the choice of animal model. These factors can be practical, technological or biological. Many different species and breeds are used, where the most common choice of animal in neurological research is either a rodent or a pig. [Biocompare, 2014; Kobeissy et al., 2016; Walters et al., 2017] Of the 9.39 million animals used for scientific purposes in the European Union in 2017, 61% were mice, 12% were rats and 0.8% were pigs. [European Commission, 2019]

Practical considerations

Animals have to be acquired, stored and maintained. Rodents have been the most common choice, because they are easy to handle, have fewer facility requirements, can be bred in large numbers at a fast pace and cost relatively little. Prices also depending on species and breed, where an outbred Swiss mouse may cost 10 Euro and an inbred Balb/c mouse may cost 16.5 Euro [The University of Adelaide, 2020]. Sprague Dawley rats may cost between 9-26 Euro dependent on age [Janvier Labs, 2017]. A livestock pig pig may cost 480 Euro [National Swine Resource and Research Center, 2020], while a mini-pig may cost 1.625 Euro [Ellegaard Góttingen Minipigs A/S, 2021]. Rodents are also easier to breed and faster at maturing, which makes them more practical, though this also makes them

unreliable for chronic experiments. [Biocompare, 2014; University of Michigan Medical School, N/A; Walters et al., 2017]

A mouse may be required to have at minimum around 0.03 square meter of space, and the cages can often be stacked. Comparatively, pigs require around one to five square meter of space depending on size, which could be a small mini-pig or a larger Hampshire pig as used for livestock. Costs which might include bedding, food, anesthesia or other facility requirements have a corresponding increase in price relative to the size, which makes pigs significantly more expensive compared to e.g. a mouse. [Biocompare, 2014; EUR-Lex, 2010; Walters et al., 2017] The use of a mini-pig instead of a larger breed can reduce the amount of drugs needed, which is useful in safety and efficacy trials using animal models [Gieling et al., 2011].

Since rodents are used more than pigs, researchers rarely have the expertise required to care for or facilitate pigs [Walters et al., 2017]. Acquiring the facilities and necessary training may not be possible for all research departments in terms of cost, although when the investment is done, it likely wouldn't require much to maintain the facilities and training. Additionally, acquiring the large breeds of pigs is relatively easy considering the widespread use of them as livestock in regards to species like Hampshire swine [Walters et al., 2017].

Technological considerations

Different species also have different advantages in regards to the equipment that is either being used in procedures or being tested. Equipment designed for human use can be used on larger animals like pigs. Magnetic resonance imaging (MRI), positron emission tomography and surgical procedures translate well from human use to pig use. Smaller versions of the mentioned diagnostic machines do exist for rodents and are cheaper, while the use of such machines for human use is common in hospitals and can be used on pigs. [Biocompare, 2014; Walters et al., 2017]

Larger animals also have the advantage of being a better model for testing devices in terms of size, since testing a pacemaker or prosthesis for humans on rodents would otherwise require shrinking of the technology, which even if done might give data that doesn't translate well into human use purely for size reasons. The same is true for procedures like organ transplantation. [Biocompare, 2014; Gieling et al., 2011; Walters et al., 2017]

Biological considerations

A large amount of clinical trials fail due to poor translation of animal trials into clinical trials caused by little biological similarity between the human and animal models and is estimated to be the cause of approximately 30% of clinical trials of drugs in development. [Berge, 2011] For this reason, some suggest replacing animal trials with human volunteers in research related to pain. [Mogil et al., 2010] Such a proposal would be in line with the first principle of the 3R's Replacement which suggests entirely removing the use of animal models when possible, while a faster change to clinical trials for the same reason of low biological simmilarity would be in line with the Reduction principle by using fewer animals. [EUR-Lex, 2010]

Pigs have many advantages over other species from a biomedical research perspective

and their similarity to humans have been more widely recognized in recent times. Pigs are already considered the optimal animal model for the fields of drug discovery, xenotransplantation and risk assessment of environmental contaminants. For regenerative medicine products in the fields of cardiovascular, orthopedic and wound-healing, pigs were chosen in 56% of cases despite their relative difficulty in maintenance compared to rodents. [Gieling et al., 2011; Walters et al., 2017] This is due to how pigs are considered more similar to humans in terms of anatomy, physiology and pathophyisology. Phylogenetically, pigs are also considered three times more similar to humans than mice are, when comparing nucleotides. [Walters et al., 2017] Pigs have similar homeostatic mechanisms e.g a similar hemorrhagic shock response [Kobeissy et al., 2016], which are important when dealing with anesthesia, as the effect of the body's response to hemorrhage can influence the effect of general anesthesia on the brain [Kurita et al., 2005]. The pharmacokinetic processes in pigs are similar to human in terms of absorption, distribution, metabolism and excretion, and are suited as a substitute for nonhuman primates [Gieling et al., 2011].

Their brain is generally more similar anatomically. The size of a human brain is 1.300-1.400 gram, compared to a pig brain of 80-180 gram (high interbreed variance), a rat brain of 2 gram and a mouse brain of 0.5 gram. The pig brain has an increased anatomical complexity relative to other animals like rodents, which may be explained by the size difference of the brain and this allows the possibility of more detailed data of regional differences. Additionally, the distribution of neurotransmitter uptake sites in pigs is comparable to that of humans, making the research more reliable to compare to humans than other animal models. The porcine brain is also gyrencephalic like a human brain, compared to rat and mice whose brains are lissencephalic. This difference in smoothness, ridges and furrows is expected to cause poor translational value research-wise, which is why pigs are the animal of choice when developing tracers for positron emission tomography due to their large gyrencephalic brain. [Gieling et al., 2011; Kobeissy et al., 2016; Walters et al., 2017]

Genetic engineering has also been done on rodents and pigs, and has proven near invaluable for some advances in biomedical research. [Walters et al., 2017] The use of genetically altered animals has seen a slight increase, although mostly in mice. Of the genetically engineered strains, 95% of them were used for basic research. 22% of this research was multisystemtic, 15% was for the nervous system and 13% for oncology. [European Commission, 2019] Thus, although it's only 15%, this is a sizable part. Especially for pigs, the technology to genetically engineer an animal has advanced a lot over the past three decades with the most recent and most significant being gene-editing technologies like clustered regularly interspaced short palindromic repeats, commonly known as CRISPR. Other major advances include somatic cell nuclear transfer[Walters et al., 2017], which most researchers would remember associate with the 1997 cloning of the first sheep "Dolly" [Edwards, 1999].

Altogether, pigs have a relatively high similarity to humans compared to rodents, especially in regards to the nervous system, which is the most widely investigated field of basic research using animal models. Other systems of the body that might affect neurological research or anesthesia, e.g. the cardiovascular system, are also similar.

Summary of advantages and disadvantages between pigs and rodents

A summary of the difference between pigs and rodents which factors in choice of animal model can be seen in tabel 2.2. As shown, rodents may be in more widespread use due to their low price and small space requirements. Pigs are more similar to human in most aspects than a rat, although they are considerably more expensive and require a lot more space.

Thus, although rodents are much more widely used, some fields have a lot to gain by using pigs instead of rodents. Regardless of whether they are used in a natural state or as a genetically engineered strain, pigs may become a biomedical model of choice because of their similarity to humans. [Walters et al., 2017]

	Pigs	Rodents
Frequency of use in the EU	0.8%	12%
Acquisition	Widespread agricultural use	Widespread scientific use and faster breeding
Cost	480 Euro for a livestock pig 1625 Euro for a mini-pig	10-16.5 Euro for mice 9-26 Euro for rats
Space	1-5 m^2	0.01 m^2 for mice
Facility and food cost	Large because proportional to size	Small because proportional to size
Technological research	Size and anatomical similarity allows testing of human pacemaker and prosthesis	Small size prevents testing of many devices for human use
Equipment opportunity	Equipment designed for humans can often be used, e.g. diagnostic equipment like MRI. Such equipment tend to be available in human hospitals.	Smaller variants of diagnostic equipment like MRI exist and are cheaper to purchase
Phylogenity	More similar to humans	Less similar to humans
Pharmacokinetics	More similar to humans, also sometimes used as a substitute for nonhuman primates	Less similar to humans
Cardiovascular	Similar responses, e.g. due to hemorrhage, making them more reliable for general anesthesia	Less similar to humans
Neurological	Gyrencephalic brain like humans, increased regional complexity for easier research, similar neurotransmitting system, large brain size (80-180 gram)	Lissencephalic brain, less similar to humans, small brain (0.5-2 gram)

Table 2.2. Summarizes the previous sections of 2.1.2, 2.1.2 and 2.1.2.

2.2 Control of anesthetic depth in laboratory animals

This section will explain different ways to assess DoA and then elaborate why measuring anesthesia is difficult.

2.2.1 Assessment of somatosensory responses

Discomfort and pain in animals is generally assessed based on visual and subjective assessment of physiology, behaviour and somatosensory reflexes. In some cases, EEG is used as a supplement. [Baars et al., 2013; Silva and Antunes, 2012].

Examples of somatosensory assessment include testing of the reflex of an animal to right itself or to respond to stimuli of specific nerves when pinching. For example, rodents are considered unconsciousness if they lose their ability to automatically try to right themselves, when they are placed on their back. Though, some anesthetic agents contain neuromuscular blocking agents, which compromises the reliability of somatosensory reflexes by paralyzing a conscious patient. The reverse can also happen, where dissociative agents can keep a patient dissociated from pain, but cause the activation of somatosensory reflexes. It is a commonly used method in both human and veterinary patients, because it provides a clear distinction between consciousness and unconsciousness, although it is subjective and can be error-prone. [Baars et al., 2013; Jahanseir et al., 2018; Silva and Antunes, 2012]

Although the activation of somatosensory reflexes can be seen as a clear distinction between consciousness and unconsciousness, it can be seen as counter-intuitive to the fact that the difference between these two states are rather gradual and continuous from the perspective of EEG, instead of binary as assumed when using somatosensory reflexes. [García et al., 2021; Martoft et al., 2002] This can cause confusion in regards to what true unconsciousness is, when using general anesthesia, and is further complicated by the appearance of a clear distinction in EEG called burst suppression patterns. Burst suppression is considered to be the distinction between light and deep anesthesia. [Haga et al., 2011] This is in contrast to activation of somatosensory reflexes, which are seen as the distinction between no anesthesia and light anesthesia.

Thus, assessment of somatosensory reflexes can be used in practice, the reflexes might be suppressed by the anesthetic agents, and the method does not sufficiently provide a graduated impression of the DoA besides the distinction between no anesthesia and light anesthesia. Therefore, it is relevant to investigate better methods of assessing the DoA in animals. [Baars et al., 2013; Silva and Antunes, 2012]

2.2.2 Assessment of physiological responses

Examples of physiological assessment includes changes in hemodynamics, breathing behaviour and gasses as a result of breathing. Besides somatosensory reflexes, the assessment of physiological parameters is also a common way to assess the DoA. Though, it is less reliable than somatosensory reflexes [Baars et al., 2013; Gu et al., 2019]. One of the most common parameters is end-tidal CO^2, which tend to be one of the first parameters to fall around a minute after the introduction of an anesthetic problem, followed by arterial hypotension (1-4 minutes late) and then changes in electrocardiography and peripheral blood saturation (around seven minutes). [Gayer et al., 2016] Other parameters

include heart rate, breathing rate [Jahanseir et al., 2018], temperature, anesthetic agent during inhalation and after expiration [Haga and Ranheim, 2005] and minimum alveolar concentration [Haga et al., 2011].

2.2.3 Assessment of electroencephalographic responses

Because the main target of hypnotic anesthetic agents is the central nervous system, it is common to look into the usage of EEG when trying to control the DoA. Decisions regarding changing thetitration of anesthetic agents is then expected to be based on the degree of cortical depression witnessed in EEG. Although physiological parameters can be used, it is deemed less useful to investigate indirect ways of measuring the anesthetic effect of the central nervous system, when EEG is expected to directly measure the effect. For this reason, recent research is suggesting to investigate the usage of EEG to assess the DoA [Silva and Antunes, 2012; Tacke et al., 2020]. Although cortical information is considered theoretically optimal for the assessment of DoA, the evaluation of such information is difficult and has no golden standard [Haga et al., 2011]. Examples of methods of using EEG to assess DoA includes the use of cortical frequencies, calculations of entropy, and burst suppression. [Silva and Antunes, 2012] Previously, decision support systems assessing the DoA in the 1990s used physiological parameters, but this was changed to EEG-related methods because of their higher accuracy. [Hashimoto et al., 2020] This may in part be due to how cortical information is nonlinear, which requires different methods of analysis than usually done with simpler physiological parameters. [Gu et al., 2019]

The use of frequency can be used to divide cortical information into different frequencies. This can be used in individual calculations used to assess DoA, e.g. spectral edge frequency 95% which represents the frequency of which 95% of the total power lies below, or the median frequency. [Haga and Ranheim, 2005; Martoft et al., 2002; Silva and Antunes, 2012] In a non-sedated state but alert state with closed eyes, the most prominent frequency band is alpha (10 Hz). [Jahanseir et al., 2018] Then, the effect of a hypnotic anesthetic agent can be seen described in table. 2.3. Besides this effect on specifically the alpha-, beta- and delta frequency bands, some studies also include the theta- and gamma band. [Hashimoto et al., 2020] Both the absolute power of the frequency bands and the power relative to other bands are used, because the EEG shows different characteristics at various states of anesthesia. [Jahanseir et al., 2018] Especially delta waves are expected to represent a lot of the information from cortical cells that fire during either sleep or anesthesia and would therefore be a sign of a deep DoA. [García et al., 2021]

Depth of Anesthesia	Characteristics
Initial sedation	Beta activity increases (13-25 Hz)
Light anesthesia	Beta activity (13-30 Hz) decreases, while alpha- (8-12 Hz) and delta activity (0-4 Hz) increases.
Moderate anesthesia	Similar effect as with light anesthesia, but bigger changes
Deep anesthesia	Burst suppression appears
Very deep anesthesia	Isolectric state appears

Table 2.3. Describes the characteristics of the DoA as described in [Jahanseir et al., 2018].

These frequency bands are then used in methods that provide simple indexes to assess the

DoA in humans, although such indexes does not exist for animals [Silva and Antunes, 2012]. The most commonly used is bispectral index (BIS), which categorises the DoA into the state of awake, light to moderate sedation, superficial anesthesia, adequate anesthesia and deep anesthesia. [Baars et al., 2013; Beydon et al., 2009; Jahanseir et al., 2018; Kurita et al., 2005, 2012; MacKenzie et al., 2018; Schmidt et al., 2002; Silva and Antunes, 2012] The BIS is designed for human patients and thus bases calculations on human data, although it is still tried on veterinary patients like pigs in laboratory settings. Additionally, its precise algorithm is a disclosed commercial method. [Beydon et al., 2009; Silva and Antunes, 2012] Although the details are not known, as the product is commercial, it is known that BIS calculates Fast Fourier Transforms and uses this to do bispectral analysis, calculate a relative beta ratio and combine it with information of burst suppression. Although it does still show reliability in pigs, especially for hypnotic-opiate combinations, BIS still suffers from some problems related to different anesthetic agents and physiological states, as also explained in section 2.2.4 [Beydon et al., 2009; Kurita et al., 2012]. Besides BIS, permutation entropy has also shown promising results, which calculates the degree of complexity of the cortical information [Gu et al., 2019; Hashimoto et al., 2020; Jahanseir et al., 2018; Silva and Antunes, 2012]

2.2.4 Why controlling general anesthesia is difficult

Research into general anesthesia experiences the problem of a combinatorial explosion when factoring in different anesthetic agents, different methods of assessment, different species and different situations, e.g. hemorrhage.

For example, Montupil et al. [2019] notes how processed EEG monitoring reduces the incidence of intraoperative awareness and postoperative recall, particularly when intravenous anesthetic agents are used, but not as effective as simply monitoring the anesthetic gas when anesthesia is administered using inhaled vapors. Kurita et al. [2012] explains how hemorrhagic shock decreases the effect of intravenous anesthesia, but mostly after a specific degree of hypotension, making EEG monitoring indexes like bispectral index (BIS) give the impression of a waking subject, even if the BIS should be lower in such a case of hemodynamic change. Baars et al. [2013] notes how propofol and similar volatile anesthetics increases low frequency activity, while ketamine increases high frequency activity, which then might give the impression of waking with BIS. Though, this was only true when ketamine was given with other anesthetics as part of balanced anesthesia, which makes the tools even more unreliable. This is supported by Liu et al. [2018] summarizing the controversy around BIS and incorrect reflection when using different anesthetic agents like ketamine and nitrous oxide. Beydon et al. [2009] shows how the pharmacodynamics of propofol differs between pigs and humans. A high amount of interspecies variation of results is observed, which makes it difficult to translate information from one species into another, as many different factors besides species also has to be controlled for [Haga et al., 2011]. Last, García et al. [2021] and Morimoto [2008] discusses the paradoxical response of visceral pain pathways that might activate during body cavity noxious stimulation. In several studies, this paradoxical response may increase delta waves and suggest the dose of a hypnotic anesthetic agent is too high if assessing the DoA using an EEG, even if the DoA is acceptable.

Overall, many different factors compound together to provide contradicting results when a single variable like species, anesthetic agent, monitoring method or alike, is changed.

2.3 Methods of decision support for general anesthesia

This section will review the use of decision support for monitoring anesthesia and exemplify some machine learning algorithms that might be useful to assess the DoA.

2.3.1 Application of decision support

The application of decision support for general anesthesia has primarily been EEG-guided monitoring. [Baars et al., 2013; Haga and Ranheim, 2005; Martoft et al., 2002; Montupil et al., 2019; Silva and Antunes, 2012]. Currently, all methods seem to investigate the current DoA instead of predicting needed changes to it, although indexes from e.g. BSI are numerical which allows observation of the index changing before it changes category. For animal models, the assessment of somatosensory reflexes is not perfect but is generally considered acceptable and is the most common method to use. EEG-guided anesthesia also focuses on the current DoA [Silva and Antunes, 2012] Though, not all investigated parameters in relation to anesthesia drop simultaneously. In regards to non-EEG parameters, an introduced anesthetic problem tends to first show in the end-tidal CO^2, while other parameters like from the cardiovascular system usually follows after. [Gayer et al., 2016] Methods like BIS also show inconsistent time delays between 15 to 66 seconds relative to changes. [Pilge et al., 2006] Thus, when the DoA changes based on physiological parameters or EEG, changes to the DoA might have been warranted for some time without knowledge to the experimenter or clinician. Considering the correlation between somatosensory reflexes and EEG, this could be the reason for unexpected and sudden somatosensory reflexes.

When working with anesthesia, it is important to not overshoot the needed anesthetic agent and need to change it again, thereby causing an oscillation between different DoA, since a stable DoA is considered better for the well-being of a patient. [Pauldine et al., 2008] Thus, although the assessment of DoA using either somatosensory reflexes or EEG is considered acceptable, they are both mostly good at considering the present DoA, which does not necessarily help prevent an oscillation between different DoA.

Thus, because the assessment of the current DoA is considered acceptable and because different parameters drop at different times or have a time delay, it is of interest to look into methods that try to predict the DoA e.g. before it's considered too light using somatosensory reflexes or using conventional processed EEG. This could be used to help make finer adjustments in preparation of a changing DoA before it becomes critical to correct, since smaller adjustments over time are preferred to multiple larger adjustment, and since large adjustments might overshoot the desired DoA.

2.3.2 Benefits of EEG-guided anesthesia

The literature obtained in the structured literature search investigating the effect of EEG-guided anesthesia in pigs is limited, and most of the literature concerns humans. Literature dealt mostly with the effect of preventing postoperative delirium and intraoperative awareness, and decreasing length of stay in intensive care units. No reduction in all-cause mortality has been found when using EEG-guided anesthesia based on data from three studies. [MacKenzie et al., 2018; Sun et al., 2020]

Postoperative delirium

Postoperative delirium is a common side-effect of too deep DoA in humans, appearing after 15-20% of major operations in patients below 65 years old. It causes a discomfort of the patient, extended hospitalization and a need of more resources for care of the patient. It is likely caused by an extended time of burst suppression, although no direct relationship has been established. [MacKenzie et al., 2018; Montupil et al., 2019]. It has been investigated whether EEG-guided anesthesia reduces the incidence of postoperative delirium, although the findings are controversial. The meta-analysis of MacKenzie et al. [2018] compiling results of five studies finds that EEG-guided anesthesia reduces the incidence by 38%, while the meta-analysis of Sun et al. [2020] compiling the results of five studies claims the evidence is insufficient. Sun et al. [2020] relies on one large randomized controlled trial of 1232 older adults, where no difference was found. These controversial results are backed up by the review of EEG-guided anesthesia by Montupil et al. [2019] suggesting the need of more large randomized controlled trials. Overall, postoperative delirium shows why the solution to good control of anesthesia is not to simply keep a constant period of burst suppression, and is one of the reasons why just because a patient is unconscious, it does not necessarily mean the DoA is acceptable. [Montupil et al., 2019]. Though, the European Society of Anesthesiology still recommends using EEG-based monitors to avoid postoperative delirium. [Sun et al., 2020]

Intraoperative awareness

Literature of intraoperative awareness and EEG-guided anesthesia also used to be controversial, as the results of many studies were often at odds. Though, the current consensus is that EEG-guided anesthesia does reduce the incidence of intraoperative awareness, although not superior compared to using end-tidal anesthetic gas monitoring instead of EEG, when the anesthetic agent is an inhaled vapor. [Montupil et al., 2019]

Length of stay

In regards to length of stay in the intensive care unit, Sun et al. [2020] found a statistically significant reduction of 0.29 days based on three studies, although no statistically significant reduction of length of hospital stay was found when compiling four studies. However, the review of Punjasawadwong et al. [2014] agrees that especially length of stay at intensive care unit is reduced, but also finds that length of hospitalization decreases.

2.3.3 Use of machine learning for monitoring of anesthesia

Machine learning is a set of methods that can be used for decision support when monitoring anesthesia. Machine learning work by using algorithms that iteratively use data to train themselves for use in pattern recognition tasks, e.g. by classifying an unlabeled data set into different DoA, or by assessing DoA of a single set of parameters. Machine learning methods can process high dimensional data in various ways even if it may be noisy, and has the ability of the algorithms to train themselves which minimizes the need for human interaction that may be biased. Becoming proficient in anesthesiology requires lengthy education, which is why it would be practical with computers assessing the DoA if possible [Morgan et al., 2006]. Some algorithms may also be able to better classify data based

on exceptions, e.g. prioritize features depending on what other features are, which can provide a synergistic effect that improves accuracy. Some algorithms require features to be prepared for them, while other can engineer their own features. The main drawback of machine learning is the requirement of a large quantity of data, although with a high variance of need between algorithms. [Goyal et al., 2018; Hashimoto et al., 2020; Tacke et al., 2020]

Some of the methods are classified as unsupervised and can be used to label data, while supervised methods can use those labels to classify future samples after having been trained.

Unsupervised learning and labeling

Unsupervised learning can be useful, when a developer does not have a clear understanding of the data. This does not mean, that the type of data is not understood, but rather that most data tend to be complex with various correlations and how this can be difficult to manually interpret, when many parameters are present at the same time. Unsupervised learning can cluster data into different groups, e.g. a defined DoA, and thus provide a label to each sample. This can help prevent human bias, e.g. if an expert inspected an EEG and were to label the data manually. It is also useful as a substitute to an expert, if a large quantity of data exist, as a machine can label the data at a much faster pace. Clustering is also being used in research to try and identify patterns, which were not already known, since a machine might see data differently compared to a human expert. [Jayatilake and Ganegoda, 2021] From the literature obtained from the structured literature search, no data from studies of either humans or animals applied clustering to label data. Few studies explained their method of labeling to an acceptable degree for replication. Some labeled their data based on cutoff points, e.g. from a histogram of BIS values [Jahanseir et al., 2018] or labeling was done based on somatosensory reflexes, which then might only include two labels (conscious and unconsciousness) [Tacke et al., 2020]. Thus, it is of interest to investigate the possibilities of better labeling using machine learning.

Examples of popular clustering methods are k-means, mean-shift and Density-Based Spatial Clustering of Applications with Noise. [Seif, 2018]

The k-means algorithm clusters based on randomly positioned centers where each sample belongs to the closest cluster based on the mean of the cluster. The clusters then moves towards the centroid of all its samples in its cluster. This causes it to potentially obtain new samples which it got from getting closer to a sample than another cluster, or lose samples to another cluster similarly. In the end, all samples have been given the most appropriate label in terms of euclidean distance to the centroid of each cluster. k-means requires a predefined number of clusters for initialization, which is both an advantage and disadvantage based on the use case. [Joshi, 2020]

Another similar method is mean-shift clustering. This method uses many randomly positioned cluster windows which moves towards the means of all its samples to incorporate a higher number of samples, thus placing themselves in a high density area when the density no longer increases. In the end, samples belong to the closest center of a cluster window after overlapping windows have been removed. This method uses a radius of the cluster

window to help convergence and assigns samples after convergence. This is opposite to k-means, where labels to a sample during convergence and can change as the center of clusters change. Also unlike k-means, this method does not require a predefined number of clusters, although the cluster radius does have to be manually chosen. [Seif, 2018]

Though, it is important to notice that especially k-means tend to cluster samples in a uniform manner like a circle in two dimensional space, which would not be appropriate to use if the data isn't in uniform clusters but instead are in non-uniform patterns as shown in figure 2.1. Although mean-shift works differently, it can still risk the same problem in a different manner. For example, if the cluster radius was too small, the rectangular cluster would receive multiple clusters, thereby segmenting each rectangular cluster into smaller clusters instead of one.

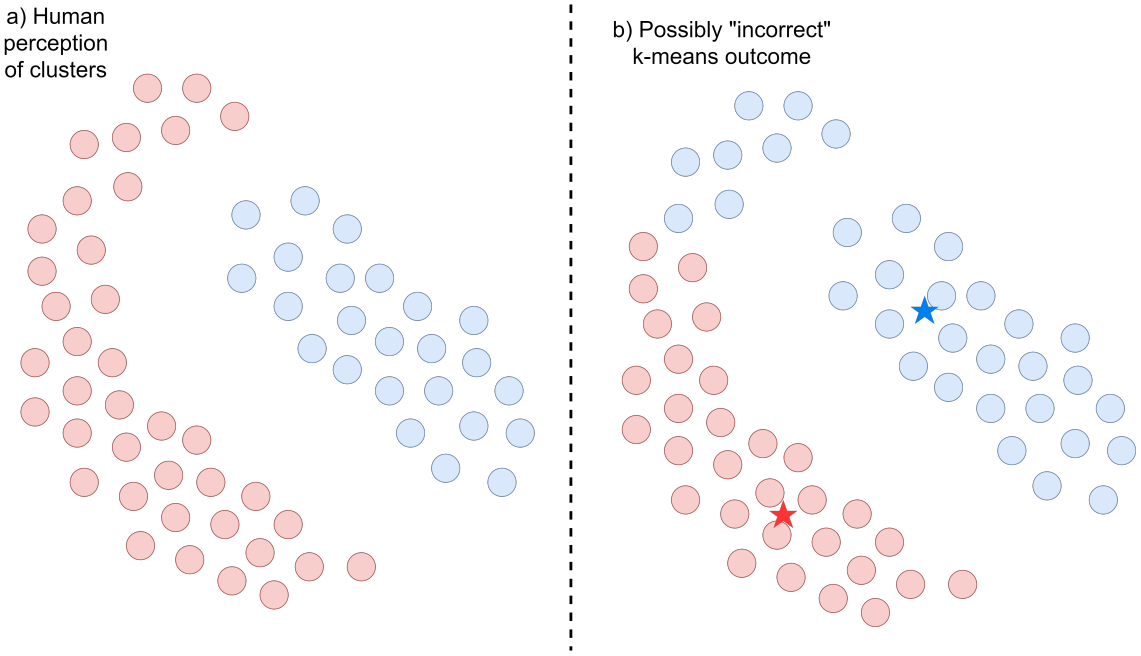

Figure 2.1. Illustrates a shortcoming of k-means in regards to non-uniform clusters. Illustration a) shows how a human might label the hypothetical samples, while illustration b) shows how k-means might cluster the data, where each star represents the centroid of the samples in its cluster. In illustration b), it can be seen how the samples originally labelled red in illustration a) are instead blue, because they are closer to the blue center than the red, despite seeming like a part of the originally red cluster.

A method like DBSCAN circumvents the problem of clustering based on centroids. DBSCAN works by selecting an unvisited sample and incorporating all points within a predefined distance of itself. If there are more samples than a predefined number, it incorporates all samples within the previously incorporated samples using the same predefined distance. This repeats until no more points are being added, at which a new unvisited point is being selected. If there are fewer samples than the predefined number, the sample is labeled as noise. Thereby, DBSCAN does not require a predefined number of clusters and can also detect potential outliers, which could confuse a supervised learning algorithm. The disadvantage of DBSCAN lies in the predefined distance and number of minimum nearby samples, which makes it difficult if clusters contain varying densities. [Seif, 2018]

Clustering methods can use any type of data composed of samples with singular values, which does make it necessary to calculate extra parameters out of data like EEG, since a whole EEG can not be processed but a set of parameters from an EEG can be used. [Jayatilake and Ganegoda, 2021; Joshi, 2020]

Other methods are more complex like hierarchical clustering, which creates many clusters and puts them into hierarchies, such that one cluster might have several clusters inside of it. [Jayatilake and Ganegoda, 2021] For anesthesia, no such hierarchies or higher complexity of the labels tend to be defined, because the different DoA are generally defined as four to five categories ordered numerically [Jahanseir et al., 2018; Silva and Antunes, 2012], which makes a method that applies a standalone label an appropriate method for labeling different DoA.

A summary of the clustering methods can be seen in table 2.4.

Algorithm	Characteristics
k-means	Clusters based on centroids. Requires a predefined number of clusters. Little human interaction, besides number of clusters.
Mean-shift	Clusters based on centroids. Does not require a predefined number of clusters. Potential need of moderate human interaction to iteratively select a cluster radius.
DBSCAN	Clusters based on points within distance. Does not require a predefined number of clusters. Potential need of a lot of human interaction, as both the distance between samples in a cluster has to be chosen along with the minimum number of points in a cluster.

Table 2.4. Summarizes the three exemplified methods for clustering.

Supervised learning and classification

For monitoring of anesthesia using machine learning, many different methods exist. Examples of popular algorithms for categorization problems include k-nearest neighbour (kNN), support vector machine (SVM), decision trees, artificial neural network (ANN). [Jayatilake and Ganegoda, 2021]

kNN compares a sample to be classified to an arbitrary number of data points from the k-means method are closest. An arbitrary number of neighbour can be defined and the cluster with most neighbours is the label, of which the sample is classified as. kNN is popular, simple and usually has a high accuracy. Though, the method can quickly become computationally expensive and require a lot of memory, both in regards to training and testing. [Jayatilake and Ganegoda, 2021] Although euclidean distance between the sample and neighbours is mostly used, but different types of distance measuring can be used, e.g. to increase or decrease the importance of outliers. [Joshi, 2020]

SVM tries to fit a decision boundary in a space that maximizes the distance between different DoA. SVM is also widely implemented in other neuroimagining analyses.

Although SVM is one of the most simple algorithms to use for anesthesia, it is also relative sensitive to choice of parameters. [Campbell et al., 2020] The study of Jahanseir et al. [2018] used it to assess DoA using measurements of the relative power of frequnecy bands to each other along with values for entropy (Approximate- permutation and Shannon entropy, fractal dimension and detrended fluctuation analysis), and achieved an accuracy of 80% when assessing four DoA equal to BIS with the exception that "Awake" and "Light to moderate sedation" were combined into one category. The method to obtain the raw EEG for the data included one channel instead of two as with BIS. Campbell et al. [2020] also showed that although SVM had a slightly higher area under the curve compared to an ANN and decision trees, it was less stable in its decisions, as some DoA had a significantly higher accuracy than others, while decision trees and ANN showed a more average accuracy across different DoA. Tacke et al. [2020] tested SVM against several classifiers like Bayesien classifiers, simple ANNs, NaiveBayes and logistic regression, where SVM performed best.

Decision trees resembles many decision nodes, where each "branch" is a simple decision rule used to classify the information. A continuous use of many branches would then lead to a final decision, much like a flowchart. The most common subtype of decision trees is Random Forest, which has shown a lot of success in multivariate neuroimaging analyses and EEG-based brain-computer interfaces. Random Forest starts by creating a single tree by creating decision nodes based on different parameters, which are decided randomly. Then, a new tree is created with a different subset of parameters. Finally, all trees are ensembled and use majority voting to decide on the classification. Random Forest is one of the state-of-the-art machine learning methods for EEG-based brain-computer interfaces, if little data is available. [Campbell et al., 2020; Lotte et al., 2018] Liu et al. [2018] utilized Random Forest using BIS as the reference value and achieved a high area under the curve (AUC) of 98%.

ANN is a category of algorithm that stand out by using many different decision boundaries that are not binary but rather like fuzzy logic. These decision boundaries are indiviual perceptrons, or neurons, which together simulate neurons of the brain. [Goyal et al., 2018; Gu et al., 2019; Hashimoto et al., 2020]. As previously mentioned, Campbell et al. [2020] and [Tacke et al., 2020] did use ANN which were outperformed by SVM. Though, investigation into the type of ANN they used, hereby the architecture and supportive training algorithms, revealed they were simple compared to recent advances in ANNs.

ANNs can be categorized as shallow or deep, where shallow ANNs tend to have few layers, e.g. input-, two intermediate- and one output layer, where as deep ANN can have any numbers, where a number of layers of 100 or 1.000 is not uncommon for some use cases. Some newer architectures can be useful to analyze an EEG, e.g. the long short-term memory architecture, since these type of architectures incorporate a memory element that can be useful for long time series signals to predict future values, where information from the past would contribute a lot. This allow for non-linear pattern recognition unlike that of human decision and can be useful in finding undiscovered patterns, though with the disadvantage that the process can be too complex for human interpretability, thereby causing a black-box situation. The same black-box problem exists when using shallow ANNs. [Goyal et al., 2018]

Obtained literature of in regards to classification of DoA has primarily been either shallow

ANNs, or the deep ANNs have been designed to only use preprocessed data, e.g. not raw EEG but values for frequency bands and entropy. [Campbell et al., 2020; Gu et al., 2019; Hayase et al., 2019; Jahanseir et al., 2018] This might explain the lacking ability of ANNs in the literature of monitoring DoA

Little focus is on the use of ANN architectures like convolutional neural network (CNN) or long short-term memory (LSTM) layers for pattern recognition, which could provide new useful information unlike that in the current literature [Hashimoto et al., 2020].

Anesthesiology is a field with much to benefit from advances in artificial intelligence, because it can assist with multiple purposes like pattern recognition of EEG and decision making of DoA. Recent resurgence of artificial intelligence is attributed to the availability of large data sets, new stronger hardware and a new wave of development of ANN architectures and other machine learning algorithms. [Hashimoto et al., 2020] [Goyal et al., 2018]

The disadvantage of ANNs is the need of large amounts of data, unlike that of kNN, SVM and Random Forest to avoid overfitting to little data and cause poor generalization. [Goyal et al., 2018]

A summary of the classification methods can be seen in table 2.5.

Algorithm	Characteristics
kNN	Categorizes based on distance to samples of nearby clusters. Computationally expensive if a lot of data is present. Allows adjustment of method for distance measuring which provides flexibility. Requires relatively little data. Little litereature regarding its use to monitor anesthesia.
SVM	Categorizes based on a decision boundary . Promising results in regards to EEG and anesthesia. Requires relatively little data.
Random Forest	Categorizes based on an ensemble of decision trees functioning as decision nodes like a flowchart. Promising results in regards to EEG and anesthesia.Useful with little data. Requires relatively little data.
Neural network	Categorizes based on a network of perceptrons. Moderately promising results in regards to anesthesia. Has recent advances not present in current literature. Can process raw time-series signals. Requires a lot of data. Black-box element, hindering interpretability of reasoning behind results.

Table 2.5. Summarizes the four exemplified supervised learning algorithms for classification.

2.4 Aim of project

ANNs have shown a lot of development in recent time. Although the literature has shown promising results using especially SVM and Random Forest, it is of note how little interest exist in the use of ANNs for pattern recognition for monitoring of anesthesia. ANNs have the ability to not only find unique patterns but also classify based on conventional parameters like physiological data or frequency bands and entropy values of. For these reason, ANNs was chosen over kNN, SVM and Random Forest.

Considering how even the most popular EEG monitoring tool, BIS, show time delays in its interpretation during assessment of DoA, it is of interest to try and predict future DoA as to give warning to the person responsible for maintaining an appropriate DoA. Based on the time of the time delays of BIS, a warning of a minute is appropriate. For this project, the available data consists of parameters commonly available during monitoring of anesthesia, along with EEG and ECG. This leads to the current aim of the study:

How can a neural network be used to predict the future depth of anesthesia by a minute or more using features extracted from an EEG in combination with time series analysis by neural network layers in a mixed-data approach?

Methods 3

This chapter will provide an overview of the solution strategy, what kind of data was used, how the data was preprocessed and how the neural network predictions were evaluated. Finally, the method of labeling with clustering will be elaborated.

3.1 Systematic literature search

To obtain knowledge for the problem analysis regarding animal models, assessing the DoA and use of neural networks in anesthesiology, both a structured literature search was completed alongside unstructured literature searches. Details of the methodology regarding the search can be seen in appendix A.

3.2 Solution strategy

Figure 3.1 illustrates an overview of the solution strategy. An EEG was preprocessed into values for the frequency bands and entropy values. These were then used to label the data with k-means. A combined neural network then utilized two neural network architectures: A simple neural network architecture processed the single-value features like frequency bands and entropy values, while a complex neural network processed the time series EEG itself. The output of both the simple and complex neural network branches were concatenated and processed by the combined neural network, which utilized the labels to make decisions during training to improve the accuracy. Finally, the combined neural network could then make a prediction as its output.

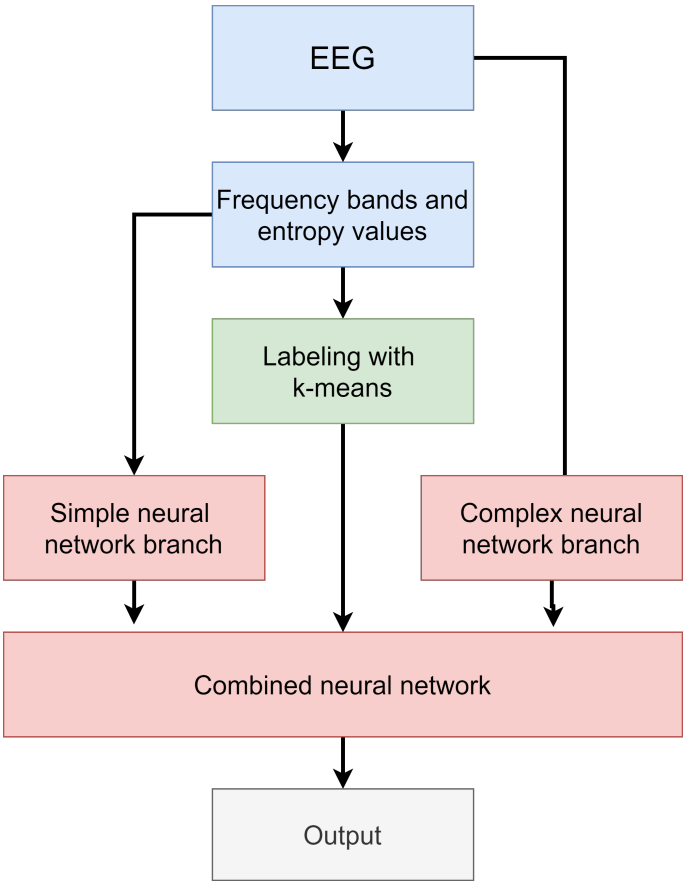

Figure 3.1. Illustrates the solution strategy. Blue boxes indicates feature data, green boxes indicate labels and red boxes indicate parts of the combined neural network

The process of labeling the data was done with clustering. A conventional neural network requires labels to detect the required changes to its neurons to train itself. For clustering data related to anesthesia, the choice of labeling method would depend on the expectation of the data. k-means is an ideal choice, if the number of clusters has to be predefined. If not, mean-shift is ideal. Though, methods do exist to evaluate different number of clusters of k-means to find the most appropriate number of predefined clusters [Novia, N/A]. If the clusters are not expected to be in uniform clusters, DBSCAN is an ideal choice. For this project, the fiddling of the distance value of DBSCAN was considered likely to introduce too much human bias. Additionally, the data is not expected to be in complex shapes that warrant the need of DBSCAN. Methods exist to evaluate the optimal number of clusters for k-means, which allow some compromise between the ability of machine learning to find its own number of clusters that might be more appropriate, and the expectation of four to five categories of DoA as is common in the literature and commercial monitors. Additionally, k-means require little human interaction, which mitigates human bias. Thus, k-means was chosen for labeling the data.

Neural networks have different architectures, which work well for different contexts. For simple data like single-value features, a simple architecture of fully connected layers called Dense layers is commonly used. Though, the use of such a simple architecture mostly does feature interpretation. An advantage of neural networks over other machine learning methods is also the capability of feature extraction alongside interpretation. [Aggarwal

et al., 2018] To utilize this capability present in some neural network architectures, a more complex neural network was also created, which tried the use of different architectures like LSTM and CNN. A review by Nagabushanam et al. [2019] highlights different architectures, where LSTM was shown to be best overall for signal classification of EEGs. Though, this does not guarantee that LSTM would prove best for classifying the DoA, which prompts the interest to still investigate other feature extracting architectures like CNN.

Last, the two neural networks were combined. This was done, because some neural networks can improve their performance by correlating information between domains. A neural network might not be able to sufficiently analyze an EEG when its given only the EEG, but it might be able to analyze it successfully when the neural network has other information to use as an information stepping stone. This is because neural networks might not be able to find existing patterns because of its method of training, where it takes steps towards the optimal solution using gradient descent. Ignoring how evolution is essentially random, this can be compared to how evolution works, where a beneficial change can not be obtained by many small changes, if the evolutionary path between the current form and the beneficial change contains disadvantageous changes. Thus, a neural network might not move towards the optimal solution, because its loss function can either fall into a local minima or it can be directed towards a less optimal but more obvious solution. Therefore, the combination of different domains or formats of information might enable a neural network to better analyze the data. [Aggarwal et al., 2018; Rosebrock, 2019]

3.3 Data Foundation

This project used electrocorticographical data from six pigs from an animal trial investigating the effect of long-term potentiation (LTP) of nerves. The animal trial is a study under protocol number 2016-15-0201-00884 and was approved by the Danish Veterinary and Food Administration under the Ministry of Environment and Food of Denmark.

The experiment consisted of sets of two EEG baseline recordings with stimulation between them, a series of 12 minute waits after each set and an intervention of LTP, which can be seen in figure 3.2. All pigs underwent three sessions followed by the intervention and then approximately ten additional sets after the intervention.

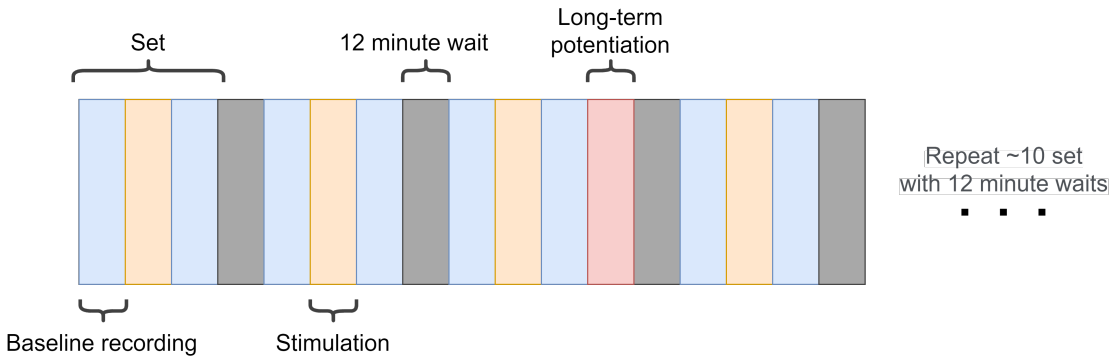

Figure 3.2. Illustrates the order of the baseline recordings, stimulation and intervention order of the animal study.

Each baseline recording lasted 30 seconds. The stimulation consisted of stimulation to two nerves. Of these two, one was 50 non-noxious electric stimulations to the motor branch with an amplitude of 1 milliampere, a duration of 500 microseconds and with 2 seconds between each stimulation. The other was 50 noxious electric stimulations to the cutaneous branch with an amplitude of 5 milliampere, a duration of 1000 microseconds and with 2 seconds between each stimulation.

The intervention of LTP was to both branches and consisted of four sweeps of electric stimulation with an amplitude of 15 milliampere, a duration of 1000 microseconds, with 10 seconds between each stimulation and four sweeps with a frequency of 100 Hz.

In total, data from six pigs were used, where four of them underwent the intervention of LTP and the last two were control pigs. This provided 157 baseline recordings of 30 seconds each.

The pigs were sedated using sevoflurane, fentanyl and propofol. After surgery, sevoflurane was decreased while fentanyl and propofol were doubled. Up to around 2% concentration of sevoflurane was used in the beginning to sedate a pig for surgery and the concentration was reduced multiple times until eventually becoming zero. Fentanyl was used in the quantities of around 300-350 µg/ml before surgery and up to around 850 µg/ml at the end of the experiment session. Propofol was used in the quantities of around 60 mg/ml before surgery and up to around 170 mg/ml at the end of the experiment session. Increase and decrease of all anesthetic agents were adjusted to the need of sedation at the moment, which was assessed using somatosensory reflexes and physiological parameters like heart rate, blood pressure, SPO^2, temperature and respiration rate. Variation in the quantities of the anesthetic agents was therefore due to individual biological variability of the pig or sudden need of additional general anesthesia.

The EEG data was sampled at 24414.062 Hz and prefiltered at 0.3 Hz to 300 Hz. 16 channels were recorded with close proximity to each other in the primary somatosensory cortex covering an area of approximately 1x1 centimeters.

3.4 Preprocessing

An example of the original EEG segments can be seen in figures 3.3 and 3.4. Each EEG was truncated, downsampled and an average EEG was calculated instead of 16 channels. Since each EEG is required to have the same dimensions, all were truncated to 30 seconds, making the length of a non-downsampled EEG 24414.062 $Hz \cdot 30\ s = 732421.86\ data\ points$, rounded to nearest integer. The signal was downsampled because its sample rate was already well over the sample rate required to capture information in the common frequencies of EEG according to the Nyquist frequency, where the frequency band with the highest frequency can be considered gamma at 30 Hz and above [Mazzoni et al., 2010]. The signal was also downsampled, because its length was well over the Nyquist frequency, meaning it was not expected to provide additional information usable for a neural network if the signal was at its full length, but it would instead incur additional unnecessary training time. Therefore, each EEG segment was downsampled to around 2.000 Hz, which is achieved by downsampling the full length 12 times, thus getting a downsampled length of 61035 data points which represents 30 seconds. Finally, the 16

channels were averaged. Because the distance between each channel is too small to yield meaningful different information and the average channel is expected to be more reliable and stable than any individual channel.

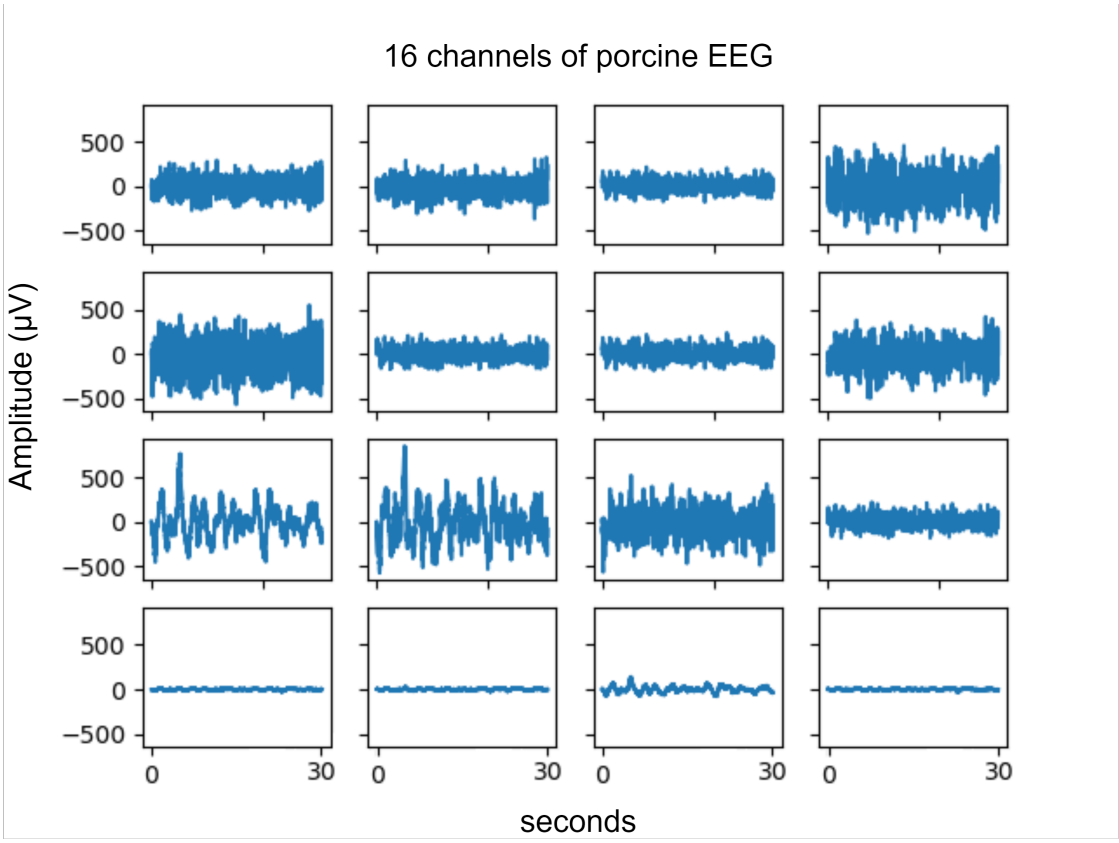

Figure 3.3. Illustrates all 16 channels of the first EEG of a pig.

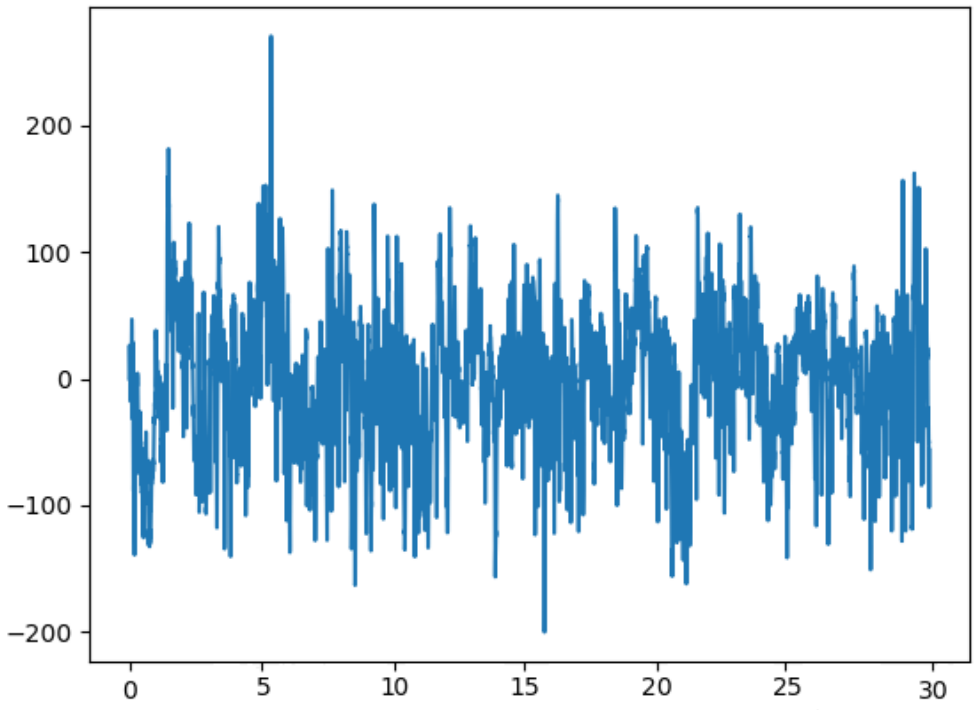

Figure 3.4. Illustrates the average channel of figure 3.3.

3.5 Method of Evaluation

Each neural network model will be evaluated using an independent part of the data separated into the test set as further explained in section 4.1.2. The evaluation will consist of a confusion matrix and performance metrics like accuracy when assessing multiclass labels. [Dankers et al., 2019]

Additionally, detecting the transition from one label to another was also evaluated instead of detecting individual labels. This was done get a better impression on the possibility of actually detecting changes instead of trying to predict the individual labels. In this case, the methodology requires binary class evaluation, i.e. change in the labels or no change in the labels, which also uses a confusion matrix. Though, it is additionally supported by the performance metrics of sensitivity and positive predictive value, as shown in figure 3.5. Sensitivity and specificity were chosen, because the focus was on detecting changes, where as specificity and negative predictive value explains the performance of the predictions not involving change.

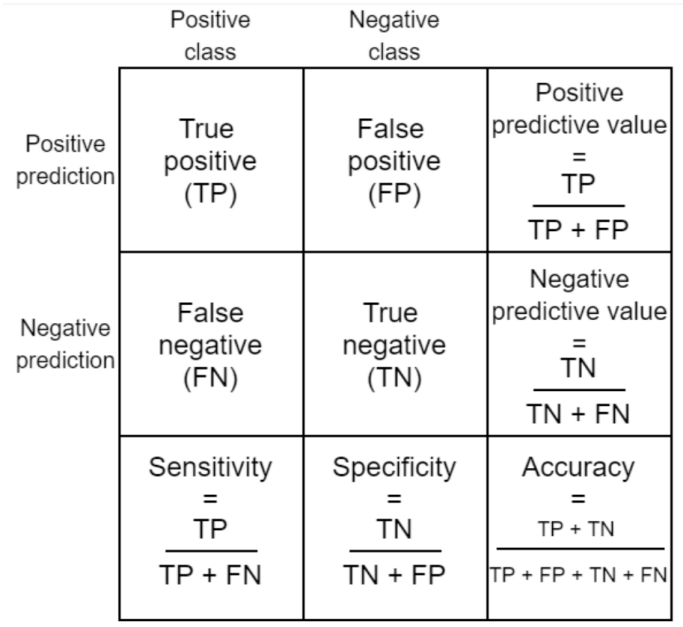

Figure 3.5. Illustrates a confusion matrix with the four squares in the top left. The remaining squares hold the performance metrics and the calculations hereof. [Dankers et al., 2019]

The evaluation was also negatively influenced by the low sample size. Because of the low sample size, there was a relatively high likelihood of the evaluation being significantly affected by chance either due to how the data was being split or due to randomly initialized values of either the k-means algorithm or the neural network. For this reason, a Monte-Carlo cross-validation (MCCV) was employed. This type of cross-validation reduces the effect of variance by averaging the result of randomly sampled training processes. A commonly used method is k-fold, which iteratively samples k number of folds, e.g. 10, and then samples the data set by the first 10%, then the next 10% and so on until ten folds of 10% each has been trained, at which the average performance is then calculated. [Dubitzky et al., 2007] Although MCCV does not necessarily remove bias better than k-fold, it does better account for the variance. This is also seen in studies trying to forecast time signal series, where MCCV provides the same average performance as k-fold, but its standard deviation is lower than that of k-fold [Fonseca-Delgado and Gómez-Gil, 2013], thus making it more reliable to evaluate results in data sets with a low amount of samples. Though, this also makes it important to mention that the clustering algorithm will also be affected similarly. Thus, small deviations in the number of samples for each label will show over some different runs.

For the simple neural network, the number of models created to find the average performance with MCCV was ten. For the complex neural network and the combined neural network, three models were made to find the average performance. This difference is related to time, as the training process of the simple neural network was significantly faster than that of the complex neural network.

3.6 Labeling using k-means

To utilize the data in a neural network, the data is required to contain labels representing the DoA. Optimally, the labels should be derived from a golden standard, which is an expert's assessment with access to relevant information such as EEG, physiological parameters and testing of reflexes. Though, the specific DoA was not noted alongside the rest of information. Although an expert could spend a substantial amount of time to manually inspect all EEGs to determine the DoA, the expert would still be disadvantaged because testing of reflexes is not possible, and because an evaluation of the DoA relies on a subjective impression in the moment, which can not easily be achieved afterwards through the data only [Flecknell, 2015].

Feature selection

First, new features had to be extracted from the EEGs, since the k-means++ algorithm can not process the entirety of an EEG or meaningfully extract relevant information from it. Instead, nine features consisting of frequency bands and entropy values were calculated from each EEG.

The frequency bands were singular values representing the total power in the range of each frequency band seen in table 3.1. These ranges are derived as suggested by Silva and Antunes [2012]. They were chosen, because the frequencies change depending on the DoA. The most prominent of these is how the beta activity decreases while delta- and alpha activity increases, as the DoA becomes deeper. though to a specific point until burst suppression appears [García et al., 2021; Jahanseir et al., 2018]. Though, some studies still use other frequency bands, like the theta- and gamma bands, with some performance gain [Hashimoto et al., 2020]. Neural networks are also designed to prioritize the input features, such that the features with little significance as evidenced during trained, would receive a lower weight in its respective neurons. Thus, the theta- and gamma bands were included anyways [Aggarwal et al., 2018].

The first six features were the total energy of each of the frequency bands. The range of each frequency band can be seen in table 3.1. The frequency bands were derived from each EEG using Fast Fourier Transform.

Delta	1 - 4 Hz
Theta	4 - 9 Hz
Alpha	9 - 15 Hz
Beta	15 - 30 Hz
Low gamma	30 - 50 Hz
High gamma	50 - 100 Hz

Table 3.1. Illustrates the ranges of categories of frequency bands as explained in Silva and Antunes [2012]

Spectral information like frequency bands are used in various formats like total power and relative power in different combinations. [Silva and Antunes, 2012] The use of large amounts of neurons can be used to calculate ratios, additions and differences of input values. [Aggarwal et al., 2018] This also removes the necessity of having to create many

different features that are essentially derivatives of the six frequency band and would create features with overlapping information.

Spectral information like frequency bands are frequency based, which does not properly take into account the non-linearity of the brain. Thus, entropy values represent the predictability of the signal which is expected to provide additional useful information from a different domain than the time- or frequency-based. [Liang et al., 2015; Silva and Antunes, 2012] Three values representing entropy were also used. Entropy values are used in information theory to represent the complexity and predictability of signals [Jeon and Chehri, 2020]. Such features have also been shown to correlate with DoA [Campbell et al., 2020]. Four commonly used entropy values used in estimating DoA are Shannon-, approximate- and permutation entropy. Shannon entropy measures the predictability of the amplitude of the signal by calculating probability distributions of previously observed amplitudes. Approximate entropy calculates a result similar to Shannon entropy and tends to have better performance, but require EEG signals of a long length and are noise sensitive. Permutation entropy was developed to measure the complexity of the EEG during coma and anesthesia and generally performs better than other methods of calculating entropy [Gu et al., 2019; Li et al., 2008; Silva and Antunes, 2012]. A fourth but less used is sample entropy, which is an alternative version of approximate entropy and has been shown to be more noise-resistant and perform better [Liu et al., 2018; Richman and Moorman, 2000]

After testing with Shannon entropy, the time to calculate the value for each EEG proved extensive. Because approximate- and sample entropy are relatively similar to Shannon entropy, it was not added as a feature. Approximate- and permutation entropy were included because of their frequent and succesful use in the literature [Campbell et al., 2020; Gu et al., 2019; Silva and Antunes, 2012]. Sample entropy does not appear as frequently in the literature in relation to anestheia but has shown better performance than its alternative approximate entropy in other use cases using biological time signals [Liu et al., 2018; Richman and Moorman, 2000]

Multiscale entropy calculations can also be used to assess the complexity and predictability at different segments of the EEG, which provide more total and varied information. Calculating the entropy at many different segments can be time extensive. [Li et al., 2019; Liu et al., 2015] This was attempted for three different multiscale entropy values of the yellowbrick Python package, which proved too time extensive for signals of the length and sample rate in this project. It is important for features to be able to be calculated within a reasonable timespan, since an algorithm predicting changes in the DoA might need to run constantly. Thus, any feature causing the calculation of features to take more than minutes would be impractical.

Splitting the features into three sets

Calculating the features in one set will provide one set of values describing the entire EEG. Thus, little proper progression can be observed in the values, since only a single value of each feature would describe the EEG. This is problematic if the DoA might be deep in the start of the EEG but light in the end of it. Thus, all features were calculated in three sets: One set of the nine features for the first 10 seconds, another for the next 10 seconds and another for the last 10 seconds of the EEG.

Thus, the final features for clustering was 27 features, where the first nine describes the first 10 seconds, the next nine describes the next 10 second and the last nine describes the last 10 seconds of the EEG.

Min-max normalization

Last, all the features were min-max normalized based on the highest and lowest value of that feature. This was done because it is disadvantageous for a neural network to work with features of different ranges. This makes it difficult to interpret the importance, when one feature might go from 0-1 while another goes from 0-100, where it would seem the latter feature is more important, as it changes more. Though, it maybe only changes more in absolute terms. Additionally, having a lower value range also speeds up the training process while not changing the performance. [Montavon et al., 2012]

Clustering

A number of five clusters was chosen for the k-means++ algorithm. Ideally, the number of clusters would be three based on the commonly used states of shallow, moderate and deep DoA [Silva and Antunes, 2012]. Though, it could be useful with additional states of DoA to provide the person responsible for anesthetic observation a finer degree of assessment. To investigate what would be an appropriate number of states, the Elbow method was implemented using a distortion- and silhouette score for determination of optimal number of clusters for k-means. The distortion score calculates the mean sum of squared distances to centers of clusters, while the silhouette score calculates the mean ratio of intra-cluster distance and nearest-cluster distance. These methods were used because of [Mahendru, 2019; Novia, N/A] because of their popular use in determination of the number of clusters in k-means. The implementation was done using the yellowbrick Python package, while the kneed Python package was used to locate the "Elbow" of the graph as shown in figure 3.6 and 3.7 instead of locating it visually.

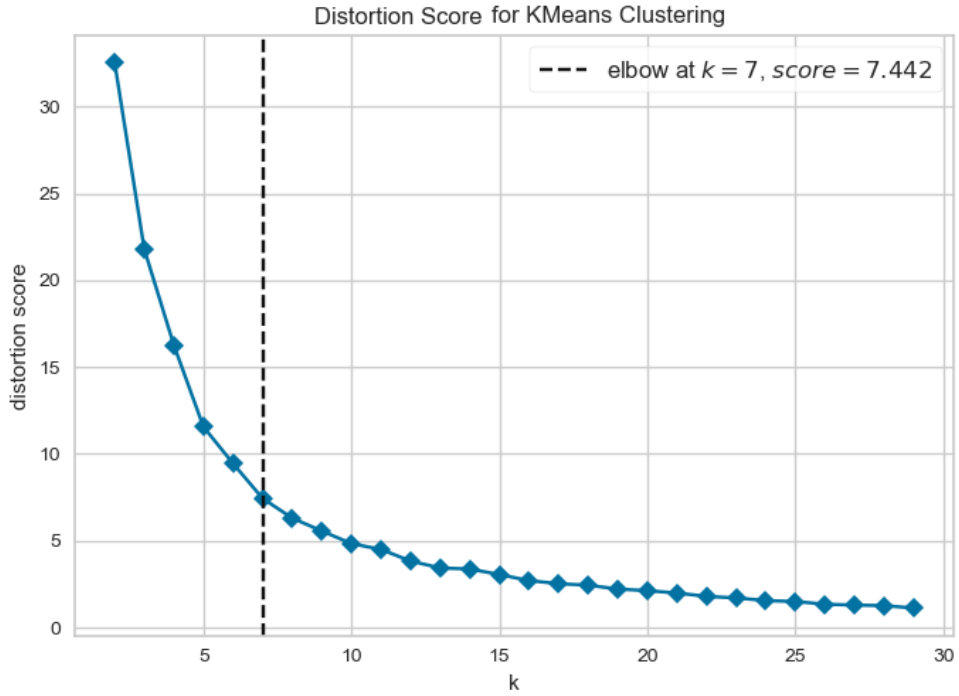

Figure 3.6. Illustrates an the optimal number of labels according to the elbow method using the distortion score.

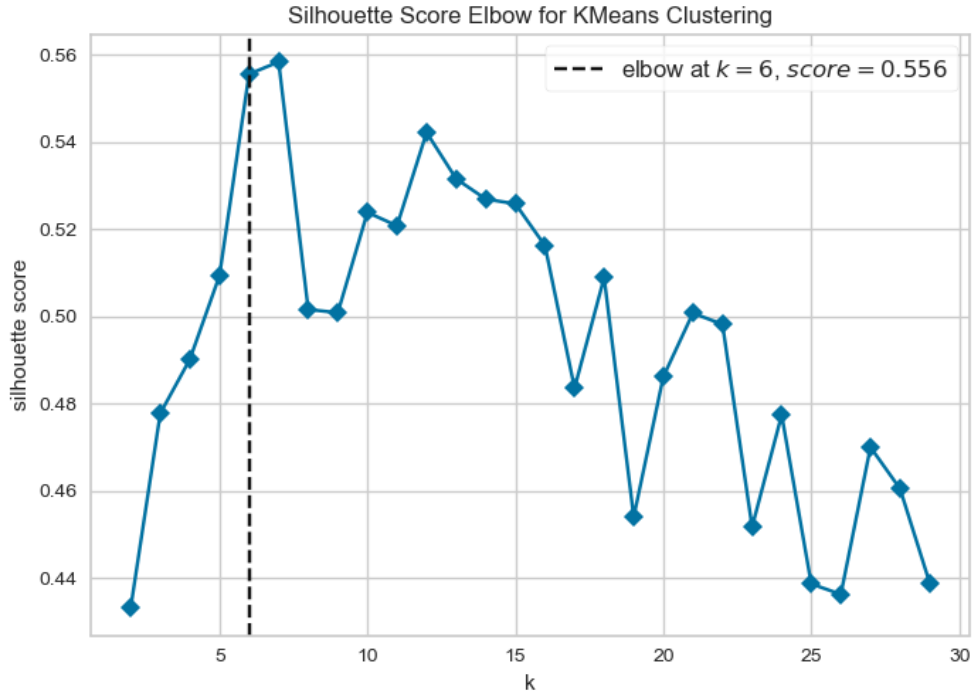

Figure 3.7. Illustrates an the optimal number of labels according to the elbow method using the silhouette score.

For final determination, the amount of samples of each cluster from a k-means++ algorithm using six and seven clusters were extracted.

For six clusters, this was: [27, 5, 25, 26, 17, 58]

For seven clusters, this was: [28, 5, 25, 26, 4, 14, 56]

Two problems existed with this amount of clusters and number of samples in each cluster. First, since the data was from an experiment where the aim was maintain the same DoA throughout the experiment, the entire range of DoA is not expected to exist in the data. Although it was unavoidable since the data has to be labelled, it is important to keep in mind that splitting the data into six or seven different DoA does not yield a range of DoA over the entire range of DoA. An illustration of this can be seen in figure 3.8. Thus, caution had to be taken since such fine division might be too fine and because it can not be guaranteed that the whole range of DoA is covered.

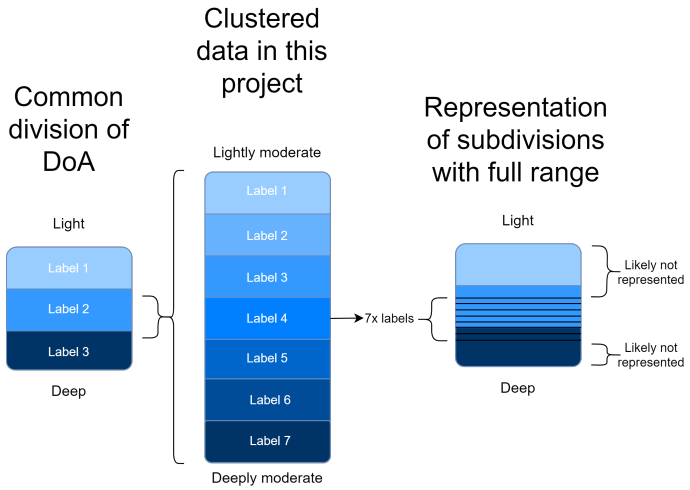

Figure 3.8. Illustrates a common division of DoA (light, moderate, deep) and how the clustering of k-means essentially clusters based only on the moderate category. Thus, the clustering is essentially divisions of moderate DoA into subdivisions of moderate DoA.

Second, the data had to be split into a training-, validation- and test set for use in a neural network. Though, it is important to keep each label proportionally represented in each relative to their size, which is why stratifying the labels was important [Sechidis et al., 2011]. The k-means++ algorithm can yield slightly different results and occasionally some results with the sample size will yield clusters with too few samples for proper stratification. After testing with four to seven clusters in k-means, it was apparent that the best compromise was using five clusters. This was the closest to the recommended number of clusters by the evaluation methods while ensuring a high likelihood of a proper clustering which was accepted by later stratification algorithm into three data sets.

For five clusters, the distribution of labels was: [27, 5, 51, 17, 58]

Development and Results 4

Because the solution strategy of this project consists of not only the final combined neural network but also the intermediary developmental steps, the development and results of both a simple neural network, a complex neural network and a combined neural network are explained in parallel in this chapter.

4.1 Simple neural network for preprocessed features

To develop the neural network for this task, Python was used with the Keras library v2.4.0 for its functional API, which provides flexibility in regards to input and output of individual layers. Being able to individually adjust each layer enables the possibility of combining two parallel architectures into the same output at the end of both architectures for a combined neural network model using mixed data. [Keras, 2020; Rosebrock, 2019]

4.1.1 Architecture of the simple neural network

The simple neural network was with an input layer, a series of hidden layers and an output layer. The hidden layers were designed using mostly Dense layers as defined by Keras terminology. Each Dense layer contained neurons which each function as shown on figure 4.1, where input values to a neuron become multiplied by an input weight that can be changed during model training. These weighted inputs are then added together along with a randomly determined value referred to as a bias. The combined value is then input to an activation function, which outputs the final value of the neuron. [Aggarwal et al., 2018]

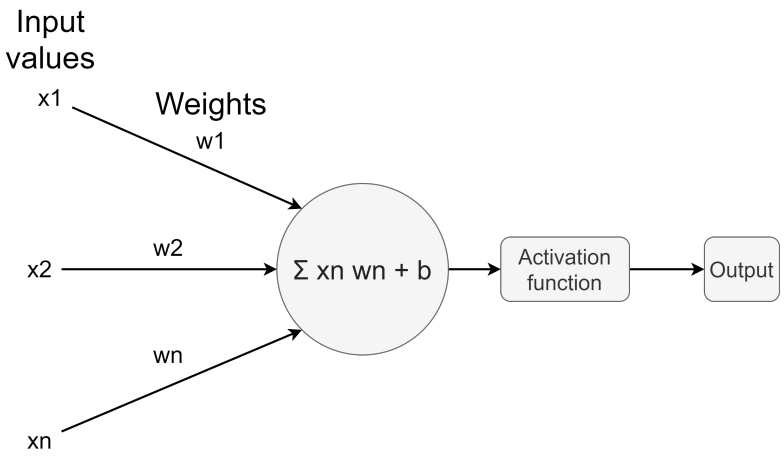

Figure 4.1. Illustrates a single neuron, also called a perceptron. It sums all inputs multiplied by a weight and then adds a bias. This is used as input in an activation function, which then delivers the output. [Aggarwal et al., 2018]

The activation function of each generic neuron in this project was chosen as Rectified Linear Unit (ReLU) and can be seen in figure 4.2. It was chosen due to its stability, sparsity and reduced likelihood of vanishing gradient during the training process. Generally networks using ReLU tend to converge in a more stable manner. The function also provides a lower range of values with most being near zero, which is referred to as sparsity, and which reduces training time and memory problems. It is also less likely to cause the vanishing gradient problem, where the change of weights change by a much lower value than they are supposed to. [Aggarwal et al., 2018]

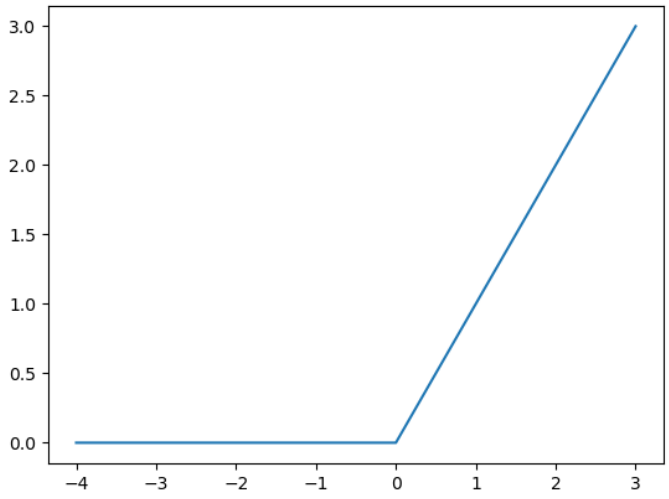

Figure 4.2. Illustrates the ReLU function which is used after each neuron in hidden layers. As with any information in neural networks, the axis are unitless. [Aggarwal et al., 2018]

The final value of each neuron is then input into each subsequent neuron of the next layer in the format referred to as Fully Connected as shown in figure 4.3. This high amount of neurons in a series of hidden layers is able to sufficiently approximate most needed functions required to properly determine a class of a sample. [Aggarwal et al., 2018]

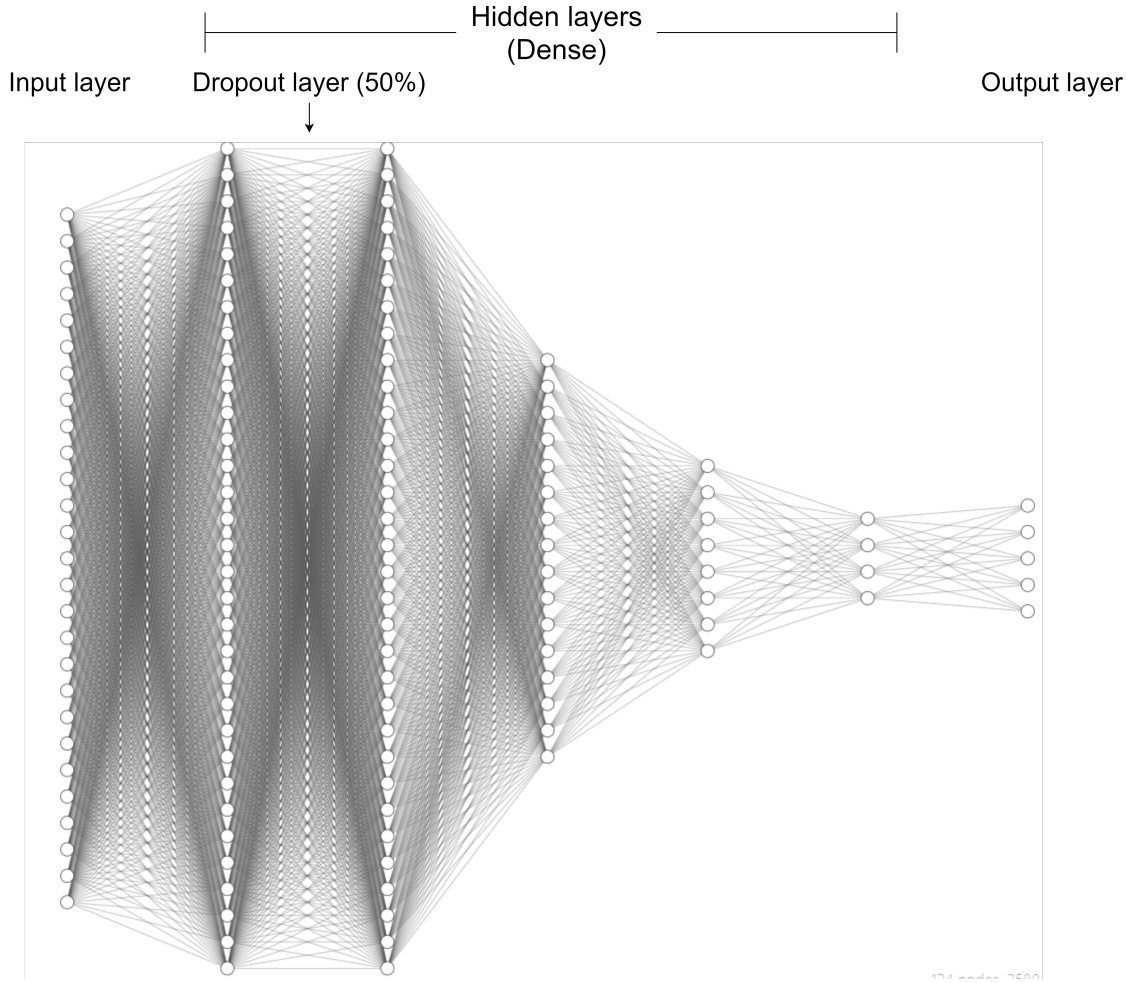

Figure 4.3. Illustrates a visual example of the Dense architecture for classification using the preprocessed features consisting of frequency bands and entropy values. Note that although the ratio of neurons between each hidden layer is 1-to-1, the actual number of neurons in the hidden layers has been reduced by a factor of eight for visual reasons.

Furthermore, the input layer contains 27 neurons due to the 27 features, and the output layer contains five neurons which correspond to the five different labels, or two neurons if the labels are "change" or "no change" as explained later in section 4.1.4. An overview of the full architecture can be seen in table 4.1.

Layer	Units or dropout rate
Input	27
Dense	256
Dropout	0.5
Dense	256
Dense	128
Dense	64
Dense	32
Output	5 or 2

Table 4.1. Illustrates the architecture of the simple neural network as defined in Keras. The output has five neurons if a test is trying to predict a specific label out of the five labels, while it has two if it is trying to predict when a change into any other label might occur which represents no change or change.

The output layer was decided to utilize the softmax activation function, because this outputs a multinomial probability distribution of the likelihood of each label, where the sum of all labels would equal 1. [Chollet, 2018] This was done to provide an opportunity for a graduated output like [0.03, 0.17, 0.05, 0.74, 0.01]. The final decision will still be the label with the highest probability, but a probability distribution can be more useful for an experimenter than a single absolute value if a nuanced output is desired. This will allow an experimenter to assess a continuous change in time of the probability distribution, where the experimenter can witness the increase of one probability of one label while another might decrease, thus representing that the DoA is slowly changing.

The model should be able to approximate the correct outcome by changing each individual weight of each neuron during its training process depending on how succesful the training process is. Though, before training, the data has to be prepared for use in a neural network. [Aggarwal et al., 2018]

4.1.2 Training, validation and test set

To train the model, it requires a training, validation and test set. The training set is used to calculate the appropriate changes to the weights, while the validation set is then used to continuously assess the performance of the current iteration of the neural network. Finally when the training process has finished, the test set is used as an independent set. This separation of data sets is necessary, since not doing so would ensure the model trains itself to remember individual samples instead of tendencies, thus reducing generalization, which would make the model less useful when presented with new data in the future. [Aggarwal et al., 2018]

The data sets were split using a 70/15/15 ratio to the training, validation and test set respectively. This ratio adheres to the requirement that the training set is larger than the validation and test set, and it was chosen to accommodate the small amount of data samples available in this project. The small amount of samples does not allow much variation in the ratio, if the data sets were to be stratified according to the labels which is preferred. This is because one of the labels with the fewest samples was not being represented equally in each of the data sets with some other tested ratios. Stratifying the data sets according to the labels was therefore necessary, because each data set should represent an equal amount

of labels to prevent the model from learning tendencies of only some labels while it is being tested later with different labels [Aggarwal et al., 2018; Chollet, 2018].

The samples in each data set is also shuffled. Along with the Monte-Carlo cross-validation, this will also ensure that the neural network will not learn any specific pattern related to inter-sample relationships, e.g. the tendency of many similar samples to be near each other.

4.1.3 Training process

The model starts with randomly determined weights and biases of each neuron, and then trains the model by following the training flowchart as shown in figure 4.4. The EEG, frequency bands and entropy values are transformed in the neural network layers, which are then used to to output predictions. These predictions are compared to the actual labels and a loss function calculates the degree of error of the model, which is referred to as loss. An optimizer function then uses the loss to calculate appropriate changes to the neurons in the layers, at which the training process then continues with the same data until no improvement is possible. [Aggarwal et al., 2018]

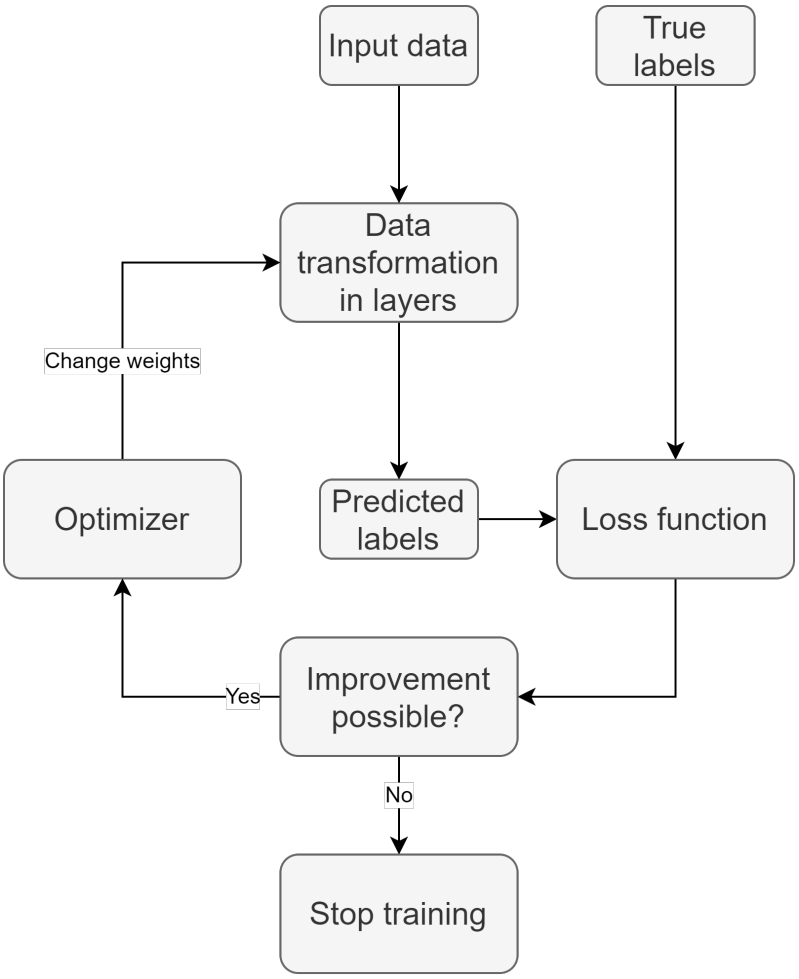

Figure 4.4. Illustrates the training process of a neural network.

Loss function

After the input has passed through the entire neural network architecture, the output layer provides values dependent on the output activation function. Then, a loss function will calculate the appropriate changes to the weights. For this project, this was chosen as categorical cross entropy. This was chosen due to the data being categorical instead of continuous values, e.g. mean square error commonly used for images and the like, and because cross entropy is a commonly used and reliable loss function. [Chollet, 2018]. The ideal probability distribution would be [0, 0, 0, 1, 0] while the output might be [0.03, 0.17, 0.05, 0.74, 0.01]. It is this difference that the loss function tries to calculate, which will be used to determine what the weights and biases should have been instead for a better prediction. [Aggarwal et al., 2018]

Optimizer

After the loss function has calculated the needed differences for a better prediction, the optimizer algorithm will calculate an appropriate change. Instead of making the changes to what the loss function suggests, smaller steps are taken instead to ensure stability of the training process. This helps developing a better performing model in the end, because the model can continuously evaluate its performance and its changes. Most optimizers are gradient-based and calculates the derivative of a loss function. The optimizer used in this project was chosen to be the Adam optimizer, because it has proven to work well for most generalized problems and combines the properties of several previously used optimizers. To determine the precise adjustments of the weights and biases, the exact gradient for each neuron is calculated using backpropagation by using the Adam optimizer. The size of the steps taken down the gradient, i.e., the size of the change to the model, is multiplied by the learning rate. In this project it was chosen as 0.01 after testing with 1, 0.1, 0.01, 0.001 and 0.0001 on the data without trying to forecast the depth. Neural network training also calculates the changes based on more than one sample at a time. With no memory constraints, this is preferred to be with all samples at once, thus the batch size was determined to be all samples of the training set. [Aggarwal et al., 2018; Chollet, 2018]

Dropout layer

Dropout layers can be used to temporarily stop some neurons from working in each iteration by setting their weight to zero. This prevents the neural network from finding tendencies in the data by using the same paths across the neurons. Instead, the temporary stop of some neurons forces neural networks to utilize different paths, which provides a variety of ways to process the data. This becomes useful when the neural network is presented with similar data though in patterns it is not used to. Thus, by using dropout layers the model becomes less overfit to specific tendencies and can instead provide a more generalized prediction that works for data not just in the training set. [Srivastava et al., 2014]

For this architecture, a single dropout layer was implemented after the first Dense layer because this was one of the layers including the highest number of parameters and was therefore the most likely layer to overfit. Multiple dropout layers did not show a noticeable change and were therefore not added. The dropout rate, which represents how many

neurons are temporarily stopped in each iteration, was set to 0.5. Thus, half of all neurons in the layer were temporarily stopped which is expected to be the most suitable ratio for generalized problems, though result may vary from use case to use case. [Srivastava et al., 2014]

Callback functions

Callback functions are non-mandatory functions that are activated before or after a training iteration. Two were used in this project: "Reduce learning rate on plateau" and "Early stopping", as provided by the Keras library.

The "Reduce learning rate on plateau" callback reduces the learning rate by a set factor after a set number of training epochs where the performance did not improve. Initially, relatively large steps of change in weights are taken for each iteration because of the the Adam optimizer. Though, the large size of the steps has the possibility of stepping over the global minima of the solution space. To help convergence, this callback ensures subsequently smaller steps are taken when the model no longer improves its performance by reducing the learning rate. [You et al., 2019]. The learning rate was set to decrease by a factor of ten, if the validation loss does not decrease over a pre-defined patience. The patience of this callback for the simple network was set to five epochs.

The "Early stopping" callback was used to stop the training process if the validation loss does not decrease over a pre-defined patience. Typically, the training process keeps going until a set number of epochs has been reached. Though, this typically shows the training and validation accuracy to keep improving until it starts to overfit, at which the training and validation accuracy does not correctly correspond to the test accuracy. Instead, it is preferable to stop the training process early.[Aggarwal et al., 2018] The patience was set to ten epochs for the simple network to allow the "Reduce learning rate on plateau" callback to see if its change will contribute something useful. Thus, five epochs with no improvement reduces the learning rate, while five additional epochs with no improvement stops the training process.

4.1.4 Tests of the simple neural network

All tests are done with ten identical runs of the code for the MCCV. Although the code for each run is identical, random initialization of the k-means algorithm and the neural network neurons, and shuffling of samples of the data sets, might provide different results.

Initial test without forecasting

First, an initial test was done where the neural network got the labels of the sample itself, i.e. no forecasting. The loss value and accuracy can be seen in figure 4.5, which demonstrates a healthy training process for this type of data and training, as evidenced by a steady and stable decrease or increase in loss or accuracy, respectively.

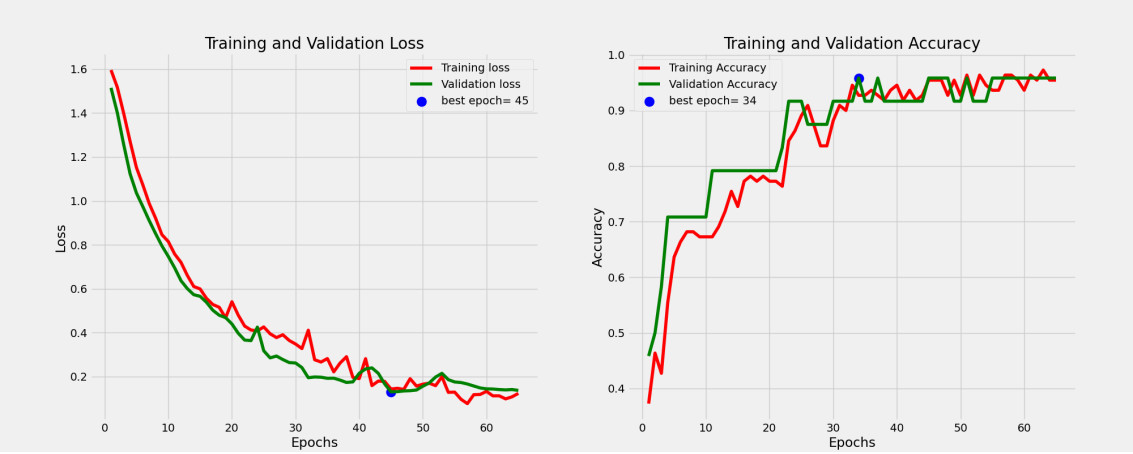

Figure 4.5. Illustrates the intended decrease in loss and increase in accuracy. The run for the graphs was chosen based on the median accuracy of the MCCV.

The results of the runs showed an accuracy and standard deviation of **95.0%** ± **3.6%**. Of these, one of the runs with the median value was extracted for investigation as shown in table 4.2. It can be seen how the labels are correctly predicted for all labels except a single sample, as evidenced by the correctly predicted labels shown in the diagonal.

	Prediction				
Truth	**9**	0	0	**1**	0
	0	**2**	0	0	0
	0	0	**6**	0	0
	0	0	0	**5**	0
	0	0	0	0	**1**

Table 4.2. Illustrates a confusion matrix of one of the runs of the initial test with no forecasting. A run with an accuracy equal to the median of the MCCV was chosen. Numbers not zero were in bold for visual reasons.

Shifting the labels for forecasting

To investigate the predictive capability, the labels were shifted pr. pig. This was done as shown in figure 4.6, where if a hypothetical pig had six samples of EEG as illustrated by the sample number and associated label, then the labels were shifted into descending order. Thus, sample number three would obtain the label of sample number four. Though, this does mean that the last sample was removed, because this sample was the latest and it could not obtain a label from a future sample. The shifting of one label will mean forecasting of 12 minutes due to the 12 minute wait between baseline recordings.

Before label shift

Sample number	0	1	2	3	4	5
Label of sample	0	0	1	1	3	2

After label shift

Sample number	0	1	2	3	4
Label of sample	0	1	1	3	2

Figure 4.6. Illustrates the method of label shifting.

Two tests were done with label shifting, as the neural network was run with both single label shifting and double label shifting, i.e. the labels were shifted two samples instead of one.

First, the labels were shifted once.

The results of the runs showed an accuracy and standard deviation of **67.0% ± 9.8%**. Of these, one of the runs with the median value was extracted for investigation as shown in table 4.3. It can be seen how the predictions begin to deviate from the diagonal. Notably, all samples of the third row were predicted wrong, and all other labels also had at least one incorrect prediction. It can be assumed that the neural network would prioritize the labels with the most samples since this would give the lowest loss, which can be seen with most of the samples from row three being predicted as samples of the two labels with the most samples (row one and two, as shown by the predictions in column one and two).

	Prediction				
Truth	**4**	**1**	0	0	0
	1	**8**	0	0	**1**
	2	**1**	0	0	0
	0	0	**1**	**3**	0
	0	0	0	0	**1**

Table 4.3. Illustrates a confusion matrix of one of the runs with single label shifting. A run with an accuracy equal to the median of the MCCV was chosen. Numbers not zero are in bold for visual reasons.

When the labels were shifted twice, the results of the runs showed an accuracy and standard deviation of **63.6% ± 9.1%**. Extracting the run with the median value for investigation provided the confusion matrix seen in table 4.4. The accuracy decreased, although only slightly despite the range of forecasting increasing from approximately 12 minutes to 24 minutes. Though, many of the labels of the EEG were similar to those around them, which is expected since the animal trial attempted to maintain a specific DoA and tried to prevent the DoA from reaching either a too light or a too deep DoA. Thus, the forecasting results here suffer slightly from the fact that most of the samples obtained the same label after label shifting as before label shifting.

	Prediction				
Truth	**6**	**2**	**0**	**1**	0
	0	**2**	0	0	**1**
	0	**1**	**0**	0	0
	0	0	0	**4**	0
	1	**1**	0	**1**	**2**

Table 4.4. Illustrates a confusion matrix of one of the runs with double label shifting. A run with an accuracy equal to the median of the MCCV was chosen. Numbers not zero are in bold for visual reasons.

Changing labels to detect changes only

To better investigate the effect of forecasting between two labels that change, the labels were changed from being multiclass to being binary, hereby 1 for change or 0 for no change. If a label was changed was determined by investigating the sample ahead of it in time, as with the label shifting method. Then, if a current sample had a similar label as the future sample, the current sample would have its label changed to 0. If it was different, it was changed to 1.

This change also required changing the output layer from five neurons to two, since now only two different labels exist.

The accuracy for this test was **76.5% ± 2.9%**. Extracting the results of one run with an accuracy closest to the median yielded the confusion matrix of table 4.5. The neural network learned to always predict the no change result, likely because this yielded the lowest loss since it was also the category with the most samples. Thus, a better accuracy was achieved by always predicting no change. This is likely due to difficulty for the neural network to find the patterns of why the label might change in the next approximately 12 minutes.

		Predicted	
		No change	**Change**
Truth	**No change**	18	0
	Change	6	0

Table 4.5. Illustrates a confusion matrix of one of the runs with labels for change or no change in labels.

To investigate this, the loss and accuracy plots were generated, which can be seen in figure 4.7 and 4.8. The ten plots of the MCCV showed different results. Some such as figure 4.7 showed a healthy training process that still ended up predicting mostly no change on all samples with a few exceptions. This was contrary to figure 4.8 which predicted solely no change. This might suggest that some runs catch on to some patterns it tries to learn, until it learns that it is still better to predict no change due to the uneven distribution of labels.

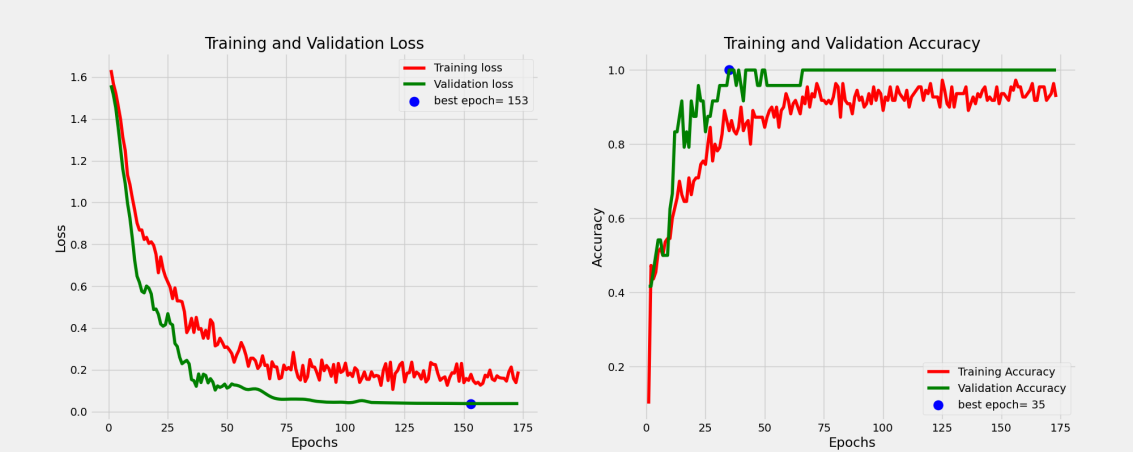

Figure 4.7. Illustrates a seemingly healthy training process of one of the runs of a test with labels indicating change or no change in labels.

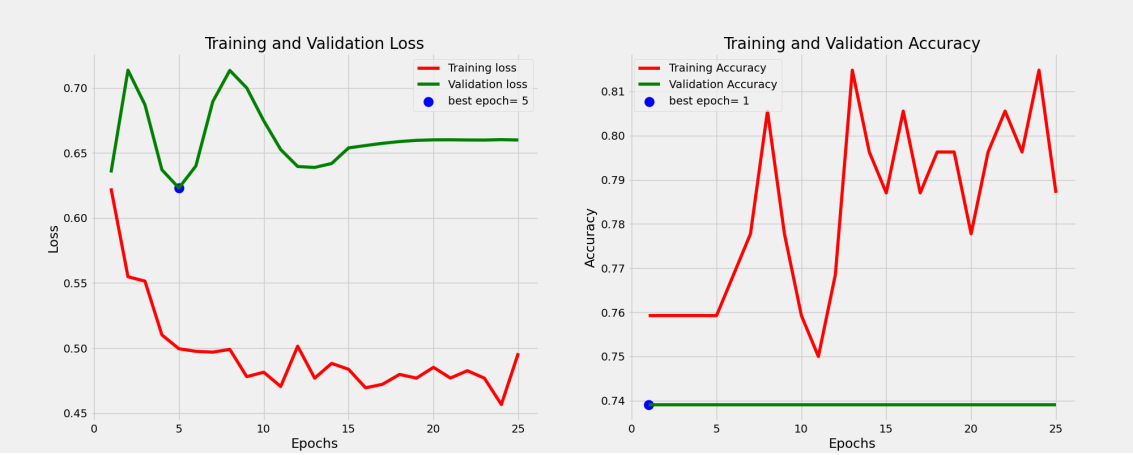

Figure 4.8. Illustrates an unhealthy training process of one of the runs of a test with labels indicating change or no change in labels.

Investigating the probability distribution of each sample with the label change yields the tabel 4.6. Here a relatively similar probability distribution can be seen of each of the samples where a change was gonna happen in the future. Though, sample six is more equal in its prediction, although it is still above 0.5 for "No change". This does however indicate that either this happened by change or the neural network was about to find some kind of pattern that might be relevant.

Sample number	No change probability	Change probability
1	0.79417	0.20583
2	0.80801	0.19199
3	0.86367	0.13633
4	0.86684	0.13316
5	0.87156	0.12844
6	0.53559	0.46441

Table 4.6. Illustrates the probability distribution of the two labels "Change" or "No change" for each of the samples with the label as "Change". The sample number is arbitrary and only for illustration.

To prevent the neural network from choosing the solution of solely predicting no change out of convenience of the uneven distribution of labels with change or no change, the labels were given different weights when calculating loss. An average ratio of the distribution over MCCV was calculated as 4:1 labels of no change pr. label with change. This was inserted into the loss function of the neural network by multiplying the losses of labels with no change with 0.25. Thus, the labels should have an equal representation in the loss function despite having an unequal distribution in the data sets.

The accuracy for this test was **69.2% ± 9.9%**. It is noticeable how the standard deviation increases nearly four-fold, as the neural network has to also dedicate its training process to learning the samples with labels for change. Though, the accuracy does decrease.

Extracting the confusion matrix from the run with a median of the MCCV yields table 4.7. Here it can be seen how the neural network prioritizes the training process to find the samples with change, because they are more rewarding for minimizing the loss function. Though, only one were found and five samples with change were predicted as no change, while reversely four samples without change were predicted as change.

The sensitivity of table 4.7 was 16.67% and the positive predictive value was 20%.

		Predicted	
		No change	**Change**
Truth	**No change**	13	4
	Change	5	1

Table 4.7. Illustrates a confusion matrix of one of the runs where the labels indicated change, but where weights were added to the loss functions. The run with an accuracy equal to the median of the MCCV was chosen.

Again, the probability distribution of the samples with the label for change can be extracted as seen in table 4.8. This time it is noticeable how there is a more even distribution of the probabiltiy distribution. This might indicate that there is a pattern to be learned, if more relevant information could be fed into the neural network. Sample two also indicates a much higher probability, suggesting a pattern was learned for this specific sample. Though, it is still only around 74%, which is still a mediocre accuracy and should be higher for practical use.

Sample number	No change probability	Change probability
1	0.54371	0.45629
2	0.25936	0.74064
3	0.54414	0.45586
4	0.54151	0.45849
5	0.55902	0.44098
6	0.55866	0.44134

Table 4.8. Illustrates the probability distribution of the two labels "Change" or "No change" for each of the samples with the label as "Change". The sample number is arbitrary and only for illustration. For this table, the weights of the labels for the loss function was changed to accommodate the uneven distribution of labels.

Additional changes to the class weight, e.g. prioritizing the samples with the label of

change, tend to quickly shift the predictions into just that label. Thus, if the class weights of no change was 0.3305 but for change it was above 1, e.g. 1,5 or 2, this would cause the neural network to prioritize all predictions to that label.

Summary of the tests

Overall, the results does indicate a possibility of something learnable in terms of pattern recognition. This could hopefully be solved by more complex information such as analyzing the whole EEG by every data point instead of by splitting its properties up into three parts as with the simple neural network.

The tests with label shifting showed a decrease of accuracy with forecasting further into the future, though this is unreliable considering samples around each other tend to have labels similar to each other.

The tests with the labels indicating change or no change does indicate that something be learned in relation to the samples immediately before a change in labels.

4.2 Complex neural network with feature extraction

Where the simple neural network was expected to lack in complexity to properly forecast, the complex neural network was expected to be able to analyze the whole signal. The complex neural network was also expected to be able to utilize the simple neural network as a metaphorical stepping stone in its training process or decision making.

This model therefore does not use the previously explained features for analyzing the samples. Instead it uses the downsampled version of the full EEG of a sample as a time-series signal where the only preprocessing done besides downsampling has been the initial filtering used in the original experiment. Though, the samples are still clustered based on the extracted features, meaning that the labels are directly associated with the features, even if the model is not.

Due to the complex neural network taking longer to train, only three runs were used for the MCCV for practical reasons.

4.2.1 Architecture of the complex neural network

The main feature of the complex neural network is the LSTM unit. An LSTM unit can be seen in figure 4.9. It works around the concept of recurrence, which is used to analyze the time domain of data. It analyzes the interdependence of data points over the entirety of a time series by processing each data point stepwise. Each LSTM unit processes one data point, meaning a time series of 100 data points would involve 100 LSTM units stacked on each other, feeding into the next unit. Figure 4.9 shows how the data point currently being processed gets concatenated with the previous data point and then undergoes a series of activation functions in the bottom horizontal processing line called the cell state, which serves a short term memory. The top horizontal line is called the hidden state and serves as long term memory from many previous data points.

In figure 4.9, the leftmost sigma activation function is called the forget gate and decides the degree of importance of the hidden state by multiplying the hidden state with a value between zero and one, which is decided based on the cell state. The middle sigma activation function similarly decides the degree of importance of the tanh activation function next to it, whose purpose is to add to the hidden state with the current cell state. Thus, these two activation functions contribute to the long term memory. Finally, the rightmost sigma activation function decides how much of the current cell state is supposed to be input into the next LSTM unit, while the tanh activation function from the hidden state similarly contributes the long term memory to the current data points output. Together, the cell- and hidden state will output to the next LSTM unit, which altogether with training and changes of neurons will be able to analyze patterns in time series useful for predicting the future DoA. [Chollet, 2018; Goyal et al., 2018; Olah, 2015]

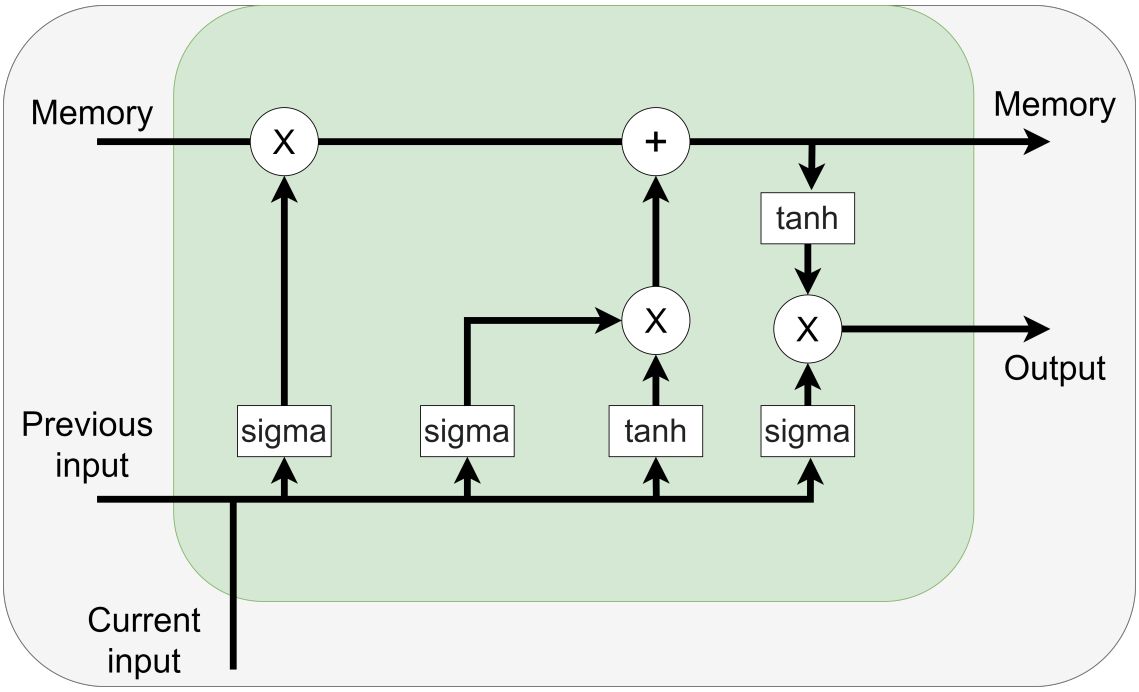

Figure 4.9. Illustrates an LSTM unit. Sigma and tanh refers to the corresponding activation functions. The X and + symbols refer to multiplicative or additive operations. The top horizontal line corresponds to learned memory from previous data points, while the bottom horizontal line corresponds to the current data point. [Aggarwal et al., 2018; Chollet, 2018]

The architecture used in this project for the complex network was inspired by the study of Michielli et al. [2019] and Nagabushanam et al. [2019], which used two LSTM layers with around 56 to 125 units. Nagabushanam et al. [2019] then used Dense layers, while Michielli et al. [2019] did not, though the difference between the two was not expected to be likely to make a significant change. A compromise of the two was created after some initial tests, which resulted in the architecture shown in table 4.9.

Layers	**Units**
Input	61035, 1 or 6
LSTM	64
LSTM	64
Dense	256
Dense	128
Dense	64
Output	5

Table 4.9. Illustrates the architecture of the complex neural network as defined in Keras. The first part of the input dimensions was 61035 due to the amount of data points in an EEG. The second part of the input dimensions was either 1 1or 6, depending on whether the ordinary EEG or the six bandpassed EEGs containing separate frequency ranges of the signal.

Thus, the full architecture consisted of two LSTM layers where the first LSTM layer returned the sequence to the next layer, i.e. the long-term memory, while the next LSTM layer did not, since this is not applicable to the next layers, only other LSTM layers. The

output layer was a Dense layer of five neurons with the softmax function like with the simple neural network.

4.2.2 Tests of the complex neural network

The EEG samples were passed to the input of the neural network, which over three runs for Monte-Carlo cross-validation yielded three identical outcomes of the same value being 37.5% in accuracy. The loss and accuracy of one of them can be seen in figure 4.10, while a confusion matrix can be seen in table 4.10. The confusion matrix was also identical for all runs. Figure 4.10 shows how the training loss barely changes by starting around 1.53 and going down to 1.38, which is contrary to a more succesful figure from the simple neural network like figure 4.5 where the loss starts around 1.6 and falls to below 0.2. Table 4.10 also shows how the label with the most samples gets all predictions, which explains the small change from a loss of 1.53 to 1.38, where the initial loss is random predictions, while the small change to 1.38 is likely the point at which the neural network figures out the optimal decision for it is to always predict the label with the most samples.

These results seem to indicate a problem in learning, where the neural network is not able to associate the EEG itself with the labels and instead always guesses the label with the most samples, because this provides the lowest loss. This means that the LSTM layers can not correctly find a correlation between the EEG and the frequency band features and entropy features. This is an unexpected result, though it might show the inability of the neural network to properly understand the frequency domain based on a time domain. Though, this does not explain the reasoning for the lack of association to the entropy features, since these are time related.

Figure 4.10. Illustrates the loss and accuracy of the complex neural network using a full EEG.

	Prediction				
Truth	0	0	4	0	0
	0	0	3	0	0
	0	0	9	0	0
	0	0	7	0	0
	0	0	1	0	0

Table 4.10. Illustrates a confusion matrix of the complex neural network using a full EEG.

To help the neural network, the EEG was split into six EEGs with their own isolated frequency ranges. The same frequency ranges were used as with the features. The six EEGs were created based on the original EEG by using a 10th order bandpass butterworth filter.

Then, these were passed into the neural network as six separate EEGs along a new dimension. The results can be seen in figure 4.11 and the confusion matrix was identical to 4.11. It is expected for the neural network to be able to associate the six bandpassed EEGs with the labels, since the labels are created using single value features of the frequency ranges that are the exact same as the bandpassed EEGs, yet the neural network seems incapable of this.

Figure 4.11. Illustrates the loss and accuracy of the complex neural network using a six bandpassed EEGs.

This created questions about the architecture choice, at which three Dense layers were added with 256, 128 and 64 neurons were added, which yielded the same result regardless of it being used with the original EEG or the six bandpassed EEGs.

A more complex architecture was then found from Xu et al. [2020], which used convolutional layers with max pool layers before two LSTM layers of 64 and 64 units, followed by three Dense layers of 256, 128 and 64 neurons. The convolutional started with a convolutional layer of 64 filters, a max pooling layer that halves the input and then three convolutional layers of 128, 512 and 1024 filters.

This approach yielded the exact same results, regardless of whether it was using the original EEG or the six bandpassed EEGs.

Overall, this shows a problem of the LSTM units to properly analyze the EEGs, considering that some of the features should be relatively simple to correlate when the six bandpasssed EEGs are used.

4.3 Combined Network

4.3.1 Architecture of the combined neural network

For the combined neural network, the architectures of the simple neural network and the complex neural network were combined into two branches of the model connecting to the same output. The tables for the combined architecture can be seen in figure 4.12. To combine the information of the two branches, the neurons of the last part of each were concatenated, essentially making it a Dense layer but with the neurons of the two previous layers. Then, the information went through three Dense layers before reaching the output layer.

Layers	Units
Input	27
Dense	256
Dense	256
Dense	128
Dense	64
Dense	32
Output for concatenate	32

Layers	Units
Input	61035, 1
LSTM	64
LSTM	64
Dense	256
Dense	128
Dense	64
Output for concatenate	64

Layers	Units
Concatenate	32 + 64
Dense	256
Dense	256
Dense	256
Output	2

Figure 4.12. Illustrates the combined neural network consisting of the simple neural network and the complex neural network. The concatenate layer combines the output of each of the simple and complex branch and forwards the combined information into Dense layers before arriving at the output layer. The output layer has two neurons for the labels of change and no change.

The output layer only had two neurons, as the tests of the combined neural network were made to detect change or no change, as with the simple neural network. This was done, because it was expected to be the most useful output for an experimenter, which merely need a warning of when to pay additional attention to a possible need in regulation of anesthesia.

4.3.2 Tests of the combined neural network

As with the complex neural network, the Monte-Carlo cross-validation was completed with three runs. All runs utilized weighted labels, since this was expected to ensure the neural network did not neglect training itself to recognize future changes in the DoA, since there are fewer samples with a change associated with it than those with no change. The weight

was set to 0.25 for no change labels because the ratio of samples with no change to those with change was 4:1.

After investigation of the probability distributions, it was decided to illustrate the performance of all three runs instead of just the run with the median performance. The confusion matrix of the first, second and third run can be seen in table 4.11, 4.13 and 4.15. Similarly, the probability distributions of the same runs can be seen in table 4.12, 4.14 and 4.16 respectively.

The confusion matrices show how the first and third run were similar in regards to the fact that the neural networks found a balanced approach to the predictions, while the second run only predicted change, because it could not find a better way to reduce its loss. This shows the variance caused by shuffling of samples and random initialization of the neural network.

Compared to the previous tests with the simple neural network, the samples had a tendency to have the probability distribution be around 0.4-0.6. However, different tendencies was seen in the combined neural network for some samples. Sample 5 and 6 of the first run showed high values of 0.897 and 0.881, respectively. For the second run, sample 1 and 5 showed a value of 0.744 and 0.947, respectively. For the third run samples were showing similar tendencies to the simple neural network. Although this does show an instability between runs, it also shows that occasionally some runs generate a high degree of confidence for some specific samples. This suggests a recognized pattern, and these high values were only found in samples with change, suggesting again that a pattern exists for when the DoA is about to change.

Considering this was not shown in the simple neural network, it suggests that a mixed data approach might be useful to provide a stepping stone by using the data from the simple neural network for the complex branch where LSTM layers analyze the full EEG.

In the simple neural network test where the aim was to identify the changes in DoA and where the weights were also equalized, the sensitivity was 16.67% and the positive predictive value was 20%.

For the complex neural network, the sensitivity for the first, second and third run was 33.3%, 100.0% and 66%. When disregarding the second run, since it decided to always predict change, the first and third run does show an average sensitivity around 50%.

The positive predictive value of the first, second and third run was 28.6%, 25.0% 33.3%, respectively. The positive predictive values show how 25.0-33.3% of the samples predicted to have a changing DoA actually does have this change in them.

Overall, the performance of the combined neural network appeared better at predicting changes than the simple neural network. Additionally, some samples had a high probability of being correct according to the neural network as shown in the probability distribution.

		Predicted	
		No change	**Change**
Truth	**No change**	13	5
	Change	4	2

Table 4.11. Illustrates a confusion matrix of the first run of the combined neural network.

Sample number	No change probability	Change probability
1	0.50867	0.49133
2	0.56778	0.43222
3	0.57835	0.42165
4	0.51747	0.48253
5	0.10307	0.89693
6	0.11902	0.88098

Table 4.12. Illustrates the probability distributions of samples that had the label for change in the first run of the combined neural network. The sample number is arbitrary and only for illustration.

		Predicted	
		No change	**Change**
Truth	**No change**	0	18
	Change	0	6

Table 4.13. Illustrates a confusion matrix of the second run of the combined neural network.

Sample number	No change probability	Change probability
1	0.25611	0.74389
2	0.47523	0.52477
3	0.48843	0.51157
4	0.41833	0.58167
5	0.05315	0.94685
6	0.49096	0.50904

Table 4.14. Illustrates the probability distributions of samples that had the label for change in the second run of the combined neural network. The sample number is arbitrary and only for illustration.

		Predicted	
		No change	**Change**
Truth	**No change**	10	8
	Change	2	4

Table 4.15. Illustrates a confusion matrix of the third run of the combined neural network.

Sample number	No change probability	Change probability
1	0.50781	0.49219
2	0.41790	0.58210
3	0.45680	0.54320
4	0.50334	0.49666
5	0.37223	0.62777
6	0.49160	0.50840

Table 4.16. Illustrates the probability distributions of samples that had the label for change in the third run of the combined neural network. The sample number is arbitrary and only for illustration.

Discussion 5

In summary, the aim of this project was to investigate how a neural network could be used to predict the future DoA by a minute or more using a mixed-data neural network. This was done by creating a combined neural network using two parts: The architecture of the simple neural network which uses the preprocessed features in Dense layers, and the architecture of the complex neural network which uses the EEG in LSTM layers.

5.1 Main findings

The tests with the simple neural network provided a baseline accuracy for detecting changes, which was useful to compare it to the combined neural network using mixed data to evaluate whether or not combining both preprocessed features and the full EEG would improve accuracy to forecast changes to the EEG.

The complex neural network showed how it was not possible to predict the labels properly with the sole use of the EEGs only. Though, it was tested to see if this problem was alleviated by using the data of the simple neural network as a stepping stone to better converge the training process and become able to succesfully process the full EEGs.

The combined neural network had a better performance than the simple neural network and also showed how some samples had a very high value in the probability distribution. This suggests that there is an unknown pattern of information in the full EEG that a LSTM unit is likely able to analyze, and which is not information found in the features used in the simple neural network. Additionally, it also shows how a mixed-data neural network can be used to improve the performance compared to a neural network using only preprocessed features.

5.2 Processing of full EEG

The complex neural network was not able analyze any patterns by itself, as evidenced by the plots of the loss and the accuracy during the training process. This was an unexpected finding, considering that LSTM often excel when using time series signals and can be used without additional layers [Aggarwal et al., 2018]. LSTM has also been used specifically for seizure detection [Hu et al., 2020; Xu et al., 2020], where it proved useful in processing the temporal information to predict the event of a seizure with sufficient warning. It has also been state-of-the-art for purposes like categorizing sleep stages [Michielli et al., 2019]. Specifically for DoA, it has been used to categorize the current DoA and has shown to have a good accuracy compared to methods solely using features instead of the full EEG [Li et al., 2020].

Though, this raises questions on why the LSTM layers failed to detect the DoA by itself as shown by the results of the complex neural network. A likely reason can be the amount of samples. Neural networks generally require large amounts of data, where as 157 samples were present in this project [Aggarwal et al., 2018]. Additionally, a few samples also had to be removed for each pig due to the method of changing the labels to make the neural network train itself to detect changes. Considering each EEG consisted of 61,035 data points long, this may also have played a factor, because longer EEGs might provide some part of information redundancy, and because the more information is present, the more needed a larger sample size is. This is especially relevant for LSTM, which [Aggarwal et al., 2018] Nonetheless, Li et al. [2020] did succesfully use LSTM for the full EEG, although this was on humans.

5.3 Labelling and golden truth

One of the most important things when working with machine learning is the availability of representative labels, preferably a golden truth [Aggarwal et al., 2018]. However, for DoA no golden truth exist and no consensus exists for the most appropriate substitute for a golden truth [Silva and Antunes, 2012]. In this project, the labels for the truth of the data was created with the kmeans algorithm based on the selected features, which has several implications. First, kmeans yields categorical values which is not necessarily comparable to the use of a linear scale of DoA [Silva and Antunes, 2012]. Contrarily, cortical information is not simple nor linear [Gu et al., 2019; Silva and Antunes, 2012]. This might support the idea of using categorical labels to simulate how the DoA can move along more dimensions than a linear scale.

Another way of assessing the labels is the use of mean square error instead of categorical labels. In this project, the data was obtained from pigs whose DoA was attempted to be kept consistent due to the experimental procedures. This would be a problem for categorical values, since they preferably need to have labels from all different DoA. Since the pigs were assumed to be kept in a consistent DoA, mean square error could be used to denote this DoA and anything deeper or lighter would provide a higher error, thus telling the experimenter to be vigilant. This would not provide the neural network with the opportunity to inform the experimenter of the direction of a change, but it will provide information of how much a sample deviates from the DoA of the training process. Another advantage is how it provides a non-categorical value for the training process, since continuous values provide a higher variability for the neural network, which would provide more information for the neural network to work with. This would help the learning process, since two samples with the same label might not be identical. One sample might be about to change into a deeper or lighter DoA, while another sample with the same label might be very likely to stay in the same DoA. A continuous value such as the mean square error would thus provide nuance for the training process and would circumvent the need for a categorical truth.

5.4 Data quality and time domain

The data consisted of baseline recordings with around 12 minutes between them. The samples with EEG consisted of 30 seconds with 12 minutes between each EEG. The results indicate, that 12 minutes may be either too late or the amount of information in 30 seconds was too low for sufficient forecasting. It is uncertain what a sufficient time frame of forecasting is for a reliable but still practical result, which prompts the need of research investigating EEGs both with fewer and more minutes of wait between them, and EEGs of a much higher recorded length. It is worthwhile to investigate the opportunity for multiple EEG samples

Similarly, each EEG was split into three sets of features to provide a set of features with a three-stage progression from the first 10 seconds to the last 10 seconds of each EEG. This was expected to provide a somewhat usable set of features for labeling, since samples could be labeled based not only on their current DoA, but also based on a progression in the features over the 30 seconds. Whether this was useful is not certain, and also prompts the need for a reliable golden truth of the DoA.

5.5 Future considerations

Three considerations are relevant for the outcome of this project: New mixed-data, theoretical potential and explainability.

Other types of information available for assessing the DoA is the physiological parameters, e.g. parameters related to the cardiopulmonary physiology, and also information such as the dosage of the anesthesia. [Haga and Ranheim, 2005; Haga et al., 2011; Jahanseir et al., 2018; Silva and Antunes, 2012] These are features that could also be added to the branch of simple branch of the combined neural network, along with whichever other features that were written measured during the experiments. Neural networks in general are able to differentiate between useful and useless features, meaning that theoretically, all features measured during an experiment can be added [Aggarwal et al., 2018]. It should not have any significant impact on performance, nor on the time of the prediction calculations since features of singular values are computationally light. Though, it does require that all samples have the same number of features, since samples need to have identical, which makes it difficult to combine data from multiple experiments, if they did not have the exact same features measured. it is possible to set unavailable features to zero in the feature array, but this might confuse the neural network into thinking the actual value is zero. Tools like masking layers exist to prevent similar situations where the length of time series signals in a data set are not identical by 'masking' the data points that are before the length of the time series signal with the longest length [Keras, N/A]. Thus, although access to higher quality and quantity of data is more important, it is also of interest to investigate adding new types of information into the model.

In regards to theoretical potential, neural networks with unlimited data and training time are expected to be able to approximate any needed functions and recognize the necessary patterns, while mostly being limited by convergence methods to reach the global minima of the solution space. Especially layers like LSTM have the potential to recognize patterns by

itself, which are not easily done with other machine learning methods. Though, in practice it can be difficult to obtain enough of the data of interest. Especially for use cases with a lower quantity of data, it can be relevant to investigate other machine learning methods like support vector machines and decision trees [Aggarwal et al., 2018; Chollet, 2018]. A turning point for when neural network might be better than other machine learning methods could be when the amount of samples are in the tens of thousands to hundreds of thousands, which can be difficult to acquire.

In regards to explainability, neural networks act like a black box and are difficult to understand. Explainability is a quickly developing subfield of neural networks, but it already has many frameworks that have shown good success like SHapley Additive ExPlanations (SHAP). Methods like SHAP can be good at calculating the contribution of each feature. For images, it is less simple to understand but doable, while for layers like LSTM, it can show each data point's contribution to the probability distribution of each label. Though, this can still be difficult for a human to understand, since it shows the degree of contribution of each feature and patterns, but does not necessarily explain why this is so [Carvalho et al., 2019; Lundberg]. Nevertheless, combining neural networks and explainability methods could be useful tools for recognition of unknown patterns related to DoA, that have not been found by manual review of features, as usually done when investigating correlation and causality in scientific studies. Thus, neural networks for DoA might be useful to not only create predictive models, but also for researching the found patterns.

Conclusion 6

To summarize, the aim of this project was:

> *How can a neural network be used to predict the future depth of anesthesia by a minute or more using features extracted from an EEG in combination with time series analysis by neural network layers in a mixed-data approach?*

In this project, a mixed-data neural network was created to predict changes to the future DoA by using preprocessed features of the EEG alongside time series analysis by neural network layers. The results indicate that pattern recognizing tools like LSTM layers in neural network are promising tools for future research as evidenced by the increase in performance between the simple neural network and the combined neural network. Notably, the complex neural network did not work, which suggests that the LSTM needed help to find patterns, which they used the simple neural network for in the combined model. Furthermore, the probability distribution of samples of the combined neural network showed higher values, prompting an interest into a future investigation of the patterns found by the LSTM layers, which can be done using AI explainability methods.

Bibliography

Aggarwal et al., **2018**. Charu C Aggarwal et al. *Neural networks and deep learning*. ISBN: 978-3-319-94463-0, Digital. Springer, 2018.

Baars et al., **2013**. Jan H Baars, Ulf Rintisch, Benno Rehberg, Karl Heinz Lahrmann and Falk von Dincklage. *Prediction of motor responses to surgical stimuli during bilateral orchiectomy of pigs using nociceptive flexion reflexes and the bispectral index derived from the electroencephalogram*. The Veterinary Journal, 195(3), 377–381, 2013.

Berge, **2011**. Odd-Geir Berge. *Predictive validity of behavioural animal models for chronic pain*. British journal of pharmacology, 164(4), 1195–1206, 2011.

L Beydon, JC Desfontis, F Ganster, J Petres, F Gautier, S Ferec, A Cailleux, C Dussaussoy, N Liu, T Chazot et al., *2009*. L Beydon, JC Desfontis, F Ganster, J Petres, F Gautier, S Ferec, A Cailleux, C Dussaussoy, N Liu, T Chazot et al. BIS response to tamponade and dobutamine in swine varies with hypnotic/opiate ratio. I *Annales francaises d'anesthesie et de reanimation*, volume 28(7-8), pages 650–657. Elsevier, 2009.

Biocompare, **2014**. Biocompare. *The Case for Large-Animal Models*. `https://www.biocompare.com/Editorial-Articles/168133-The-Case-for-Large-Animal-Models/`, 2014. Visited: 10/03-2021.

Campbell et al., **2020**. Justin M Campbell, Zirui Huang, Jun Zhang, Xuehai Wu, Pengmin Qin, Georg Northoff, George A Mashour and Anthony G Hudetz. *Pharmacologically informed machine learning approach for identifying pathological states of unconsciousness via resting-state fMRI*. NeuroImage, 206, 116316, 2020.

Carvalho et al., **2019**. Diogo V Carvalho, Eduardo M Pereira and Jaime S Cardoso. *Machine learning interpretability: A survey on methods and metrics*. Electronics, 8(8), 832, 2019.

Chollet, **2018**. François Chollet. *Deep Learning with Python*. ISBN: 9781617294433, Digital. Manning Publications Co., 2018.

Dankers et al., **2019**. Frank JWM Dankers, Alberto Traverso, Leonard Wee and Sander MJ van Kuijk. *Prediction modeling methodology*. I *Fundamentals of Clinical Data Science*, pages 101–120. Springer, Cham, 2019.

Dubitzky et al., **2007**. Werner Dubitzky, Martin Granzow and Daniel P Berrar. *Fundamentals of data mining in genomics and proteomics*. Springer Science & Business Media, 2007.

Edwards, **1999**. Jeanette Edwards. *Why dolly matters: kinship, culture and cloning*. Ethnos, 64(3-4), 301–324, 1999.

Ellegaard Góttingen Minipigs A/S, **2021**. Ellegaard Góttingen Minipigs A/S. *Price list 2021.* https://minipigs.dk/fileadmin/filer/Price_list/Price_list_2021.pdf, 2021. Visited: 10/03-2021.

EUR-Lex, **2010**. EUR-Lex. *Directive 2010/63/EU of the European Parliament and of the Council of 22. september 2010 on the protection of animals sued for scientific purposes.* https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32010L0063, 2010. Visited: 02/03-2021.

European Commission, **2019**. European Commission. *2019 report on the statistics on the use of animals for scientific purposes in the Member States of the European Union in 2015-2017.* https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020DC0016&from=EN, 2019. Visited: 04/03-2021.

Flecknell, **2015**. Paul Flecknell. *Laboratory animal anaesthesia.* Academic press, 2015.

Rigoberto Fonseca-Delgado and Pilar Gómez-Gil, *2013*. Rigoberto Fonseca-Delgado and Pilar Gómez-Gil. An assessment of ten-fold and Monte Carlo cross validations for time series forecasting. I *2013 10th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)*, pages 215–220. IEEE, 2013.

García et al., **2021**. Paul S García, Matthias Kreuzer, Darren Hight and James W Sleigh. *Effects of noxious stimulation on the electroencephalogram during general anaesthesia: a narrative review and approach to analgesic titration.* British Journal of Anaesthesia, 126(2), 445–457, 2021.

Gayer et al., **2016**. Steven Gayer, Howard D Palte, Thomas A Albini, Harry W Flynn Jr, Ricardo Martinez-Ruiz, Nelson Salas, Andrew J McClellan, Nidhi Relhan and Jean-Marie Parel. *In vivo porcine model of venous air embolism during pars plana vitrectomy.* American journal of ophthalmology, 171, 139–144, 2016.

Gieling et al., **2011**. Elise T Gieling, Teun Schuurman, Rebecca E Nordquist and F Josef van der Staay. *The pig as a model animal for studying cognition and neurobehavioral disorders.* Molecular and functional models in neuropsychiatry, pages 359–383, 2011.

Goyal et al., **2018**. Palash Goyal, Sumit Pandey and Karan Jain. *Deep learning for natural language processing.* Deep Learning for Natural Language Processing: Creating Neural Networks with Python [Berkeley, CA]: Apress, pages 138–143, 2018.

Gu et al., **2019**. Yue Gu, Zhenhu Liang and Satoshi Hagihira. *Use of multiple EEG features and artificial neural network to monitor the depth of anesthesia.* Sensors, 19 (11), 2499, 2019.

Haberham et al., **1999**. ZL Haberham, WE Van den Brom, AJ Venker-van Haagen, V Baumans, HNM De Groot and LJ Hellebrekers. *EEG evaluation of reflex testing as assessment of depth of pentobarbital anaesthesia in the rat.* Laboratory animals, 33(1), 47–57, 1999.

Haga and Ranheim, **2005**. Henning A Haga and Birgit Ranheim. *Castration of piglets: the analgesic effects of intratesticular and intrafunicular lidocaine injection.* Veterinary anaesthesia and analgesia, 32(1), 1–9, 2005.

Haga et al., **2011**. Henning A Haga, Birgit Ranheim and Claudia Spadavecchia. *Effects of isoflurane upon minimum alveolar concentration and cerebral cortex depression in pigs and goats: an interspecies comparison.* The Veterinary Journal, 187(2), 217–220, 2011.

Hashimoto et al., **2020**. Daniel A Hashimoto, Elan Witkowski, Lei Gao, Ozanan Meireles and Guy Rosman. *Artificial intelligence in anesthesiology: current techniques, clinical applications, and limitations.* Anesthesiology, 132(2), 379–394, 2020.

Hayase et al., **2019**. Kazuma Hayase, Kazuko Hayashi and Teiji Sawa. *Hierarchical Poincaré analysis for anaesthesia monitoring.* Journal of clinical monitoring and computing, pages 1–10, 2019.

Hu et al., **2020**. Xinmei Hu, Shasha Yuan, Fangzhou Xu, Yan Leng, Kejiang Yuan and Qi Yuan. *Scalp EEG classification using deep Bi-LSTM network for seizure detection.* Computers in Biology and Medicine, 124, 103919, 2020.

Jahanseir et al., **2018**. Mercedeh Jahanseir, Seyed Kamaledin Setarehdan and Sirous Momenzadeh. *Automatic anesthesia depth staging using entropy measures and relative power of electroencephalogram frequency bands.* Australasian physical & engineering sciences in medicine, 41(4), 919–929, 2018.

Janvier Labs, **2017**. Janvier Labs. *Price Catalogue 2017.* `https://clin.medarbejdere.au.dk/fileadmin/www.clin.au.dk/Dyrefaciliteter/CATALOGUE_JANVIERLABS_NORDIC_COUNTRIES_2017.pdf`, 2017. Visited: 10/03-2021.

Jayatilake and Ganegoda, **2021**. Senerath Mudalige Don Alexis Chinthaka Jayatilake and Gamage Upeksha Ganegoda. *Involvement of Machine Learning Tools in Healthcare Decision Making.* Journal of Healthcare Engineering, 2021, 2021.

Jeon and Chehri, **2020**. Gwanggil Jeon and Abdellah Chehri. *Entropy-Based Algorithms for Signal Processing*, 2020.

Joshi, **2020**. Ameet V Joshi. *Machine learning and artificial intelligence.* Springer, 2020.

Keras, **2020**. Keras. *The Functional API.* `https://keras.io/guides/functional_api/r`, 2020. Visited: 17/05-2021.

Keras, **N/A**. Keras. *Masking layer.* `https://keras.io/api/layers/core_layers/masking/`, N/A. Visited: 27/05-2021.

Kobeissy et al., **2016**. Firas Kobeissy, C Dixon, R Hayes and S Modello. *Injury models of the central nervous system.* Methods in Molecular Biology, 1462, 2016.

Kurita et al., **2005**. Tadayoshi Kurita, Koji Morita, Kazushige Fukuda, Masahiro Uraoka, Kotaro Takata, Yoshimitsu Sanjo and Shigehito Sato. *Influence of hemorrhagic shock and subsequent fluid resuscitation on the electroencephalographic*

effect of isoflurane in a swine model. The Journal of the American Society of Anesthesiologists, 103(6), 1189–1194, 2005.

Kurita et al., **2012**. Tadayoshi Kurita, Masahiro Uraoka, Koji Morita and Shigehito Sato. *Influence of progressive hemorrhage and subsequent cardiopulmonary resuscitation on the bispectral index during isoflurane anesthesia in a swine model*. Journal of Trauma and Acute Care Surgery, 72(6), 1614–1619, 2012.

Li et al., **2020**. Ronglin Li, Qiang Wu, Ju Liu, Qi Wu, Chao Li and Qibin Zhao. *Monitoring Depth of Anesthesia Based on Hybrid Features and Recurrent Neural Network*. Frontiers in neuroscience, 14, 26, 2020.

Li et al., **2019**. Weijia Li, Xiaohong Shen and Yaan Li. *A Comparative Study of Multiscale Sample Entropy and Hierarchical Entropy and Its Application in Feature Extraction for Ship-Radiated Noise*. Entropy, 21(8), 793, 2019.

Li et al., **2008**. Xiaoli Li, Suyuan Cui and Logan J Voss. *Using permutation entropy to measure the electroencephalographic effects of sevoflurane*. The Journal of the American Society of Anesthesiologists, 109(3), 448–456, 2008.

Liang et al., **2015**. Zhenhu Liang, Yinghua Wang, Xue Sun, Duan Li, Logan J Voss, Jamie W Sleigh, Satoshi Hagihira and Xiaoli Li. *EEG entropy measures in anesthesia*. Frontiers in computational neuroscience, 9, 16, 2015.

Liu et al., **2015**. Quan Liu, Yi-Feng Chen, Shou-Zen Fan, Maysam F Abbod and Jiann-Shing Shieh. *EEG signals analysis using multiscale entropy for depth of anesthesia monitoring during surgery through artificial neural networks*. Computational and mathematical methods in medicine, 2015, 2015.

Liu et al., **2018**. Quan Liu, Li Ma, Shou-Zen Fan, Maysam F Abbod and Jiann-Shing Shieh. *Sample entropy analysis for the estimating depth of anaesthesia through human EEG signal at different levels of unconsciousness during surgeries*. PeerJ, 6, e4817, 2018.

Lotte et al., **2018**. Fabien Lotte, Laurent Bougrain, Andrzej Cichocki, Maureen Clerc, Marco Congedo, Alain Rakotomamonjy and Florian Yger. *A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update*. Journal of neural engineering, 15(3), 031005, 2018.

Lundberg. Scott Lundberg. *SHAP*. https://github.com/slundberg/shap.

MacKenzie et al., **2018**. Kristen K MacKenzie, Angelitta M Britt-Spells, Laura P Sands and Jacqueline M Leung. *Processed electroencephalogram monitoring and postoperative delirium: a systematic review and meta-analysis*. Anesthesiology, 129(3), 417–427, 2018.

Mahendru, **2019**. Khyati Mahendru. *How to Determine the Optimal K for K-Means?* `https://medium.com/analytics-vidhya/how-to-determine-the-optimal-k-for-k-means-708505d204eb`, 2019. Visited: 25/04-2021.

Martoft et al., **2002**. L Martoft, L Lomholt, C Kolthoff, BE Rodriguez, EW Jensen, PF Jørgensen, HD Pedersen and Anders Forslid. *Effects of CO2 anaesthesia on central nervous system activity in swine*. Laboratory Animals, 36(2), 115–126, 2002.

Mazzoni et al., **2010**. Alberto Mazzoni, Kevin Whittingstall, Nicolas Brunel, Nikos K Logothetis and Stefano Panzeri. *Understanding the relationships between spike rate and delta/gamma frequency bands of LFPs and EEGs using a local cortical network model*. Neuroimage, 52(3), 956–972, 2010.

Michielli et al., **2019**. Nicola Michielli, U Rajendra Acharya and Filippo Molinari. *Cascaded LSTM recurrent neural network for automated sleep stage classification using single-channel EEG signals*. Computers in biology and medicine, 106, 71–81, 2019.

Mogil et al., **2010**. Jeffrey S Mogil, Karen D Davis and Stuart W Derbyshire. *The necessity of animal models in pain research*. Pain, 151(1), 12–17, 2010.

Montavon et al., **2012**. Grégoire Montavon, G Orr and Klaus-Robert Müller. *Neural networks-tricks of the trade second edition*. Springer, DOI, 10, 978–3, 2012.

Montupil et al., **2019**. Javier Montupil, Aline Defresne and Vincent Bonhomme. *The raw and processed electroencephalogram as a monitoring and diagnostic tool*. Journal of cardiothoracic and vascular anesthesia, 33, S3–S10, 2019.

Morgan et al., **2006**. G Edward Morgan, Maged S Mikhail, Michael J Murray and C Philip Larson. *Clinical anesthesiology*, volume 361. Lange Medical Books/McGraw-Hill New York, 2006.

Morimoto, **2008**. Yasuhiro Morimoto. *Usefulness of electroencephalogramic monitoring during general anesthesia*. Journal of anesthesia, 22(4), 498–501, 2008.

Nagabushanam et al., **2019**. Perattur Nagabushanam, S Thomas George and Subramanyam Radha. *EEG signal classification using LSTM and improved neural network algorithms*. Soft Computing, pages 1–23, 2019.

National Swine Resource and Research Center, **2020**. National Swine Resource and Research Center. *NSRRC Distribution Fees*. `https://nsrrc.missouri.edu/fees/`, 2020. Visited: 10/03-2021.

Novia, **N/A**. Data Novia. *Determining The Optimal Number Of Clusters: 3 Must Know Methods*. `https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/`, N/A. Visited: 11/03-2021.

Olah, **2015**. Christopher Olah. *Understanding LSTM Networks*. `http://colah.github.io/posts/2015-08-Understanding-LSTMs/`, 2015. Visited: 27/05-2021.

Pauldine et al., **2008**. Ronald Pauldine, George Beck, Jose Salinas and David W Kaczka. *Closed-loop strategies for patient care systems*. Journal of Trauma and Acute Care Surgery, 64(4), S289–S294, 2008.

Pilge et al., **2006**. Stefanie Pilge, Robert Zanner, Gerhard Schneider, Jasmin Blum, Matthias Kreuzer and Eberhard F Kochs. *Time delay of index calculation: analysis of cerebral state, bispectral, and narcotrend indices.* The Journal of the American Society of Anesthesiologists, 104(3), 488–494, 2006.

Punjasawadwong et al., **2014**. Yodying Punjasawadwong, Aram Phongchiewboon and Nutchanart Bunchungmongkol. *Bispectral index for improving anaesthetic delivery and postoperative recovery.* Cochrane database of systematic reviews, (6), 2014.

Richman and Moorman, **2000**. Joshua S Richman and J Randall Moorman. *Physiological time-series analysis using approximate entropy and sample entropy.* American Journal of Physiology-Heart and Circulatory Physiology, 2000.

Rosebrock, **2019**. Adrian Rosebrock. *Keras: Multiple Inputs and Mixed Data.* `https://www.pyimagesearch.com/2019/02/04/keras-multiple-inputs-and-mixed-data/`, 2019. Visited: 09/05-2021.

Schmidt et al., **2002**. M Schmidt, T Marx, C Papp-Jambor, U Schirmer and H Reinelt. *Effect of xenon on cerebral autoregulation in pigs.* Anaesthesia, 57(10), 960–966, 2002.

Konstantinos Sechidis, Grigorios Tsoumakas and Ioannis Vlahavas, *2011.* Konstantinos Sechidis, Grigorios Tsoumakas and Ioannis Vlahavas. On the stratification of multi-label data. I *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 145–158. Springer, 2011.

Seif, **2018**. George Seif. *The 5 Clustering algorithms Data Scientists Need To Know.* `https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68`, 2018. Visited: 11/03-2021.

Silva and Antunes, **2012**. Aura Silva and Luis Antunes. *Electroencephalogram-based anaesthetic depth monitoring in laboratory animals.* Laboratory animals, 46(2), 85–94, 2012.

Speaking of research. Speaking of research. *The Animal Model.* https://speakingofresearch.com/facts/the-animal-model/.

Srivastava et al., **2014**. Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever and Ruslan Salakhutdinov. *Dropout: a simple way to prevent neural networks from overfitting.* The journal of machine learning research, 15(1), 1929–1958, 2014.

Sun et al., **2020**. Yi Sun, Fan Ye, Jing Wang, Pan Ai, Changwei Wei, Anshi Wu and Wuxiang Xie. *Electroencephalography-guided anesthetic delivery for preventing postoperative delirium in adults: an updated meta-analysis.* Anesthesia & Analgesia, 131(3), 712–719, 2020.

Tacke et al., **2020**. Moritz Tacke, Eberhard F Kochs, Marianne Mueller, Stefan Kramer, Denis Jordan and Gerhard Schneider. *Machine learning for a combined electroencephalographic anesthesia index to detect awareness under anesthesia.* Plos one, 15(8), e0238249, 2020.

Tannenbaum and Bennett, **2015**. Jerrold Tannenbaum and B Taylor Bennett. *Russell and Burch's 3Rs then and now: the need for clarity in definition and purpose.* Journal of the American Association for Laboratory Animal Science, 54(2), 120–132, 2015.

The University of Adelaide, **2020**. The University of Adelaide. *Laboratory Animal Services - Current price list.* `https://www.adelaide.edu.au/animal-services/products-services/current-price-list#price-list`, 2020. Visited: 10/03-2021.

University of Michigan Medical School, **N/A**. University of Michigan Medical School. *Mouse Breeding.* `https://brcf.medicine.umich.edu/cores/transgenic-animal-model/training-education/breed/`, N/A. Visited: 22/05-2021.

Walters et al., **2017**. Eric M Walters, Kevin D Wells, Elizabeth C Bryda, Susan Schommer and Randall S Prather. *Swine models, genomic tools and services to enhance our understanding of human health and diseases.* Lab animal, 46(4), 167–172, 2017.

Xu et al., **2020**. Gaowei Xu, Tianhe Ren, Yu Chen and Wenliang Che. *A One-Dimensional CNN-LSTM Model for Epileptic Seizure Recognition Using EEG Signal Analysis.* Frontiers in Neuroscience, 14, 1253, 2020.

You et al., **2019**. Kaichao You, Mingsheng Long, Jianmin Wang and Michael I Jordan. *How does learning rate decay help modern neural networks?* arXiv preprint arXiv:1908.01878, 2019.

Literature Search A

To obtain information about the problem, what physiological data could be used to solve it, hereby the anesthetic effect on ECoG, ECG and other systemic parameters, and of what methods could be used on the data, information was obtained both structured and unstructured. The information obtained structured was composed of three structured literature searches: A Swine/Anesthesia-related search, an Anesthesia Review-search, and a Neural network/Supervised learning/Anesthesia-search. Documentation of the process of the structured searches can be seen in figure A.1.

The process of the structured searches consisted of the following steps:

- Block search
- Inclusion criteria
- Removal of duplicates
- Abstract reading and exclusion
- Full-text reading and exclusion

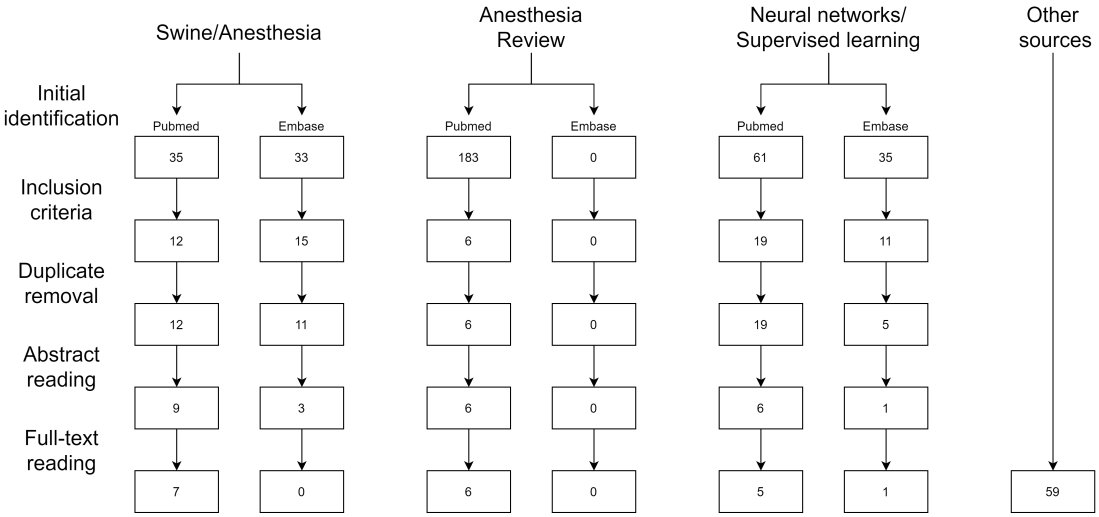

Figure A.1. Illustrates a flowchart of the process of obtaining literature in the three structured literature searches. Additionally, the literature identified outside of these searches is specified as "Other Sources".

The searches were done in the databases of Pubmed and Embase.

Block searches were done using only thesaurus terms to make sure the studies related directly to these words of interest and were not merely mentioned secondhand in an article. The specific choice of thesaurus terms are described in the sections of the respective search.

Inclusion criteria contained criteria that were common to all structured searches and criteria that were specific to each search. Common for all searches was for material to be in English, Danish, Norwegian or Swedish. The full-text also had to be available. For the Swine/Anesthesia-search and Anesthesia Review-search, the publication year was set 2001-2021, because this provided a manageable amount of results to sift through, and because this time frame was sufficiently large to obtain the newest and most relevant information. Though, the Neural network/Supervised learning/Anesthesia-search had the publication year set to 2016-2021, because of the number of obtained results would have been too high and because the field of neural network and supervised learning is seeing fast development, thus quickly making older material obsolete. The inclusion criteria would utilize the filter functions of each database and if this was not possible, the criteria would be implemented manually.

Besides the structured searches, the literature for this project was supplemented by unstructured searches and literature already identified prior to the project as seen in figure A.1 as "Other sources". This literature was obtained from either unstructured searches in scientific databases, other students or university employees, previous structured literature search made during a similar use case related to the depth of anesthesia in rats, chain searches, neural network textbooks, specific searches for sources on personally known information, and other sources of information. The majority of "Other Sources" were derived from chain searches.

A.1 Swine/Anesthesia-search

The first structured search was the Swine/Anesthesia-related search, whose purpose was to obtain information related to anesthesia in swine, and to the available data of this project, which were ECoG, ECG, a set of physiological parameters and dosages related to the anesthetic agents. The search focused on the EEG and ECG data, since the data of the physiological parameters and anesthetic agents were singular values, they were expected to be ready to be implemented without preprocessing. This search was also expected to provide the most literature, since it was designed to be very general in relation to the combination of anesthesia and swine. This can also be seen in the block searches of table A.1 and table A.2, which contained terms related to anesthesia, e.g. "Consciousness Monitors" for Pubmed, as "OR" boolean operators, thus allowing a larger number of results to be found in any cases where the literature was catalogued under "Consciousness Monitors" instead of "Anesthesia, general" or both. Of importance was also that the terms anesthesia were limited to general anesthesia, as originally a large amount of literature was found unrelated to the purpose of this project, as they often had to do with local anesthesia. This had not been a problem, if the terms like "Consciousness Monitors" and "Anesthesia" were separated by "AND" boolean operators, but since this was not the case for the reason explained before, the term anesthesia was limited to general anesthesia only.

This search was limited specifically to literature like articles, trials, studies, reviews, meta-analysis and alike. This was done to cover the most relevant information and the search was not limited much to publication types, since this search was expected to be the largest and provide the most information. Though, some formats of literature were deemed unwanted, e.g. preprints, letters and conference material. The terms for ECoG were not used, but

were substituted with ECG instead. Although a conventional ECG was not used in this project but ECoG was, their behaviour in literature was expected to be similar with the exception of ECoG being a cleaner signal.

Pubmed	OR	
AND	Anesthesia	"Anesthesia, General"[Mesh] "Intraoperative Awareness"[Mesh] "Consciousness Monitors"[Mesh] "Monitoring, Intraoperative"[Mesh]
	Parameters	"Electroencephalography"[Mesh] "Electrocardiography"[Mesh]
	Species	"Swine"[Mesh]

Table A.1.

Embase	OR	
AND	Anesthesia	'general anesthesia'/exp 'intraoperative awareness'/exp 'consciousness monitor'/exp 'intraoperative monitoring'/exp
	Parameters	'electroencephalography'/exp 'electrocardiography'/exp
	Species	pig'/exp

Table A.2.

A.2 Anesthesia review-search

The second structured search was the General Anesthesia-search, whose purpose was to obtain information related to the monitoring of anesthesia in general, thus not only in swine. No specific species was set, but the majority of results was expected to be humans, which happened to be true. Results related to humans were included, because the biology of pigs is sufficiently similar to humans, although important interspecies differences still exist [Kobeissy et al., 2016; Walters et al., 2017]. This search was expected to have too many results, because the field of anesthesia related to humans is very large. For this reason, the search was limited to reviews, systematic reviews and meta-analyses in an attempt to obtain information about the most important trends related to anesthesia, e.g. new methods that were proving very successful in humans, and avoid less well-tested meticulous details from standalone studies. These details could be useful too, although the immense amount of literature of general anesthesia in humans alone would be overwhelming, and thus this search focused on summaries of the literature.

As with the Swine/Anesthesia-related search, the anesthesia was limited to general anesthesia. For this search, the term of anesthesia and the terms related more to monitoring of anesthesia were split into two separate "AND" blocks. This was done to minimize the number of results and obtain literature more specific to the use case of monitoring the depth of anesthesia instead of anesthesia in general. Additionally, the monitoring and parameter terms were also split into two new "AND" blocks to utilize the "Major topic" function of Pubmed and Embase. Thus, with the new arrangement as seen in table A.3

and table A.4, the search focus was intensified on literature catalogued very thoroughly with a must-have inclusion of many of these terms and with involvement of both ECG and EEG, where one must be a major topic.

Pubmed	OR	
AND	Major topic: Monitoring	"Consciousness Monitors"[Majr] "Monitoring, Intraoperative"[Majr]
	Monitoring	"Consciousness Monitors"[Mesh] "Monitoring, Intraoperative"[Mesh]
	Major topic: Anesthesia	"Anesthesia, General"[Majr]
	Anesthesia	"Anesthesia, General"[Mesh]
	Major topic: Parameters	"Electroencephalography"[Majr] "Electrocardiography"[Majr]
	Parameters	"Electroencephalography"[Mesh] "Electrocardiography"[Mesh]

Table A.3.

Embase	OR	
AND	Major topic: Monitoring	'consciousness monitor'/exp/mj 'intraoperative monitoring'/exp/mj
	Monitoring	'consciousness monitor'/exp 'intraoperative monitoring'/exp
	Major topic: Anesthesia	'general anesthesia'/exp/mj
	Anesthesia	'general anesthesia'/exp
	Major topic: Parameters	'electroencephalography'/exp/mj 'electrocardiography'/exp/mj
	Parameters	'electroencephalography'/exp 'electrocardiography'/exp

Table A.4.

A.3 Neural network/Anesthesia-search

The third structured was the Neural network/Supervised learning/Anesthesia-search, whose purpose was to obtain information related to using machine-learning to guide decisions of anesthesiology. It was supposed to focus on the use of neural networks, but also included supervised machine learning, since one of the most important things in neural networks is the use of labels to indicate the truth and labels from non-neural network methods could still be useful. Thus, other supervised machine learning methods might provide useful information regarding what kind of truth was being used and how successful its use was. Contrary to the other searches, this search only looked at material from 2016-2021 because it provided a more focused perspective on the newest research, and because a larger frame of time provided too many results. It was possible to limit this search to reviews and meta-analyses as in the Anesthesia Review-search to compensate for the amount of results, though it is common in the field of neural networks for individual papers to contribute significantly by small important details in the neural networks. Furthermore, specific elements i the field of neural network have a harder time establishing trends because of the development process, and thus obtaining only reviews and meta-analyses would be

undesired. For these reasons, the publication time was decreased to 2016-2021 instead. The search blocks can be seen in table A.5 and table A.6.

Pubmed	OR	
AND	Machine learning	"Neural Networks, Computer"[Mesh] "Supervised Machine learning"[Mesh]
	Monitoring	"Consciousness Monitors"[Mesh] "Intraoperative Awareness"[Mesh] "Anesthesia, General"[Mesh] "Monitoring, Intraoperative"[Mesh]

Table A.5.

Embase	OR	
AND	Machine learning	'artificial neural network'/exp 'supervised machine learning'/exp
	Monitoring	'consciousness monitor'/exp 'general anesthesia'/exp 'intraoperative monitoring'/exp 'intraoperative awareness'/exp

Table A.6.