Evaluation of Machine Learning Tools for Predicting Sand Production

A thesis submitted to the faculty at African University of Science and Technology in partial fulfilment of the requirements for the degree of Master of Science in the Department of Petroleum Engineering

By

NGWASHI Ronald Afungchwi

Supervised by

Professor David Ogbe

African University of Science and Technology

www.aust.edu.ng

P.M.B 681, Garki, Abuja F.C.T, Nigeria

Abuja, Nigeria

March, 2021

CERTIFICATION

This is to certify that the thesis titled "EVALUATION OF MACHINE LEARNING TOOLS FOR PREDICTING SAND PRODUCTION" submitted to the school of postgraduate studies, African University of Science and Technology (AUST), Abuja, Nigeria for the Award of the Master's degree is a record of original research carried out by NGWASHI RONALD AFUNGCHWI in the Department of Petroleum Engineering.

EVALUATION OF MACHINE LEARNING TOOLS FOR PREDICTING SAND PRODUCTION

By

Ngwashi Ronald Afungchwi

A THESIS APPROVED BY THE PETROLEUM ENGINEERING DEPARTMENT

RECOMMENDED: .....

Supervisor, Prof. David Ogbe

.....

Committee Member, Dr. Alpheus Igbokoyi

.....

Head, Department of Petroleum Engineering

APPROVED: .....

Chief Academic Officer

.....

Date

COPYRIGHT

ABSTRACT

Data analytics has only recently picked the interest of the oil and gas industry as it has made data visualization much simpler, faster, and cost-effective. This is driven by the promising innovative techniques in developing artificial intelligence and machine-learning tools to provide sustainable solutions to ever-increasing problems of the petroleum industry activities. Sand production is one of these challenges faced by the oil and gas industry. Understanding whether a well will produce sand or not is the foundation of every completion job in sandstone formations. The Niger Delta Province is a region characterized by friable and unconsolidated sandstones; therefore, it is more prone to sanding. It is economically unattractive in this region to design sand equipment for a well that will not produce sand.

This study is aimed at developing a fast and more accurate machine-learning algorithm to predict sanding in sandstone formations. A two-layered Artificial Neural Network (ANN) with back-propagation algorithm was developed using PYTHON programming language. The algorithm uses 11 geological and reservoir parameters that are associated with the onset of sanding. These parameters include depth, overburden, pore pressure, maximum and minimum horizontal stresses, well azimuth, well inclination, Poisson's ratio, Young's Modulus, friction angle, and shale content. Data typical of the Niger Delta were collected to validate the algorithm. The data was further split into a training set (70%) and a test set (30%). Statistical analyses of the data yielded correlations between the parameters and were plotted for better visualization.

The accuracy of the ANN algorithm is found to depend on the number of parameters, number of epochs, and the size of the data set. For a completion engineer, the answer to the question of whether or not a well will require sand production control is binary-either a well will produce sand or it does not. Support vector machines (SVM) are known to be better suited as the machine-

learning tools for binary identification. This study also presents a comparative analysis between ANN and SVM models as tools for predicting sand production. Analysis of the Niger Delta data set indicated that SVM outperformed ANN model even when the training data set is sparse. Using the 30% test set, ANN gives an accuracy, precision, recall, and F1- Score of about 80% while the SVM performance was 100% for the four metrics. It is then concluded that machine learning tools such as ANN with back-propagation and SVM are very simple, accurate, and easy-to-use tools for effectively predicting sand production.

Keywords: Machine Learning, Artificial Neural Networks, Support vector Machines, Data Analytics, Python programming, Sand production

DEDICATION

I dedicate this work to the Almighty God for giving me the strength to take one day at a time, to the African University of Science and Technology (AUST) community, for giving showing me love and companionship, to my family and friends for guiding me spiritually, morally and financially.

Acknowledgements

The completion of this work would not be possible if it weren't the Grace of God Almighty. Furthermore, this work would have been impossible to complete if it weren't for the encouragement, positive comments and quick response from the people around me. My gratitude goes to my supervisor and mentor, Prof. David Ogbe for always pushing me to test my limits and emerge a better person. Also, special thanks go to my committee member and Head of Department, Dr. Alpheus Igbokoyi for taking the time to go through this work and guiding me throughout my stay at AUST. Special gratitude goes to my amazing friends in AUST especially my roommate Kinnouhezan Marius Dona, my brother from Benin Republic, Azah Daizy-Clay Che, my greatest motivator and rock, Kawu Musa Idris Idah, Kanfoui Felicienne Djossou, and Opeyemi Ogunkuade, whose smile and fun nature kept AUST tolerable, and most importantly, Chiamaka Mary Igwilo, my greatest fan.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1. General Overview

The application of data analytic tools in the oil and gas industry is rapidly growing. Applying machine learning (ML) and artificial intelligence (AI) tools for solving problems has been shown to be cost effective, time saving, and efficient. Sand production is one of the challenges faced by well completion engineers in developing sandstone reservoirs.

Sand failures may be due to poor completion practices or completions in unconsolidated reservoirs. It is known that formations may be susceptible to sand production when in-situ rock strength is reduced by erroneous completion, poor production practices, or improper fluid production rate. Several problems may occur in the life of the production system as results of sand production, including abrasion of downhole tubular or casing, subsurface safety valve and surface equipment, casing or tubing buckling, failure of casing or liners from the removal of surrounding formation, compaction, and erosion. This leads to losses in oil production because of sand bridging in the tubular and or flowlines (Osisanya, 2010). The cost of handling and disposal of produced sand is extremely high especially in offshore platforms. Therefore, it is vital to predict sand production in order to design the proper completion system as a proactive solution to this problem. Machine learning and AI algorithms have been used to develop promising tools in identifying, classifying, or predicting similar problems in the oil and gas industry. There is the need to know whether a well will produce sand or not; what tools to use to best predict the onset of sanding; and how effective the tool is.

The premise of this research is outlined in the following section as the research problem is identified, the research question posed, and the objectives are clearly written. Furthermore, the

significance of this study is discussed as well as the scope of the study. The utility of this work to not only the industry but the research community, is clearly identified.

1.2. Problem Statement

The following sanding problems often occur during the life of a well especially in sandstone formations. Firstly, abrasion of downhole tubular or casing, subsurface safety valve and surface equipment, casing or tubing buckling, failure of casing or liners from the removal of surrounding formation, compaction and erosion, and its loss in production caused by sand bridging in tubular and or flowlines. Secondly, cost of handling and disposal of produced sand can be significant especially in offshore platforms. It is therefore crucial to predict sand production with accuracy to design the fit-for-purpose completion system as a proactive solution to these problems.

1.3. Research Question

Machine Learning (ML) and Artificial Intelligence (AI) algorithms have been used to derive promising results in identifying, classifying, or predicting such problems in the E&P Industry (Olatunji and Micheal, 2017). So, the following questions are posed for the purpose of this research:

1. In a given field, will Well X produce sand?
2. Is it possible to predict the sanding in Well X?
3. What algorithm is best for predicting the problem of sanding and why?
4. How effective is this algorithm?

This study aims at answering the above research questions. The approach is to use ML and AI tools to solve the problem of predicting sand production in sandstone formations.

1.4 Objectives

To understand and predict sand production, this project aims at achieving the following objectives:

- To conduct a thorough review of Machine Learning (ML) and Artificial Intelligence (AI) tools used in petroleum engineering especially those related to sand production studied in this research.
- To review causes, effects, and techniques for predicting sanding in sandstone reservoirs
- To use PYTHON programming language to develop the machine learning algorithms to predict sand production; specifically, Artificial Neural Network (ANN) and Support Vector Machines (SVM) algorithms.
- To use a case study to validate the models built with the ANN and SVM algorithms.
- To perform a comparative study to evaluate the performance and efficiency of both algorithms in predicting sand production based on the data from the case study.

1.5. Significance of Study

The expected outcome of this research is a tool for predicting sand production in sandstone formations. The application of this study includes the fields of reservoir management, well completion design, well perforation design, sand monitoring strategy, design of surface facilities and pipelines, and analysis of field development economics. In addition, the research is designed to contribute to our knowledge of sand production, and the need to handle it rapidly, efficiently, and cost effectively before the deployment of any well completion job. For this work, ML and AI approaches are proposed to provide a proactive and cheap solution to the problem. Also, this work

aims to contribute to the field of data analytics in the E&P industry, which is rapidly becoming an important topic for engineer and geoscientists. Moreover, this work will pave a way to concretize the fusion of modern computing and engineering skills to solve real world problems in the E&P industry. It should serve as a guide to students and other professionals on how to transform traditional methods into less time-consuming and effective tools for getting the job done.

To achieve these objectives, we evaluate the application of machine learning tools for predicting the onset of sanding. Two algorithms, Artificial Neural Network (ANN) with back-propagation and Support Vector Machine (SVM) are developed using PYTHON programming language. A comparative study is conducted to evaluate the performance of the models based on the two algorithms. Data from the Niger Delta is used to validate the performance of the models.

This work is presented in five Chapters. First, Chapter 1 contains the introduction to this study and a theoretical background is given in Chapter 2 to improve our understanding of sanding, prediction methods, as well as the application of machine learning and artificial intelligence discussed in this work. In the third Chapter is a description of the methodology used in developing the two models based on the ANN and SVM algorithms. Then, the results are analyzed and comparative performance of the two models are discussed in the fourth Chapter. Finally, the major observations and conclusions derived from this study as well as a set of recommendations are presented in Chapter 5.

Chapter 2: Theoretical Background of Sand Production and Machine Learning

2.1. Literature Review

In order to understand more of sanding and machine learning, it is important to review some of the work done by past researchers on this topic. Udebhulu (2014) reviewed in detail, various methodologies on predicting sanding. Most of these require knowing certain reservoir and well data such as the reservoir depth, flow rate; formation cementation; compressibility and natural permeability; surface exposed to flow (interval length, open versus plugged perforations); produced fluid types and phases; formation sand characteristics (angularity and shaliness); pressure drawdown; impairment of natural permeability; and reservoir pressure.

Also, Osisanya (2010) came up with methodologies for predicting sand production which required the following: reservoir depth; flowrate; formation cementation; compressibility and natural permeability; surface exposed to flow (interval length, open versus plugged perforations); produced fluid types and phases; formation sand characteristics (angularity and shaliness); pressure drawdown; impairment of natural permeability; and reservoir pressure. He went further to classify certain factors that can influence a well's tendency to produce sand. (Osisanya, 2010).

Until recently, five methods have been known for predicting sand production. These include, production rate methods, data from well logs, laboratory testing method, reservoir formation classification, and the analogy method. These methods have some flaws which rendered them inapplicable in some cases, less accurate to predicting sanding, and more costly among other handicaps. Production rate method requires rates from test wells or other wells in the same area as

the well under investigation. Factors like reservoir heterogeneity limit the accuracy of the results of rate gotten from this method. Even though the use of well logs was considered the most effective means of determining in-situ mechanical properties, the cost of logging a lot of wells, and the accuracy of logs from older wells limit the application. Laboratory testing methods and the reservoir formation classification method depend on obtaining well logs and using correlations which were also approximations with their own errors. The analogy method for predicting sand production is based on information obtained from other wells. This method is useful only in virgin fields with limited data on sand production (Osisanya, 2010).

Isehunwa et al. (2017) carried out a research on sand production prediction in gas and gas condensate wells. They laid out the notion that gas and gas condensate reservoirs are not believed to be prone to sand production. So, they used the concept of erosional failure mechanism to come up with a model which predicted sand production from gas and gas condensate wells. This model showed that sand production was as a result to certain factors which were: flowrate, fluid density and viscosity, density of sand, particle size, and borehole radius. He further concludes that production rates above 10,000 MSCF/day could lead to cavity heights between 30ft to 60ft. moreover, high production rates in gas condensates wells lead to increase in cavity arch radius and hence, sand production (S. et al., 2017).

Udebhulu (2014) developed a mechanistic approach for predicting sand production which incorporated the concept of dimensionless quantities associated with sanding. These quantities include loading factor, Reynold's number, water cut and gas-liquid ratio (GLR). This study showed that every reservoir has a unique sand production rate (SPR) correlation index which represents its propensity to producing sand (Udebhulu & Ogbe, 2015). The data required for prediction sand production include: production data, formation intrinsic strength, dynamic elastic

constants (Shear modulus, bulk modulus, Young's modulus, Poisson's ratio), and log data (sonic and bulk density). It should also be noted that failure, post-failure sand mechanics and flow-dominated erosion mechanisms are important in sand production processes (Udebhulu & Ogbe, 2015).

In recent years, the necessity of answering the question of sanding in a cost-effective manner has fostered the implementation of computer algorithms, Machine Learning and Artificial Intelligence (AI) tools.

With the evolution of technology in E&P, companies have ventured into the use of machine learning and AI (Artificial intelligence) to model and simulate reoccurring problems. Sand production is not exempted from such innovations. The concept of machine learning, artificial intelligence, computational neuroscience and many other advanced fields have provided very robust, efficient, effective platforms for solving perennial problems in today's world (Olantunji & Micheal, 2017). The question of the effectiveness of machine learning to predict sand production for proactive mitigation has always been an issue and is important in the current discussion. In 1999 Kanj and Abousleiman (1999) modeled, for the first time, the northern Adriatic basin using artificial neural networks (ANN) to predict important sanding indication parameters for gas wells. Using this machine learning algorithm, they showed that the predictions from ANN gave better results compared to those derived from analytical models (Kanj and Abousleiman, 1999).

M. Azad et al (2011), used models to develop an approach for predicting sand production using artificial neural networks. The implementation of ANN was tailored to predicting bottom hole flowing pressures which inhibited sand production. After using data from 38 wells from 3 oil fields producing from the same source rock, they developed an exact method for predicting sand initiation with a high degree of accuracy. For simplicity, they used two of the three factors

considered to have a greater impact on sanding. These were formation strength characteristics, and production and reservoir characteristics. Furthermore, the 38 wells used in the study were already known to be sand-prone, and Azad et al. used the critical drawdown pressure to determine corrections for sand production (Azad et al, 2011).

Khamehchi et al (2014), utilized a total of 23 field datasets collected from problematic wells from the North of the Adriatic Sea to perform a simple regression to recognize the statistically important parameters, performed multiple linear regression (MLR) and generic algorithm MLR (GA-MLR) to develop an estimation of for the critical total drawdown (CTD). Then, they utilized artificial neural networks (ANN) with back propagation (BP) and particle swam optimization (PSO) algorithm to correlate CTD to all the other parameters. This study proved that artificial intelligence or machine learning is particularly useful in efficiently predicting sand production and minimizing uncertainties (Khamehchi et al, 2014).

Rahmati et al (2013) reviewed various sand production prediction models. Their premise was based on knowing that sand production could occur in oil and gas wells if the fluid flow exceeds a certain threshold governed by factors such as stress state, and the completion type used around the wellbore, as well as inconsistences of the reservoir rock. They showed that a little quantity of solids could pose substantial risk to the well in the long run. They looked into some models which could only predict whether sand production is possible or not by using analytical formulae, and some models which could make volumetric predictions by utilizing a numerical approach in calculating the sanding rate. Their results showed major improvements in predicting sand control, but issues such as the ability of critical-state-based constitutive models in providing more realistic representations of sand production, amoagst a host of others, were explained in the paper. (Rahmati et al, 2013).

In 2017, Olatunji and Michael used a more advanced machine learning algorithm—Support Vector Machines (SVM) to develop a new model to identify the sanding prevalent in the Niger Delta Basin. By using SVM, they were able to determine whether sand will be produced from a well or not (Olatunji and Micheal, 2017).

To provide better understanding of the research work, the following machine learning and AI concepts are discussed. Detailed discussion of Machine Learning can be found in several publications (Smith & Frank, 2016; Geron, 2019; Hamid Rahmanifard, 2018; and Agatonovic-Kustrin & Beresford, 2000). Also, detailed understanding of sanding, causes of sanding and predicting methods can be found in the following works (William and Joe, 2003; Completion tech., 1995; Qui et al., 2006; Veeken et al., 1991; Bellarby, 2009); Khaksar et al., 2009, and Navjeet, 2004). The following section is a summary of the work of these researchers on sand production, causes, methods of prediction of sanding, and ML concepts, theories, tools, and their application specifically to predicting sanding in the E&P industry.

2.2. An Overview of Sand Production

Sanding is a very destructive problem in the life of a well in sandstone formations. The level of destructiveness ranges from blocked subsurface equipment, to completely damaging the well. The following are the various effects of sanding.

1. Erosion of surface and sub-surface equipment: Produced sand travels at high velocity through the well and production equipment. This high velocity sand causes abrasion of the walls of the equipment leading to complete damage or frequent maintenance of these equipment. Potential sites for abrasion include blasts joints, tubing opposite perforations, screens or slotted liners not packed in the gravel pack installations (William & Joe, 2003).

Human and environmental hazards can arise if the erosion is severe and not properly checked over time.

2. Formation subsistence: The cumulative production of sand can cause caving-in and collapse of the formation. With time, large quantities of sand are produced at the surface creating a void behind the casing. As more sand is produced, the void widens leading to collapse of formation sand or shale above the void (Completion tech., 1995). This formation collapse causes the sand grains to rearrange, leading to reduction in permeability. The reduction in permeability often at times results in the loss of productivity. The collapse of the formation is particularly important if the formation material fills or partially fills the perforation tunnels. Even a small amount of formation material filling the perforation tunnels will lead to a significant increase in pressure drop across the formation near the well bore for a given flow rate (Completion tech., 1995).
3. Sand accumulation in surface equipment: In some situations, the production velocity of the reservoir fluid is sufficient to carry sand up the tubing to the surface. Sand particles often settle in surface facilities as separators, heaters, pumps, condensers (Completion tech., 1995). As the accumulation builds to appreciable volume in these facilities, equipment clean-up becomes inevitable. This causes deferred production (well is shut-in) and additional cost is incurred as a result of the sand clean-up activity (Completion tech., 1995). Production capacity of the separator is reduced if partially filled with sand. This is as a result of its reduced ability to handle gas, oil and water (Completion tech., 1995).
4. Subsurface accumulation: This occurs when the production flow velocity is not sufficient to carry the sand particles to the surface. The sand accumulates in the casing or bridges off in the tubing, and with time the production interval might be filled with sand. This reduces

the production rate for such wells which might eventually cease as the sand accumulation makes it impossible for production to continue (Completion tech., 1995). Work-over activities are often required in such occurrences for the well to resume production. If sand production is continuous, the well clean-out operations may be required regularly. This causes increased maintenance cost and loss of production which in turn reduces economic returns from the well (Completion tech., 1995).

5. Sand disposal: This constitutes a problem in formations producing sand especially in areas where there are stringent environmental constraints. For offshore processing systems that do not satisfy anti-pollution regulation of the separated sand, the produced sand must be transported onshore for disposal, constituting additional production cost.

2.3. Causes of Sand Production

Factors promoting the tendency of a well to produce sand can be categorized into rock strength effects and fluid flow effects (Completion tech., 1995). Production of sand particles consists of formation fines and load bearing solids. The production of formation fines which is not considered as part of the formation's mechanical framework may be beneficiary as they can move freely through the formation instead of plugging it. Production rates are often kept to low levels to avoid the production of the load bearing particles, and in many cases, however, low production rates are uneconomical. The factors that affect sand production in a well include:

1. Degree of consolidation: The ability to maintain open perforation tunnels is closely tied to how strongly the individual sand grains are bound together. The cementation of sandstone is typically a secondary geological process, and as a general rule, older sediments tend to be more consolidated than newer sediments. This indicates that sand production is normally a

problem when producing from shallow, geologically younger tertiary sedimentary formations. Such formations are in the Gulf of Mexico, California, Nigeria, France, Venezuela, Trinidad, Egypt, Italy, China, Malaysia, Brunei, Indonesia and others. Young tertiary formations often have little matrix material (cementation material) bonding the sand grains together and these formations are generally referred to as being "poorly consolidated" or "unconsolidated". A mechanical characteristic of rock that is related to the degree of consolidation is called "compressive strength". Poorly consolidated sandstone formations usually have a compressive strength that is less than 1,000 pounds per square inch (Completion tech., 1995).

2. Production rate: Increasing the well production rate creates large fluid pressure gradient near the wellbore (perforation) which tends to draw sand into the wellbore. Generally, production of the reservoir fluids creates pressure differential and frictional drag forces that can combine to exceed the formation compressive strength. This indicates that there is a critical flow rate for most wells below which pressure differential and frictional drag forces are not great enough to exceed the formation compressive strength and cause sand production. The critical flow rate of a well may be determined by slowly increasing the production rate until sand production is detected. One technique used to minimize the production of sand is to choke the well flow rate down to the critical flow rate where sand production does not occur or has an acceptable level. In many cases, this flow rate is significantly below the acceptable production rate of the well (Completion tech., 1995).
3. Pore pressure reduction: Reservoir fluid production overtime depletes the reservoir pressure resulting in pore pressure reduction. As the reservoir pressure is depleted throughout the producing life of a well, some of the support for the overlying rock is removed. Lowering

the reservoir pressure creates an increasing amount of stress on the formation sand itself, i.e., the effective overburden pressure increases (Completion tech., 1995). The formation sand particles may be crushed or break loose from its matrix at some time in the reservoir life which could be produced along with the reservoir fluids. The formation might subside if the effective stress exceeds the formation strength due to compaction of reservoir rock from reduction in pore pressure.

4. Reservoir fluid velocity: The frictional drag force exerted on the formation sand grains is created by the flow of reservoir fluid. This frictional drag force is directly related to the velocity of fluid flow and the viscosity of the reservoir fluid being produced. High reservoir fluid viscosity will apply a greater frictional drag force to the formation sand grains than will a reservoir fluid with a low viscosity. The influence of viscous drag causes sand to be produced from heavy oil reservoirs which contain low API gravity, high viscosity oils even at low flow velocities (Completion tech., 1995).
5. Increasing water production: Increase in water cut increases sand production or as water production begins sand production begins too. These occurrences can be explained by two mechanisms. In a typical water-wet sandstone formation, some grain-to-grain cohesiveness is provided by the surface tension of the connate water surrounding each sand grain. At the onset of water production, the connate water tends to adhere to the water produced, resulting in a reduction of the surface tension forces and subsequent reduction in the grain-to-grain cohesiveness. The stability of the sand arch around the perforation has been shown to be limited greatly by the production of water resulting in the production of sand. Figure 2.1 taken from Completions tech (1995) illustrates this situation. An arch is a hemispherical cap of interlocking sand grains that is stable at constant drawdown and flow rate preventing sand

production (Jon Carlson et al., 1992). A second mechanism by which water production affects sand production is related to the effects of relative permeability. As the water cut increases, the relative permeability to oil decreases. This result in an increasing pressure differential being required to produce oil at the same rate. An increase in pressure differential near the wellbore creates a greater shear force across the formation sand grains. Once again, the higher stresses can lead to instability of the sand arch around each perforation and subsequent sand production (Completion tech., 1995).

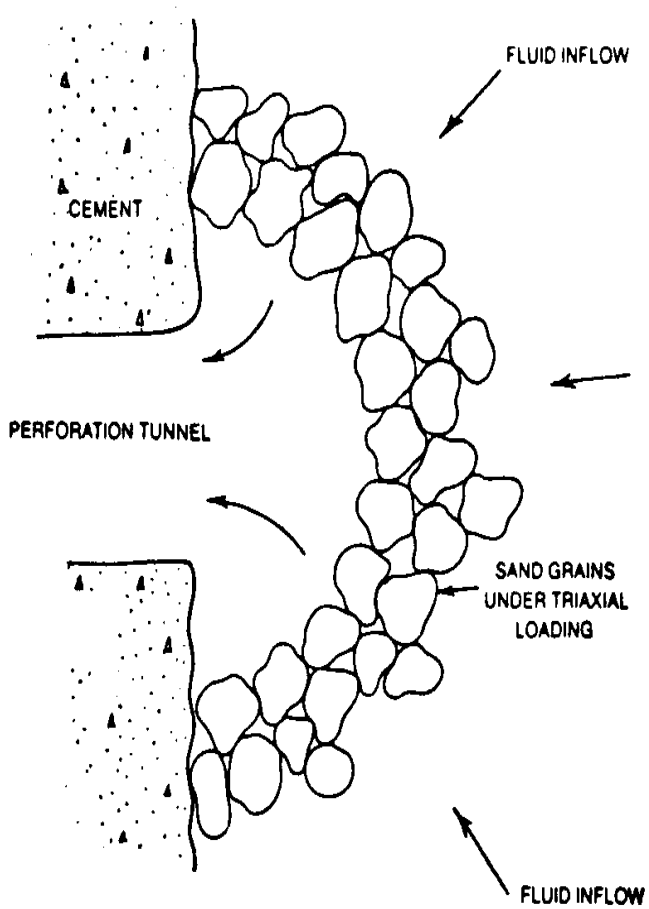


Figure 2.1: Geometry of a stable arch surrounding a perforation (Source: Completion tech., 1995)

2.4. Prediction of Sand Production

It is important for the completion engineer to know under what conditions a well produce sand to predict if the well will require a method of sand control. Sand prediction is usually done at the initial stage of reservoir development. It involves development of completion design, reservoir management strategy, perforation strategy, sand monitoring strategy, planning of the surface facilities and field economics. This task is not an easy one as the process of sand prediction is more of art than a science. At best performances of nearby offset wells are observed or the well is completed conventionally and flowed to observe if sand production will occur. The many published techniques to predict the onset of sanding can be categorized into four basic approaches: Empirical methods using field observations and well data, Laboratory simulation, Numerical methods and Analytical methods (Qui et al, 2006). Often two or more techniques are used in combination for prediction.

2.4.1. Empirical Methods Using Field Observations and Well Data

This technique uses a correlation between sand production well data and field operational parameters in prediction. Typically, one or a group of parameters are used to evaluate the sanding potential and to establish a benchmark for sanding or no sanding. This is due to the practical difficulties of monitoring and recording several years' worth of data for all the wells involved in a study (Veeken et al., 1991). Parameters such as porosity, drawdown or flow rate, compressional slowness etc. are often used. Veeken et al., (1991) presented a list of the parameters that may influence sand production. These include:

1. Formation
    a. Rock (strength, vertical and horizontal in-situ stresses, depth)

b. Reservoir (far field pore pressure, permeability, fluid composition, drainage radius, reservoir thickness, heterogeneity)

2. Completion (wellbore orientation, wellbore diameter, completion type, perforation policy, sand control, completion fluids, stimulation, tubular size).

3. Production (flow rate, drawdown pressure, flow velocity, damage, build-up / shut-in policy, artificial lift technique, depletion, water / gas coning, cumulative sand volume).

In its simplest form, the field data-based sand prediction tool uses only one parameter. Examples include avoiding porosities higher than 30% (Bellarby, 2009), using a cut-off depth criterion for the installation of sand control measures in several deltaic environments: sand control is not installed below a certain depth. A depth of 12,000ft and 7,000ft were mentioned by Tixier et al and Lantz et al respectively. This critical depth is regionally dependent. Another example is applying a compressional sonic wave transit time **(Δtc)** below which sand control is not required; the limit Δ**tc** is again field or regionally dependent and may vary from 90 to 120μs/ft (Veeken et al, 1991). Tixier et al., 1975 derived a log-based technique using mechanical properties log to predict sanding. A limit value for the sonic and density log derived parameter ratio of G (the dynamic shear modulus) to C_b the bulk compressibility i.e. (G/C_b) was established. When G/C_b exceeds $0.8 * 10$ psi, no sanding problem is expected. At ratios less than $0.7 * 10$ psi sand influx will occur. These mechanical properties log method works 81% of the time (Osisanya, 2010) but seems to be dependent on regional environment too. The one parameter method is practical but conservative. The two parameters method considers the depletion of the reservoir pressure (ΔP_{de}) and the drawdown pressure (ΔP_{dd}) not accounted for in the one parameter model. Stein et al., (1972) provided a method to estimate the maximum production sand free rate from density and acoustic velocity log data by relating drawdown to the dynamic shear modulus, E_s. Data from wells

producing sand were used to relate to new wells.

On the basis of data from many fields Veeken et al., (1991) plotted the total drawdown pressure, ($\Delta\Delta Ptd$ versus sonic transit time, Δt_c, for sand and no-producing sand wells. From the plot shown in Figure 2, a risk region possible to produce sand was established. To the left of the region, sand-free production can be realistically expected. It was also inferred that increasing total drawdown may trigger sand production. The position of the risk region is field dependent and its position can be determined from sand production tests or routine monitoring.

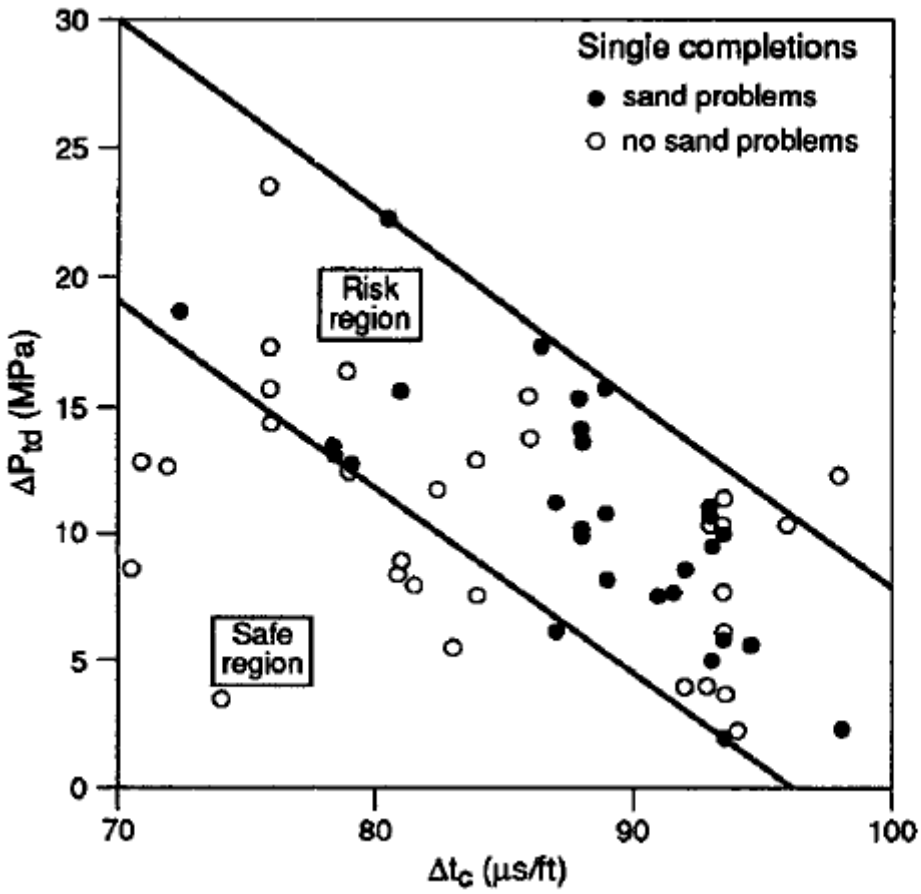


Figure 2.1: Total drawdown versus transit time for intervals with and without sand problem (Source: Completion tech., 1995).

2.4.2. Laboratory Simulation

This approach is also used widely to establish correlation between the risk of sanding and

measurable parameters like stress, flow rate and rock strength and to develop an insight into the mechanism of sanding in the formation involved. Laboratory experiments involve the use of available reservoir core samples or outcrop rock samples (with similar mechanical properties). Two types of laboratory sand production are common: laboratory sand production experiments and hollow cylinder collapse tests (Qui et al, 2006). Typically, laboratory experiments formulate sand production phenomenon in a controlled environment. Laboratory sand production test involves the use of cores to produce a small-scale simulation of flow through perforations or cylindrical cavities contained within a stressed cylindrical core sample. The technique offers the investigation of factors such as drawdowns, stress boundary conditions, flow rates, water cuts and rock properties. Expected conditions during the producing life of the well can be chosen as test parameters. This method is widely used to calibrate and validate predictions from analytical and numerical models. However, considerable number of cores and well-equipped facilities are needed for the test.

Thick wall cylinder tests (TWC) are also used for sanding evaluation and calibration, easier to perform than sand production test. In this test a hollow cylindrical core plug is loaded axially and laterally under increasing hydrostatic stress ($\sigma 1=\sigma 2=\sigma 3$) until collapse occurs in the walls of the cylinder. The hydrostatic stress at which failure initiates in the internal wall is reported as the TWC-internal and the stress that causes external wall failure is called TWC-external or TWC-collapse. The external wall catastrophic failure pressure corresponds to the perforation failure condition that causes continuous and catastrophic sand production. The internal wall failure pressure is less than the catastrophic failure and normally corresponds to the onset of transient sanding. TWC internal can be defined by an increase in fluid volume expelled during constant loading or by monitoring and measuring the internal hole deformation during tests using internal

gauged or camera. However, such measures require large plug sizes which are not routinely available (Khaksar et al, 2009). BP reports using plugs that have a 1.5 in. outside diameter (OD), a 0.5 in. internal diameter (ID) and are 3 in. long (Willson et al., 2002), whereas Shell use plugs that have a 1 in. OD, 0.33 in. ID and are 2 in. long (Veeken et al., 1991), (Bellarby, 2009). Results from TWC test can be used to predict the depths and conditions at which sanding might occur in the field, if the stresses corresponding to failure are considered representative of stresses at the sand-face or perforation cavity. Veeken et al, (1991) gave a relationship between the near-wellbore vertical effective stress ($\sigma_{v,w}$) and the TWC collapse pressure (σ_{twc}) from many experiments carried out on friable-consolidated sandstone.

$$\sigma_{v,w} = 0.86 \times \sigma_{twc} \qquad [2.1]$$

The results from TWC can however be influenced by sample size/hole size ratio of the hollow cylinder.

Figure 2.2: TWC machine (Source: Bellarby, 2009)

From laboratory experiments, important findings as stresses and rock strength are dominant factors controlling sanding initiation and sand production, flowrate only plays a role in weak and unconsolidated rocks and rocks under excessive stresses, increase in drawdown causes sand production increase, due to changes in boundary conditions (i.e., stresses of fluid flowrate) bursts of sand production are frequently observed after which sand production may gradually decline to some background, there are significant nonlinear scale effects related to the size of a perforation or open hole and their stability against sanding, with smaller diameter cavities being most stable

(Qui et al, 2006) have been reached. In the 1970s, Exxon conducted an experiment to establish the relationship between the rock compressive strength and sand production potential of the rock. The studies revealed that the rock failed and began sand production when the fluid flow stresses exceeded the formation compressive strength. As a rule of thumb from the research, sand production or rock failure will occur when the drawdown pressure is 1.7 times the compressive strength. Figure 4 shows the equipment used in the test to determine the magnitude of the pressure

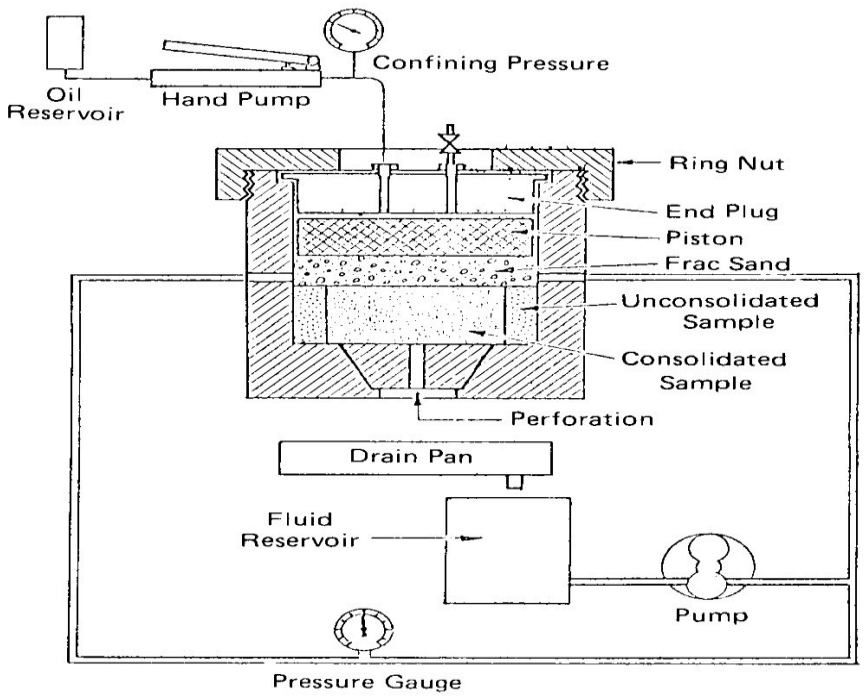


Figure 2.3: Exxon Equipment for Drawdown-to-Rock Failure Test

drops that core samples could withstand before sand production starts. This relationship holds for consolidated formations. Non-destructive test like impact and scratch test are also used for measuring the strength properties of a rock.

The main disadvantage of this approach is the amount and availability of core samples needed, time and cost for preparing the core, conducting the experiments, processing and analyzing the

data from the test. A question of how well a laboratory simulation represents on field scenarios is also raised.

2.4.3. Analytical Methods

This method has gained more popularity in the petroleum industry due to its computational simplicity, readily implementable calculations, and the ease of running multiple realizations to compare many different scenarios. Analytical sand prediction models are based on modeling of perforations and production cavity stability. This tool requires a mathematical formulation of the sand failure mechanism. Production cavity stability under producing conditions is related to the stresses imposed on the formation matrix and the complex manner in which the matrix accommodates these stresses. Stresses imposed are due to overburden pressure, pore pressure, flowing fluid pressure gradient near the wellbore, interfacial tension effects and viscous drag forces. At the mechanical failure of the load bearing sand grain matrix, sand is assumed to be produced. Prediction accuracy depends more on how the rock constitutive behaviour is modelled, the failure criterion chosen and whether the materials and other parameters affecting the rock failure are determined precisely. Moore, (1994) highlighted some engineering and geologic parameters to be considered in a complete evaluation of the sand production potential of a formation based on different sand prediction models and techniques available in the industry. These include:

1. Field data
2. Cyclic loading
3. Directional in-situ stresses
4. Quality of cementation
5. Perforation geometry and spacing

6. Perforation cavities geometry and shot density

7. Cavity evolution effect of varying perforation geometry

8. Well pressure

9. Flow rate (fluid forces)

10. Permeability, viscosity, and relative permeability for two and three phase flow

11. Rock deformation characteristics

12. Rock strength characteristics

13. Flow through porous media where non-Darcy flow is included

14. Log-derived rock mechanical properties

15. Laboratory triaxial measurements of core samples

16. Regional tectonic forces

However, no single sand prediction method can accommodate all these data reason being that the process of data acquisition is extensive and such information are not available during field development.

The process of sand production starts with the mechanical failure of rock near the wellbore. Sand particles become loose from the formation matrix due to shear opening of the rock and become available to be transported by formation fluid to the wellbore. This process is governed by the formation intrinsic strength and the effective in-situ stresses at that depth in the formation. Once the sand grains are loose, the rate and amount of erosion of the matrix depends on factors such as production rate, the fluid velocity and the fluid viscosity (Navjeet, 2004). The mechanisms responsible for sand production (i.e., sand failure mechanisms) are:

1. Compressive or Shear failure

2. Tensile failure due to pressure drawdown

3. Erosion or Cohesive failure due to cementation degradation

2.4.3.1 Compressive or Shear Failure

Compressive failure refers to an excessive, near cavity wall, tangential stress which causes shear failure of the formation matrix. Compressive failure occurs predominantly in consolidated sandstones (Veeken et al, 1991). Shear failure condition can be triggered by far- field stresses (depletion) and drawdown pressure. Rock strength criterion plays an important role in sand production from shear failure. Shear strength consists of two components; cohesion or physical bonds between the adjoining sand grains and friction. As a result of shear failure, reduction in hole size due to plastic failure near the perforation tunnel might occur. Around the perforation tunnel a stress concentration field is established. This stress concentration field causes the rock to respond either elastically as in strong formation or yield (weak formation), in which case a plastic zone is developed around a perforation tunnel. Large and small sand grains are generated and formation starts deteriorating at the failure plane once shear failure occurs.

Various failure criterions can be used to predict the shear failure mechanism of a rock. Among which are Von Mises, Drucker-Prager, Mohr-Coulomb, Hoek-Brown, Modified Lade and Modified Weibols & Cook. The choice of the failure criterion can be guided by a laboratory experiment to understand behaviour of the rock. The Mohr-Coulomb criterion is the most widely used for shear failure prediction. This criterion considers only the effects of the maximum and minimum principal stresses. It postulates that the failure occurs when shear stress on a given plane within a given plane within the rock reaches a critical value given by:

$$\tau = C + \sigma_n \tan\theta \qquad [2.2]$$

Where, τ is the shear strength, psi

σ_n is the stress normal to the failure plane, psi

C is the cohesive strength, psi

θ is the internal friction angle, degrees

The equation consists of two components; cohesion (C) and friction ($\sigma_n \tan \theta$). Cohesion failure produces the sand particles while shear failure breaks the rock along the shear plane. Assumed material behaviour for shear failure models include: linear elastic brittle, elastic plastic.

2.4.3.2. Tensile Failure

Tensile failure refers to a tensile radial stress exceeding the tensile failure envelope and is triggered exclusively by drawdown pressure. From the tensile failure criterion when a fluid flows into a cavity at high production flow rates, tensile net stresses can be induced near the cavity resulting in formation failure. The mechanism of tensile failure occurs at the perforation tunnel; here the radial stress is controlled by the reservoir pressure and wellbore pressure. Sudden pressure changes can exceed the tensile strength of the formation, causing sand production and subsequent enlargement of the perforation tunnel. Tensile failure may occur at the perforation tip or the perforation wall which is usually penetrating within the plastic zone (Navjeet, 2004). Weingarten and Perkins, 1995 studied the conditions necessary for formation stability around a spherical cavity in weakly consolidated rock. An equation describing tensile failure condition using pressure drawdown, formation rock cohesion and frictional angle was derived. They provided dimensionless curves for determining the pressure drawdown at a specified wellbore

pressure.

2.4.3.3. Erosion or Cohesion Failure

Erosion refers to a gradual removal or production of individual sand particle from the cavity surface (perforation tunnel, wellbore surface in open-hole completion etc.). Erosion is controlled by the cohesive strength. Erosion will take place if the drag force exerted on a surface particle exceeds the (apparent) cohesion between surface particles. The frictional drag is directly related to the velocity of the fluid flow. Hence, fluid velocity becomes an important parameter. This is confirmed by field experience, in loosely consolidated formations, sand production from open holes tends to be less than from perforated completion. This is in line with the fact that the fluid velocity at the open hole surface is three orders of magnitude smaller than the velocity at the (intact) perforation surface. Erosion is related to tensile failure, but needs to be considered as a separate mechanism due to its particulate nature (Veeken et al., 1991). Analytical approach captures the mechanisms of sand production, they can be implemented and calibrated more easily compared to numerical methods. Important aspects of sand production captured by analytical approaches are: stresses, rock strength, in-situ stresses. The time and effort needed for analyses are reduced and the difficulties of obtaining complex input parameters are overcome by analytical methods.

2.4.4. Numerical Methods

These are finite element analysis models that incorporate the full range of formation behaviour during plastic, elastic and time-dependent deformation. Numerical models provide a detailed description of the stress state and can be accurate. In comparison to other methods of prediction, numerical method is regarded as superior because it accounts for more factors influencing rock

failure and sand production. However, the main disadvantage of the method is its complexity and time consumption. Time, resources and data needed for the method might not be available. When properties needed in the numerical modelling are assumed or approximated due to lack or real data, results from the complex modelling are not necessarily more accurate or reliable than that from other approaches that uses simpler easily accessible data.

Another method used in sand prediction is the analogy or historical method. This relies on production experiences such as rate, drawdown, water-cut etc. from other wells in the same reservoir or nearby fields (offset data) to arrive at a choice between sand control and sand prevention. The most critical factors to determine the sand production potential of a reservoir formation are (1) formation strength (2) in-situ stresses (3) production rate. Formation intrinsic strength is however the key information needed. Zhang et al., 2000 developed a simple and efficient approach to evaluate formation strength. They found out that to construct a universal failure envelope, the only parameter needed is the critical pressure. Conventional logs data (compressional wave velocities) can be used to obtain the failure envelope of a sandstone formation. The generality of their observation is still explored.

2.5. Sand Control

The concept of sand control is based on the absolute exclusion of sand; zero tolerance of sand production at the surface. Problems associated with sand production have provided justification for downhole sand control devices. Once it has been established through sand prediction that at the desired production rate the reservoir will produce sand. The question of the best completion

practice to mitigate sand is raised. The choice of the sand control method to be used in a reservoir depends on operating practices, conditions of the field (formation sand characteristics), successful field experiences and economic considerations. Traditionally, the main classes of sand control techniques are mechanical and chemical. Available sand control techniques in the industry include:

1. Rate control or exclusion
2. Non-impairing completion techniques
3. Selective perforation practices
4. Screens (without gravel packs)
    - Slotted liners
    - Wire-wrapped screens
    - Premium screens
    - Expandable screens
    - Pre-packed screens
5. Gravel packs
6. Frac packs
7. Chemical sand consolidation
    - In situ formation consolidation
    - Consolidated gravel

The techniques highlighted above can be further divided into two groups: mechanical exclusion

methods and arch stabilization methods (Najveet, 2004). Arch stabilization methods can be further divided into natural arches and reinforced arches. Classification of the above sand control methods given by Najveet, (2004) are presented in table 2.3. The mechanical exclusion methods are designed to prevent sand production through bridging type retention or filter type retention. Bridging type retention allows a certain quantity of sand production until a bridge is formed against a filtration medium such as a screen or sized gravel or the two in combination (e.g., gravel packs). These bridges are disturbed easily by abrupt changes in production rate, resulting in sand production until a new bridge form. In filter type retention, sand production is excluded and does not depend on the formation of bridges. Filter type sand control is attained simply by further reducing slot size of the screen and size of gravel, below that required for bridging type retention (Najveet, 2004).

Arch stabilization depends on the formation of stable arches near the wellbore to prevent sand production. Natural arch stabilization is produced by avoiding arch destabilizing actions that can induce sanding. Stability of natural arches is sensitive to changes in flow rate. Reinforced arches are produced using chemical bonding agents such as plastic resin etc. to create new bonds and strengthen existing ones between adjoining sand grains. Often a combination of techniques is used to ascertain reliability of sand control. Such situation may occur when the parameters of the well exceed the design applicability of a specific control method.

2.6. A review of ML and AI Tools in petroleum engineering

Machine learning is the science of getting computers to learn, without being explicitly programmed (Andrew, 2015). Machine learning is all about getting a computer to perform a task

by watching you do it (Andrew, 2015). In this case, the computer mimics your actions and perform it later when you might not even be there.

Learning algorithms are applied in every aspects of our lives from reading emails, to google search. Machine learning is all about creating intelligent machines. This field has grown so large that it has gain so much recognition and it has stood out from the shadow of AI (artificial intelligence) (Andrew, 2015). Simple machine learning algorithms can be utilized to solve problems in every aspect of our lives. The problem with machine learning is not developing the said algorithms but actually applying these algorithms to problems of interest (Andrew, 2015).

Fields where machine learning is applied is very vast ranging from advanced robotics, to biological analytics, to even the E&P industry where this project is geared towards. Examples of application of machine learning algorithms includes: Data mining, as a result of large datasets from growth of automation and web (web click data, medical records, biology, and engineering); Applications that cannot be programmed by hand (e.g. Autonomous helicopter, handwriting recognition, most of natural language processing (NLP), and computer vision); self-customizing programs (e.g. Amazon, Netflix product recommendations, iTunes music recommendations); and Understanding human learning (brain and AI) (Andrew, 2016).

It is somehow complex to define machine learning as there are several reasons to this. According to Arthur Samuel (1959), Machine learning is a field of study that gives computers the ability to learn without being explicitly programed (Samuel, 1959). In 1997, Tom Mitchell (1997) proposed the following functional definition: “A computer program is said to learn from experience, E, with respect to some task, T, and some performance, P, if its performance on T, as measured by P, improves with experience, E” (Mitchell, 1997). There are two main types of machine learning algorithms; Supervised learning and Unsupervised learning. Other less used types of machine

learning algorithms are; Reinforcement learning, and recommender systems. All these will be discussed in details in the upcoming section.

There are several types of machine learning tools: Artificial Neural Networks (ANNs) (Agatonovic-Kustrin & Beresford, 2000), Support Vector Machines (SVMs) (Smith & Frank, 2016), Least Squares Support Vector Machines (LSSVMs) (Geron, 2019), Convolutional Neural Networks (CNNs) (Shin, 2020), and a host of others. Some of these machine learning tools are better than others in solving different kinds of problems. For the purpose of these work, ANNs and SVMs concepts are summarized.

2.6.1. Artificial Neural Networks (ANNs)

An artificial neural network is an information processing system that has certain performance characteristics in common with biological neural networks (Geron, 2019). A neural network is characterized by; its pattern of connections between the neurons (known as its architecture), its method of determining the weight on the connection (known as training or learning algorithm), and its activation function (Fausett, 1994).

The artificial neural network consists of an enormous number of interconnected simple processing elements called neurons, unit, cells, or nodes. The connection between neurons are weighted, and the weights represent information used by the network to solve problems. Neural networks have been applied in several areas including storing and recalling of data and pattern, performing general mapping from input patterns to output patterns, grouping of similar patterns, or finding solutions to the constrained optimization problem (Fausett, 1994).

The characteristic mode of operation and function of the brain inspired the development of the neural network. The brain operates in a very complex manner consisting of approximately 10^{11}

neurons of highly connected elements, approximately 10^4 connections per element (Demuth *et al.,* 2014). These neurons have three principal components: the dendrites, the cell body (also known as soma) and the axon. The figure below shows a schematic representation of a biological neuron.

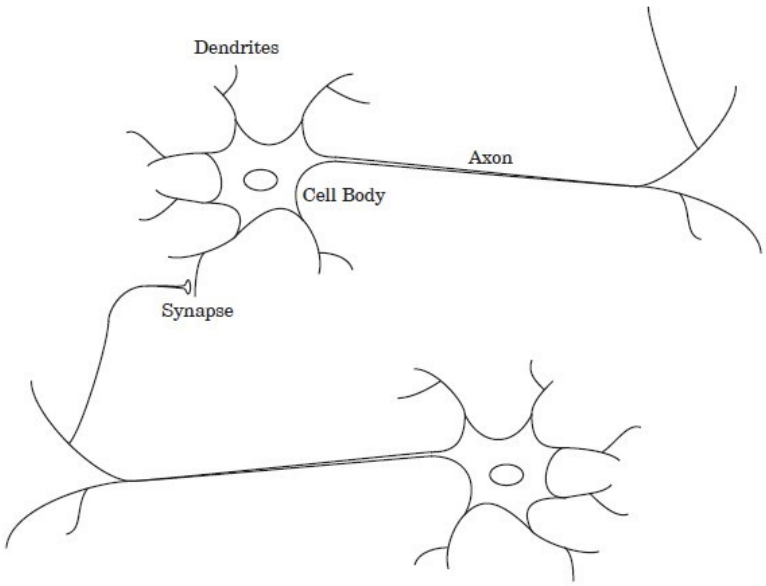


Figure 2.4: The Neuron (Source: Demuth et al., 2014)

The dendrites are receptive network of nerve fibers that carry electrical signals into the cell body, the summation of these incoming signal threshold is done in the cell body, and the axon, which is a single long fiber, carries the signal from the cell body out to other neurons. A synapse is a point at which the axon of one neuron contacts the dendrite of another neuron.

The complexity in the manner in which the brain operates is enormously superior to that of an artificial neural network. However, there exist some similarities between the artificial and biological neural network. To start with, the building pieces of both networks are basic computational gadgets that are exceptionally interconnected, although artificial neural network

is considerably less complex than biological neural network. Secondly, the connections between neurons decide the capacity and function of the network. Artificial neural networks are developed to mimic the biological neural system based on the following assumptions (Fausett, 1994).

1. Information is processed in many basic components called neurons.
2. Connection links enable transfer of signals between neurons.
3. Weights are associated with connection links; the weights multiply the signals transmitted.
4. To determine the output signal, each neuron applies an activation function to the net input signal.

Mathematically, the input signal can be expressed as the sum of the weighted inputs and bias.

$$network\ k = \sum_{i=1}^{n}(w_{ki} * x_i) + b_k \qquad [2.3]$$

Table 2.1: Terminological relationship between artificial and biological neurons (Source: BuKhamseen, 2014)

Biological Neurons	Artificial Neurons
Neurons	Node / Unit / Cell / Neurons
Synapse	Connection / Link / Edge
Synaptic Efficiency	Connection
Strength / Weight	Strength / Weight
Firing Frequency	Node Output

Notwithstanding, if one prefers not to have a bias in a specific neuron, it can be discarded. It is worthy of note that both *w* and *b* are adjustable parameters of the neuron. The processed signal is sent to the output node through a transfer function. The terminological relationship between

artificial and biological neural network is shown in Table 1.

2.6.1.1. Brief History of Artificial Neural Network

The advancement in neural network development has occurred with several hitches rather than a smooth evolution. Some of the significant background work for the field of neural network occurred in the late 19th and early 20th century. Around that period, emphasis was mainly on the general theories of learning, vision, conditioning, etc., and did not include specific mathematical models of neuron operation (Demuth *et al.,* 2014).

A new phase of research in neural network began in the 1940s with the work of McCulloch, W.S. & Pitts, W. (1943). They suggested that the relation among neural events can be treated using propositional logic and further discussed the various application of calculus involved. Their work is often acknowledged as the origin of the neural network field (Demuth *et al.,* 2014).

Rosenblatt in (1958) was the first to practically apply the concept of artificial neural network by inventing the perceptron network and associated learning rule. He tried to understand the capability of higher organisms to recognize, generalize, recall, and think by trying to provide logical explanations to two fundamental questions: what form is information stored or remembered, and how does information in memory influence recognition and behavior? Rosenblatt built a perceptron network that was able to perform the simple task of pattern recognition, and his success generated much interest in the research area. Unfortunately, the basic perceptron network developed by Rosenblatt could only solve a limited class of problems. Neural network-related research was suspended for about a decade because there were no powerful digital computers on which to experiment. In the 1970's, there was limited progress in the development of new neural network ideas within the scientific community. Fortunately, things turned around for the better in the 1980's. Research in the neural network was reinvigorated because of the computational

revolution that occurred during this period, and two new main concepts were developed: the use of statistical machines to explain the operation of a particular class of recurrent network, and the development of backpropagation algorithm for training multilayer perceptron networks. Recent advances in the neural network have had to do with new concepts, such as innovative architecture and training rules (Demuth *et al.,* 2014).

2.6.1.2. Artificial Neural Network Architecture

Complex problems may require layering of several neurons operating in parallel to obtain reasonable solutions. The layered arrangement of these connected neurons defines the architecture of the artificial neural network (Alghazal, 2015). There are three classes of a neural network: single layer feedforward network, multilayer feedforward network, and recurrent network.

Single Layer Feedforward Network

A single layer network includes the weight matrix, the summers, the bias vector *b*, the transfer function boxes, and the output vector. The weight matrix *w* connects every element in the input vector *p* to each neuron. Each neuron has a bias *bi*, a summer, a transfer function *f*, and an output *ai*. The output from each neuron when put together forms the output vector *a*. A single layer network has only one hidden layer as suggested by its name. It is referred to as feedforward because the hidden layer is connected to the input vector and for the case of multi-layered network, the subsequent layer is connected to the previous layer. A schematic representation of a single-layered network is shown in Figure 2.5.

Multilayer Feedforward Network

Multilayer neural networks usually have one or more layers between the input and output layer known as hidden layers (Fausett, 1994). Each layer has its own weight matrix, bias vector, a net

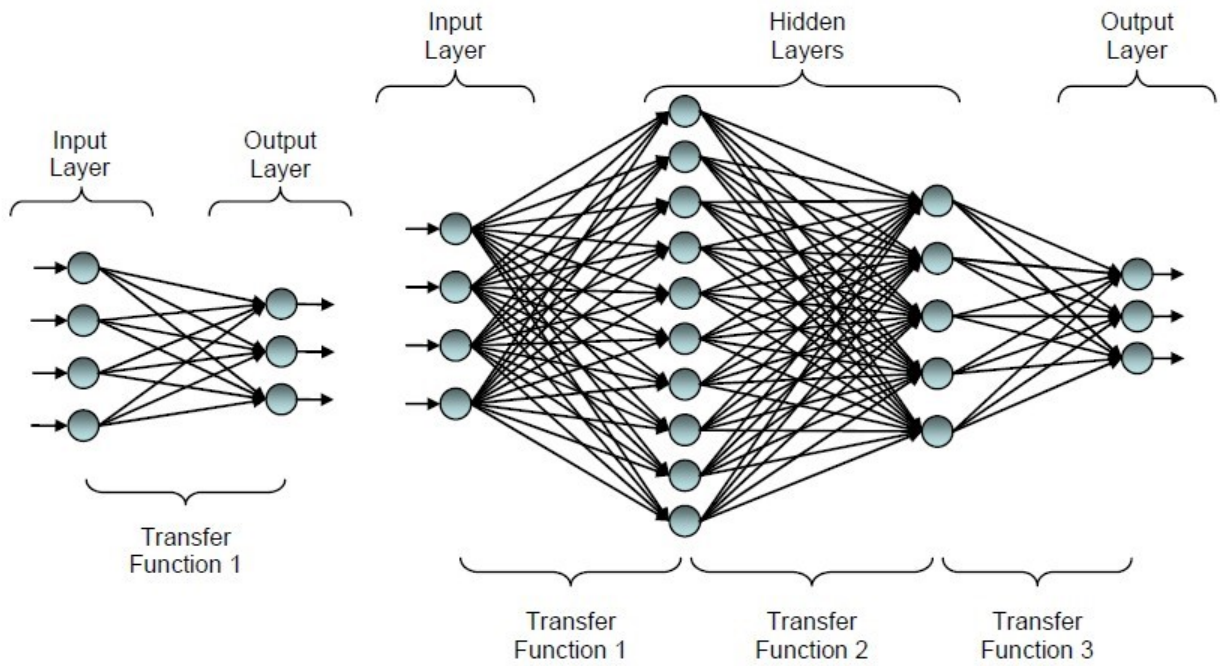


Figure 2.5: Single and Multilayer neural network architecture (Parada, 2008).

input vector and an output vector. Problems that are complex to handle by a single layer network can be handled by a multilayer network. However, training of the network may be difficult. Figure 2.5 shows the schematic representation of a multilayer network architecture.

<u>Back Propagation</u>

Back propagation was originally introduced in the 1970s, but its usefulness was not appreciated until 1986 after a paper publication by David Rumelhart, Geoffrey Hinton, and Ronald Williams. In this paper, several neural networks were described to work faster when back propagation was implemented. Therefore, problems thought to be unsolvable became possible to solve. So, back propagation became the backbone for neural networks. At the center of back propagation is an expression for the partial derivatives $\partial C/\partial w$ of the cost function C with respect to any weight, w (or bias, b) in the network (Nielson, 2006). The expression tells us how quickly the cost changes when we change the weight and biases.

The goal of back propagation is to compute the partial derivatives $\partial C/\partial w$ and $\partial C/\partial b$ of the cost function C with respect to any weight w or bias b in the network. For back propagation to function, Nielson (2006) brought out 2 main assumptions about the form of the cost function. Firstly, the cost function could be written as an average $C = \frac{1}{n}\sum_x C_x$ over cost function C_x for individual training examples, x. The reason for this assumption is that, what back propagation actually lets us do is to compute the partial derivatives $\partial C/\partial w$ and $\partial C/\partial b$ for a single training example. The second assumption is that the cost can be written as a function of the outputs from the neural network (Nielson, 2006). In order to compute the partial derivatives $\partial C/\partial w_{jk}$ and $\partial C/\partial b_j$, Nielson (2006) first introduced an intermediate quantity, δ_j, which we call the error in the j-th neuron in the l-th layer. Back propagation will give us a method to compute the error δ_j, and then will relate δ_j to $\partial C/\partial w_{jk}$ and $\partial C/\partial b_j$.

The first equation is an equation for the error in the output layer δ^L. The components of δ^L are given by

$$\delta_j^L = \frac{\partial C}{\partial a_j^L}\sigma'(z_j^L) \qquad [2.4]$$

Where:

$\frac{\partial C}{\partial a_j^L}$ is the measure of how fast the cost is changing as a function of the j-th output activation.

$\sigma'(z_j^L)$ is the measure of how fast the activation function σ is changing at z_j^L.

It should be noted that the first equation for back propagation can easily be calculated by finding z_j^L while computing the behavior of the network, and it is only a small additional overhead to

compute $\sigma'(z_j^L)$. The exact form of $\frac{\partial C}{\partial a_j^L}$ will, of course, depend on the form of the cost function. However, provided the cost function is known there should be no problem computing $\frac{\partial C}{\partial a_j^L}$. Converting the equation to a matrix-wise form necessary for back propagation, we have

$$\delta^L = \nabla_a C \circ \sigma'(z^L) \qquad [2.5]$$

$\nabla_a C$ is a vector whose components are the partial derivatives $\frac{\partial C}{\partial a_j^L}$ (that is the rate of change of C with respect to the output activations).

The second back propagation equation is the equation for the error δ^l in terms of the error in the next layer δ^{l+1}. (Nielsen, 2006). So,

$$\delta^l = ((w^{l+1})^T \delta^{l+1}) \circ \sigma'(z^l) \qquad [2.6]$$

Where: $(w^{l+1})^T$ is the transpose of the weight matrix w^{l+1} for the (l+1)-th layer.

We can compute the error δ^l for any layer in the network by combining the first two equations.

The third back propagation equation is the equation for the rate of change of the cost function with respect to any bias in the network (Nielsen, 2006). That is,

$$\frac{\partial C}{\partial b_j^l} = \delta_j^l \qquad [2.7]$$

Finally, the fourth and last equation is the equation for the equation for the rate of change of the cost with respect to any weight in the network (Nielsen, 2006). That is,

$$\frac{\partial C}{\partial w_{jk}^{l}} = a_k^{i-1}\delta_j^l \quad [2.8]$$

This tells us how to compute the partial derivatives $\frac{\partial C}{\partial w_{jk}^{l}}$ in terms of the quantities a_k^{i-1} *and* δ_j^l, which we already know how to compute (Nielsen, 2006). In a less-heavy notation,

$$\frac{\partial C}{\partial w} = a_{in}\delta_{out} \quad [2.9]$$

Where it is understood that a_{in} is the activation of the neuron input to the weight w, and $\delta_{out}$is the error of the neuron output from the weight w. The consequence of this equation is that the weight output from low-activation neurons learn slowly (Nielsen, 2006).

2.6.1.3. Recurrent Networks

Recurrent networks have at least one feedback loop; this makes them different from the feedforward network in the sense that some of the outputs are connected to the input. The output of the recurrent network is the input delayed by one step, and the future outputs of the network are computed from future outputs. Recurrent networks are potentially more powerful than feedforward networks (Demuth *et al.,* 2014).

2.6.1.4. Transfer function

A transfer function basically scales the response of a neuron to an external stimulus and generates the neuron activation (Maren *et al.,* 1990). Transfer function could be broadly classified as linear or nonlinear. The transfer fraction employed for solving problems depends on the complexity of the problem. Complex problems require the use of a multilayer neural network architecture, and a multilayer network requires the use of a nonlinear activation function. Most times, for a given layer, the same activation function is applied to all neurons, but for special cases different

activation functions may be applied to a single layer. The most frequently used transfer functions are in multilayer networks are the linear (or purelin) and the sigmoid (logsig) transfer functions (Alghazal, 2015).

The purelin transfer function is a linear function that is mostly associated with the output layer since the network can produce within the desired limit without having to denormalize them (Minakowski, 2008). Sigmoid functions (S-shaped curves) are powerful continuous nonlinear transfer functions. The most commonly used sigmoid function are the logistic function and the hyperbolic tangent function. They are most helpful when used in a neural network trained by backpropagation, this is because there exists a simple relationship between the value of the function and its derivative at a point which reduces the burden of computation during training. The logistic and hyperbolic tangent function respectively take the input signal, which may lie between plus or minus infinity, activates the neuron, and scales the output signal within the range of (0 and 1), and (-1 and 1). Table 2 shows a list of activation functions.

Table 2.2: Transfer functions (Source: Demuth et al., 2014)

Name	Input/Output Relation	Icon	MATLAB Function
Hard Limit	$a = 0 \quad n < 0$ $a = 1 \quad n \geq 0$		hardlim
Symmetrical Hard Limit	$a = -1 \quad n < 0$ $a = +1 \quad n \geq 0$		hardlims
Linear	$a = n$		purelin
Saturating Linear	$a = 0 \quad n < 0$ $a = n \quad 0 \leq n \leq 1$ $a = 1 \quad n > 1$		satlin
Symmetric Saturating Linear	$a = -1 \quad n < -1$ $a = n \quad -1 \leq n \leq 1$ $a = 1 \quad n > 1$		satlins
Log-Sigmoid	$a = \frac{1}{1 + e^{-n}}$		logsig
Hyperbolic Tangent Sigmoid	$a = \frac{e^{n} - e^{-n}}{e^{n} + e^{-n}}$		tansig
Positive Linear	$a = 0 \quad n < 0$ $a = n \quad 0 \leq n$		poslin
Competitive	$a = 1$ neuron with max n $a = 0$ all other neurons	C	compet

Where w is the weight, x is the input data, and b is the bias. The bias is like a weight, with the exception of its constant value of one that shifts the transfer or activation function to the right or left.

2.6.1.5. Learning Methods and Training Algorithm

Learning is the end product of a successful training procedure that is achieved by modifying the connection weights between nodes and biases of a neural network. Learning methods can be classified into supervised and unsupervised learning, there's, however, a third category where the weights are fixed without an iterative training process (Fausett, 1994). Supervised learning requires that an input vector or pattern be provided with its associated target or output vector or pattern. The weights are initialized to a small random number or zero; the weights are then iteratively manipulated until the calculated output vector is like the initial output vector within some acceptable range of error. The unsupervised learning does not require an output vector; the weights are modified such that the most similar input vector is assigned to the same target output.

Mathematical functions that are responsible for adjusting the weights and biases of the network are referred to as Training Algorithm. A number of training algorithms exist, including the Levenberg-Marquart (LM) algorithm, the gradient descent methods, the resilient backpropagation algorithm, the conjugate gradient methods, Newton's method and the Quasi-Newton's method. Most oil and gas applications employ the feedforward network with backpropagation supervised learning algorithm. The backpropagation algorithm is a gradient descent method that minimizes the total error between the output of the network and the desired targets by adjusting weights and biases using a generalized rule of the least mean square error (LMS) (Faussett, 1994).

Convergence, Training Efficiency, and Overfitting

Convergence is said to take place when the total error of the current iteration is lower than the previous iteration. When the error between the desired and calculated outputs is minimized to a certain threshold, final convergence is said to have occurred. Convergence problem arises when the network calculates a lower total error as compared to that of the previous iteration. However, a global minimum is not reached due to the presence of at least local minima on the error surface. Training efficiency can be improved and made less sensitive to local minima on the error surface by combining techniques such as functional links, batching, adaptive learning and momentum (BuKhamseen, 2014, Beale *et al.,* 2017).

Training efficiency can be improved upon by altering the learning rate. Large learning rates lead to faster training, although some degree of fluctuation may be introduced. Small learning rates provide a more stabilized but delayed learning process, and this may direct convergence to a local minimum (Parada, 2008). The number of hidden layers chosen, and the number of neurons in each hidden layer affects the convergence speed. Choosing a number of neurons less than the required number leads to underfitting, while too many neurons may lead to overfitting (BuKhamseen, 2014).

Overfitting is one of the major problems during neural network training. Overfitting occurs when the error on the training set is driven to a small value but results in a large error when the network is presented with new data. This is as a result of the memorization of the trained examples rather than learning to generalize to new situations. Regularization and early stopping are techniques used for preventing memorization and improving generalization (Beale *et al.,* 2017).

2.6.2. Support Vector Machines (SVMs)

Unlike ANNs, SVMs gives a cleaner and better way to solve complex non-linear regression functions. Support vector machines is a popular supervised machine learning algorithm that is used to analyzed and recognize patterns of input/output data. (Vapnik, 1995). In the data mining and machine learning communities, SVMs are considered quite effective in terms of its mathematical approaches. The SVM classification model uses the optimal separation rule, whereby the input parameters are non-linearly mapped in a high-dimensional features space. This rule selects an optimal separating hyperplane with the maximum distance between linearly separable classes. (Olatunji & Micheal, 2017)

SVMs is used for both classification and regression problems but are quite popular in solving classification problems. The main objective of a SVM is to create the best line or decision boundary that can segregate n-dimensional space into classes. Separating the space into classes can make it easily to place a new data point to the correct category in future. There can be multiple lines or decision boundaries to segregate classes in n-dimensional space, hyperplane refers to the best decision boundary. The dimensions of the hyperplane depend on the features present in the dataset. If there are 2 features, the hyperplane will be a straight line, if there are 3 features, the hyperplane will be a 2-dimensional plane.

SVMs work by choosing the extreme points/vectors that help is creating this hyperplane. The extremes are referred to as support vectors, and hence the term support vector machines. There are two types of SVMs; Linear SVM, used for linearly separable data; and Non-linear SVM used for non-linearly separated data.

Data is said to be linearly separable if it can be classified into two classes by using a single straight line. The classifier used for this type of SVM is called a Linear SVM Classifier. Also, a dataset is

said to be non-linearly separable if it cannot be classified by using a single straight line. The classifier used here is known as a Non-linear SVM Classifier.

2.6.2.1. How SVMs Work

Unlike ANN, SVM give a cleaner and better way to solve complex non-linear regression functions. According to Cortes and Vapnik (1995) the support vector machines is a popular supervised machine learning algorithm that is used to analyze and recognize patterns of input/output data (Cortes and Vapnik, 1995). The SVM classification model uses the optimal separation rule, whereby the input parameters are non-linearly mapped in a high-dimensional features space. This rule selects an optimal separating hyperplane with the maximum distance between linearly separable classes. (Olatunji and Micheal, 2017). SVM is used for both classification and regression problems but are quite popular in solving classification problems.

The SVM algorithm finds the closest points (called support vectors) of the lines from both classes. The distance between the vectors and the hyperplane is referred to as the margin. The main objective of the SVM is to maximize this margin. The optimal hyperplane is the hyperplane with maximum margin.

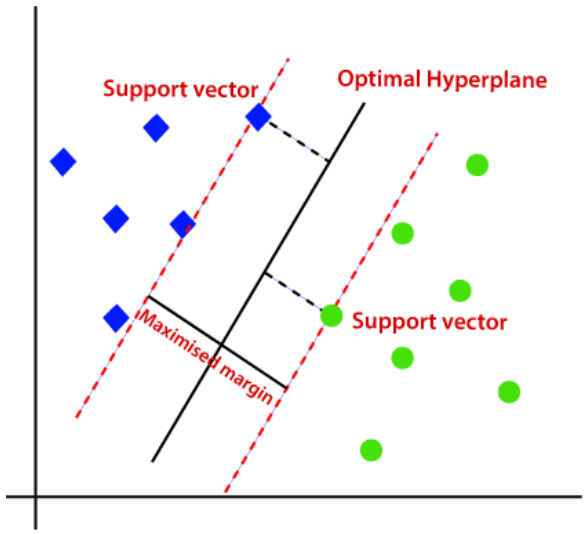


Figure 2.6: Linear Classification (Source: Shin, 2020)

For a non-linear classification, it will be impossible to separate the dataset by a straight line. Therefore, to separate such datapoints, another dimension, say z, is added and calculated as:

$$z = x^2 + y^2 \qquad [2.10]$$

The resulting plane will be as shown below:

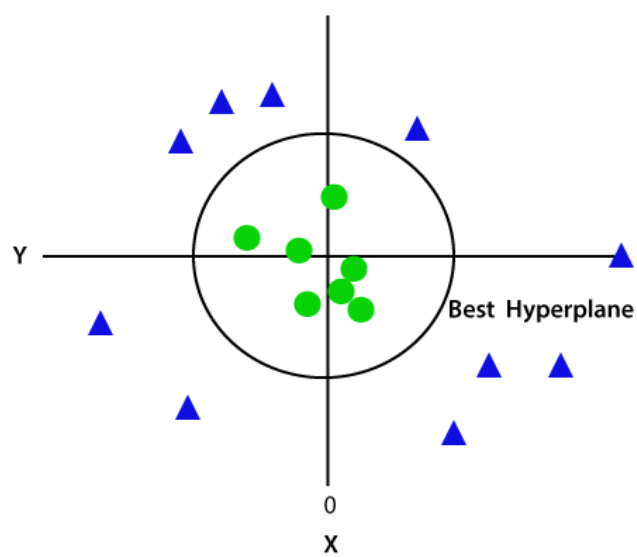


Figure 2.7: Non-Linear Classification (Source: Shin, 2020)

Chapter 3: Methodology

3.1 Overview

In this study, a comparative analysis is conducted to evaluate the performance of two algorithms, SVM vs. ANN with back propagation. The intent is to benchmark the effectiveness of the algorithms in predicting sand production especially in areas where the available training data is sparse. The following methodology is adopted to develop the algorithms and evaluate their performance. The methodology consists of four major steps, i.e., data pre-processing, data analysis, algorithm implementation, and defining the criteria for evaluating the performance of the algorithms. Figure 3.1 shows a schematic of the research methodology for the ANN algorithm. A similar schematic can be used to describe the methodology for the SVM.

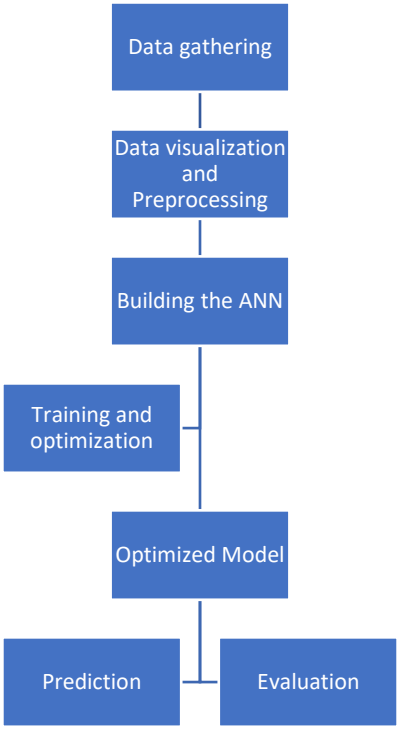


Figure 3.1: Schematic of Research Methodology for Developing and Applying the ANN

3.2 Data pre-processing

This involves splitting the data into the training set, test set, and validation set, as well as applying feature scaling. The percentages of the training set, test set, and that used for validation are altered during the development of the algorithm to achieve an optimum result. As part of the pre-processing, the data encoding is done by converting all text data into a numeric code. For the data used in this study, wells that produce sand are labeled “0”and those that do not are labeled “1.” Feature scaling or data normalization is applied to ease computations and to prevent one variable from dominating other variables. With the use of PYTHON, the data is imported, encoded, and displayed as shown in Figure 3.2.

Index	Depth	Overburden	Pore_Pressure	Min_Horizontal_Stress	Max_Horizontal_Stress	Well_inclination	well_azimuth	poissons_ratio	youngs_modulus	friction_angle	shale_content	Output	type
0	10080	0.871	0.439	0.86	0.91	8.89	156.22	0.25	0.351	30	0.354	No_Sand	1
1	10232	0.881	0.439	0.86	0.91	18.45	156.22	0.25	0.467	30	0.354	No_Sand	1
2	10863	0.896	0.439	0.89	0.94	18.45	156.22	0.25	0.584	30	0.354	No_Sand	1
3	11414	0.923	0.415	0.89	0.94	18.45	156.22	0.25	0.547	26.5	0.543	No_Sand	1
4	11995	0.928	0.47	0.89	0.94	18.45	156.22	0.25	0.598	26.89	0.573	No_Sand	1
5	12291	0.928	0.47	0.89	0.94	18.45	156.22	0.25	0.683	26.89	0.573	No_Sand	1
6	12544	0.928	0.477	0.89	0.94	18.45	156.22	0.25	0.683	27.13	0.67	No_Sand	1
7	13214	0.928	0.465	0.89	0.94	18.45	156.22	0.25	0.683	27.13	0.443	No_Sand	1
8	13691	0.928	0.466	0.89	0.94	18.45	156.22	0.25	0.683	27.13	0.6891	No_Sand	1
9	14075	0.928	0.466	0.89	0.94	18.45	156.22	0.25	0.683	27.13	0.6891	No_Sand	1
10	14200	0.928	0.464	0.89	0.94	18.45	156.22	0.25	0.683	27.89	0.6891	No_Sand	1
11	10080	0.871	0.47	0.88	0.93	11.09	115.08	0.25	0.351	30	0.354	No_Sand	1
12	10232	0.881	0.47	0.88	0.93	11.09	115.08	0.25	0.467	30	0.354	No_Sand	1
13	4000	0.822	0.435	0.78	0.83	2.08	156.22	0.25	0.2234	26.77	0.321	Sand	0
14	5243	0.841	0.435	0.78	0.83	2.08	156.22	0.25	0.2234	26.77	0.321	Sand	0

Figure 3.2: Encoded Dataset

3.3 Data Analysis

A total of 11 parameters are used to develop the algorithms. They are the geological and reservoir properties which are found to be the major factors affecting sanding in sandstone formations. These parameters include depth, overburden, pore pressure, maximum and minimum horizontal stress, well azimuth, well inclination, Poisson’s ratio, Young’s modulus, frication angle, and shale

content. Data typical of the Niger Delta are used to validate the models. Due to difficulty in obtaining data, 4 wells from the Niger Delta are used for model validation. The data is statistically populated using the knowledge gotten from literature about the location of the wells and overall depths of the reservoirs in the Niger Delta. Using PYTHON and the Pandas library, the data is encoded as “1” representing “no sanding” and “0” representing “sanding”. Figure 3.3 shows the correlation between the input parameters. It can be observed that the relationship between the parameters is fairly weak, except for the correlation between the shale content and friction angle with a value of -0.45. Sample size and statistical significance of the sample is considered in evaluating the correlation matrix. The low correlation values are probably a result of the small sample size used in the study.

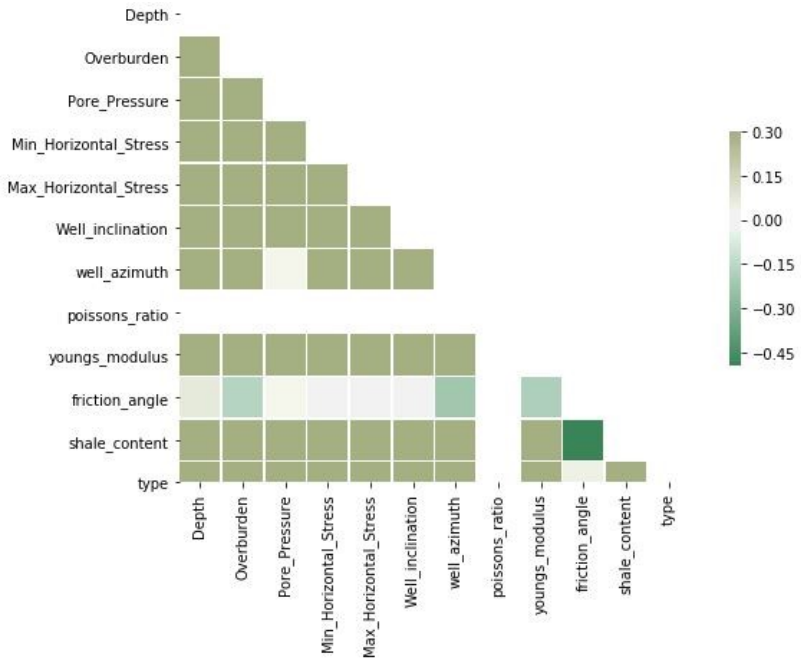


Figure 3.3:Correlation Matrix of Input Parameters

It is observed from Figures 3.3 and 3.4 which show the individual correlations, that parameters

such as overburden, pore pressure, maximum and minimum horizontal stress, well inclination, and Young's modulus have a strong correlation with reservoir depth. This is true as these parameters are influenced by the overburden pressure of the overlying rock in the reservoir. Furthermore, friction angle is seen to have little effect on other parameters as can be seen from its low correlation value. Also, the Poisson's ratio is not influenced by any of the parameters, and therefore, it is assigned NaN in this analysis. This means that Poisson's ratio is an independent parameter in our data set. It is known that Poisson ratio is only dependent on the transverse and axial strain of the reservoir rock. To summarize, parameters with a correlation value greater than 0.5 are considered to have good correlation whereas those less than 0.5 are considered to have weak correlation between the parameters.

Index	Depth	Overburden	Pore_Pressure	Min_Horizontal_Stress	Max_Horizontal_Stress	Well_inclination	well_azimuth	poissons_ratio	youngs_modulus	friction_angle	shale_content	type
Depth	1	0.946362	0.569595	0.830093	0.830093	0.919052	0.397043	nan	0.915294	0.0714576	0.632175	0.789614
Overburden	0.946362	1	0.580026	0.88482	0.88482	0.920812	0.465082	nan	0.968262	-0.165879	0.767932	0.829048
Pore_Pressure	0.569595	0.580026	1	0.591543	0.591543	0.450819	0.0234402	nan	0.593679	0.019193	0.535951	0.546619
Min_Horizontal_Stress	0.830093	0.88482	0.591543	1	1	0.823711	0.378818	nan	0.889796	-0.0176901	0.784902	0.987392
Max_Horizontal_Stress	0.830093	0.88482	0.591543	1	1	0.823711	0.378818	nan	0.889796	-0.0176901	0.784902	0.987392
Well_inclination	0.919052	0.920812	0.450819	0.823711	0.823711	1	0.375792	nan	0.905731	0.0120861	0.642144	0.788644
well_azimuth	0.397043	0.465082	0.0234402	0.378818	0.378818	0.375792	1	nan	0.471165	-0.220572	0.425998	0.370734
poissons_ratio	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan
youngs_modulus	0.915294	0.968262	0.593679	0.889796	0.889796	0.905731	0.471165	nan	1	-0.183402	0.78051	0.837512
friction_angle	0.0714576	-0.165879	0.019193	-0.0176901	-0.0176901	0.0120861	-0.220572	nan	-0.183402	1	-0.492861	0.0457662
shale_content	0.632175	0.767932	0.535951	0.784902	0.784902	0.642144	0.425998	nan	0.78051	-0.492861	1	0.751748
type	0.789614	0.829048	0.546619	0.987392	0.987392	0.788644	0.370734	nan	0.837512	0.0457662	0.751748	1

Figure 3.4: Data Representation of correlation matrix

Table 3.1 shows the results of the statistical analysis of the input parameters under study. Note that in this table, Std dev is the standard deviation.

The statistics from this set of input data (Table 3.1) indicate that the average depth of reservoirs in the Niger Delta is 9729 ft, overburden is 0.88 psi/ft, pore pressure is 0.45 psi/ft,

Table 3.1: Statistical Analysis

Index	Depth	Overburden	Pore Pressure	Min Horizontal Stress	Max Horizontal Stress	Well Inclination	Well Azimuth
	(ft)	(psi/ft)	(psi/ft)	(psi/ft)	(psi/ft)	(degree)	(degree)
Count	25	25	25	25	25	25	25
Mean	9729.70	0.8804	0.448	0.83	0.88	12.23	143.06
Std dev	2941.5	0.0368	0.0184	0.0528	0.0528	5.86	19.59
Min	4000	0.822	0.415	0.78	0.83	2.08	115.08
25%	8118.7	0.857	0.436	0.78	0.83	8.89	115.08
50%	10080	0.871	0.439	0.86	0.91	11.09	156.22
75%	11995	0.928	0.466	0.89	0.94	18.45	156.22
max	14200	0.928	0.477	0.89	0.94	18.45	156.22

Table 3.1 continued

Index	Poisson's Ratio	Young's Modulus	Friction Angle	Shale content
		(Mpsi)	(degree)	(%)
Count	25	25	25	25
Mean	0.25	0.432	28.15	0.379
Std dev	0	0.1798	1.32	0.186
Min	0.25	0.2234	26.50	0.112
25%	0.25	0.334	26.89	0.321
50%	0.25	0.351	28.31	0.354
75%	0.25	0.598	30.00	0.543
max	0.25	0.683	30.00	0.689

minimum horizontal stress is 0.83 psi/ft, maximum horizontal stress is 0.88 psi/ft, Poisson's ratio is 0.25, and Young's modulus is 0.43 Mpsi. Figures 3.5a to 3.5k show the visual distributions of the input parameters.

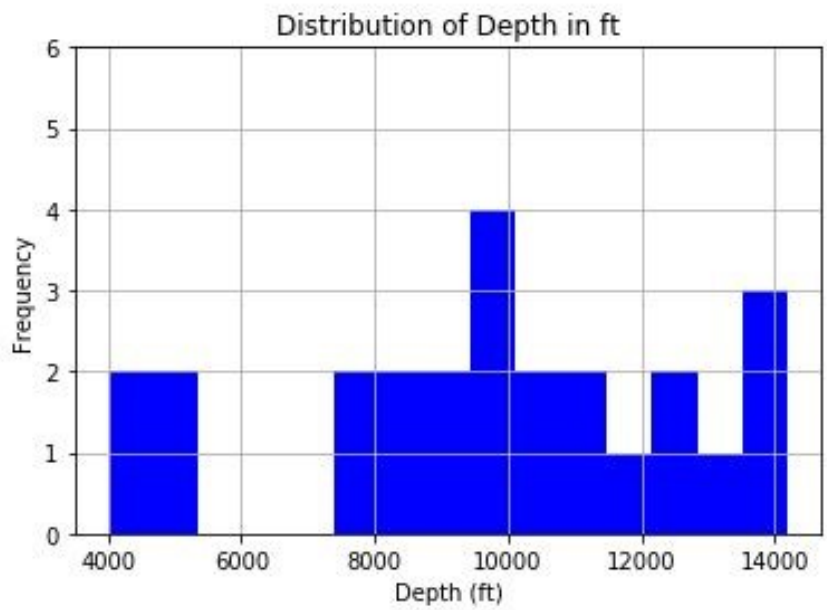


Figure 3.5a: Distribution of Depth

Figure 3.5a shows that the distribution of depth is skewed to the right with a mean depth of 9729.70ft.

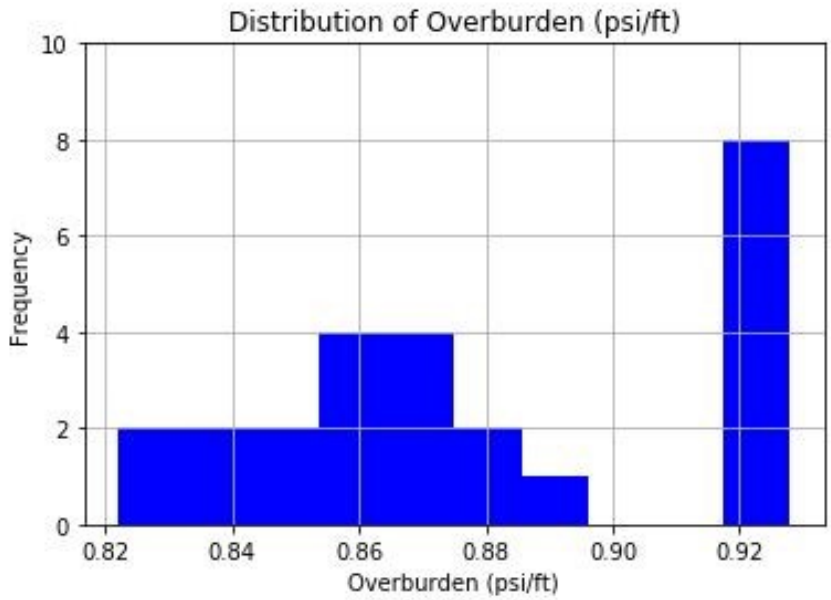


Figure 3.5b: Distribtuion of Overburden stress

Figure 3.5b shows the distribution of the overburden which is seen to be positively skewed. This means that more of the data is to the left hand side of the distribution with a few large ones on the right. The mean overburden is seen to be 0.8804 psi/ft.

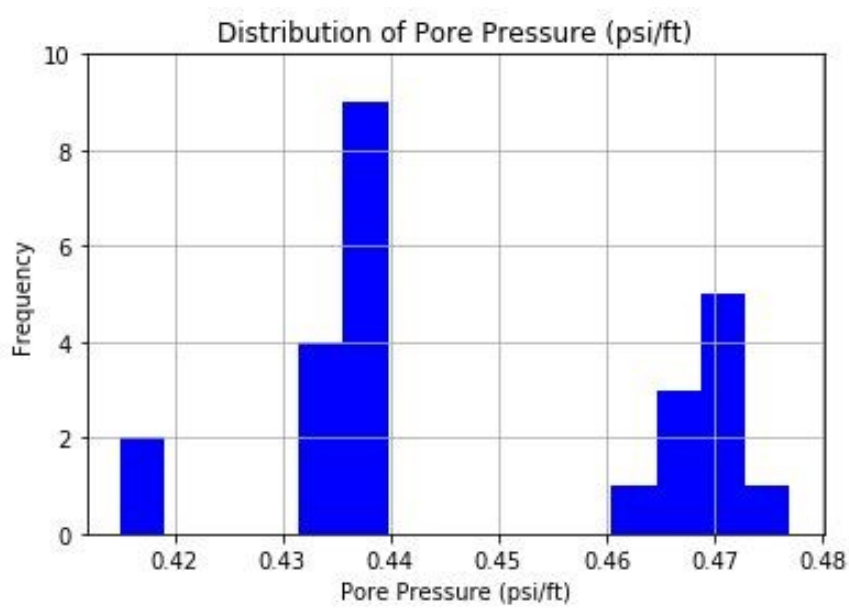


Figure 3.5c: Distribution of Pore Pressure

Figure 3.5c shows the distribution of the pore pressure with mean a pore pressure of 0.448 psi/ft.

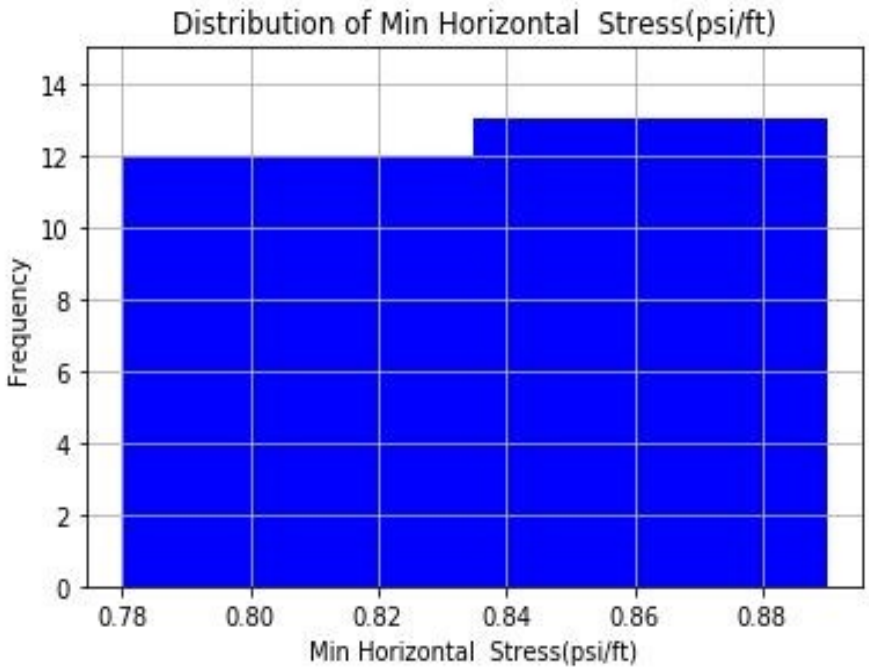


Figure 3.5d: Distribution of Minimum Horizontal Stress

Figure 3.5d shows the distribution of the minimum horizontal stress which is seen to be uniform. This means that the data is observed to be constant with mean value of 0.83 psi/ft.

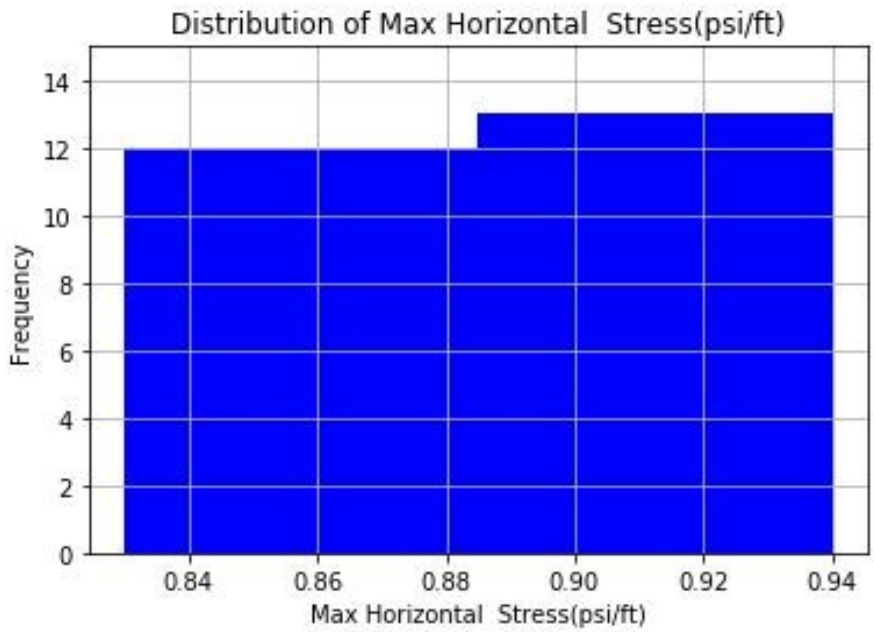


Figure 3.5e: Distribution of the Maximu Horizontal Stress

Figure 3.5e shows the distribution of the maximum horizontal stress which is seen to be uniform. This means that the data is observed to be constant with a mean stress of 0.88 psi/ft.

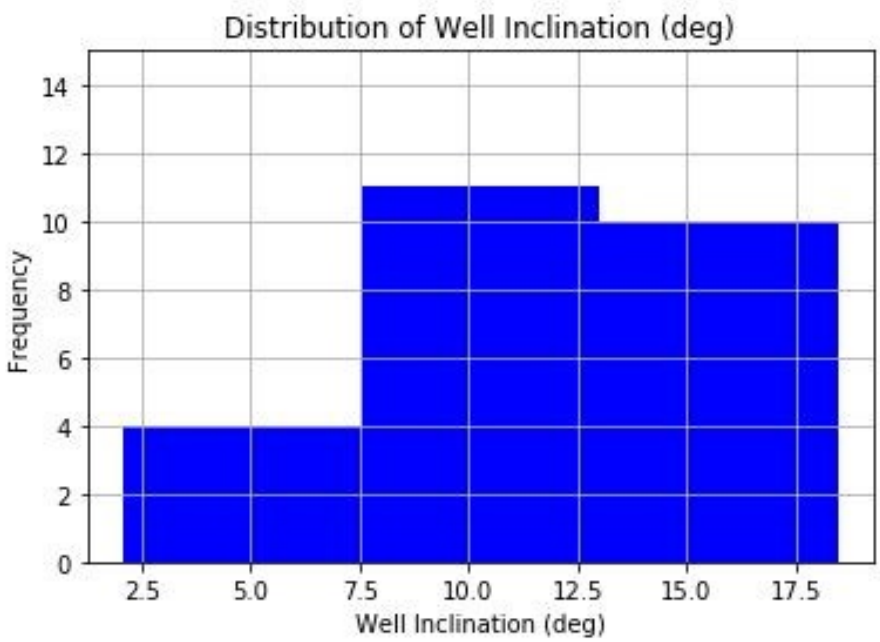


Figure 3.5f: Distribution of Well Inclination Angle

Figure 3.5f shows the distribution of the well inclination which is seen to be negatively skewed. This means that more of the data is to the right hand side of the distribution with a few large ones on the left. The mean value of the well inclination angle is 12.23 degrees.

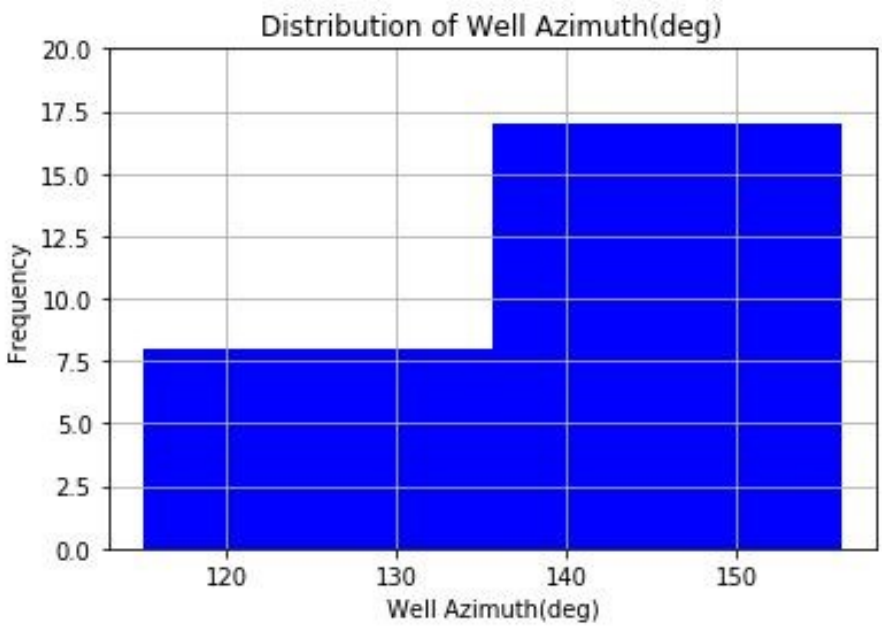


Figure 3.5g: Distribution of Well Azimuth

Figure 3.5g shows the distribution of the well azimuth which is seen to be negatively skewed. This means that more of the data is to the right hand side of the distribution with a few large ones on the left. The mean value of the well azimuth is 143.06 degrees.

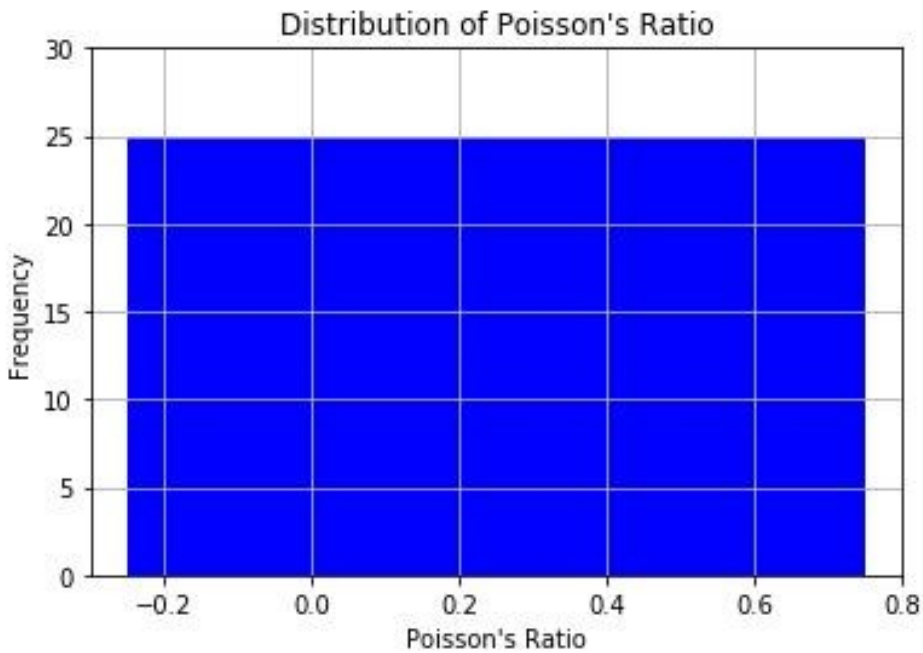


Figure 3.5h: Distribution of the Poisson's Ratio

Figure 3.5h shows the distribution of Poisson's ratio which is uniformly distributed. This means that the frequency of the data is observed to be constant with a mean value of 0.25.

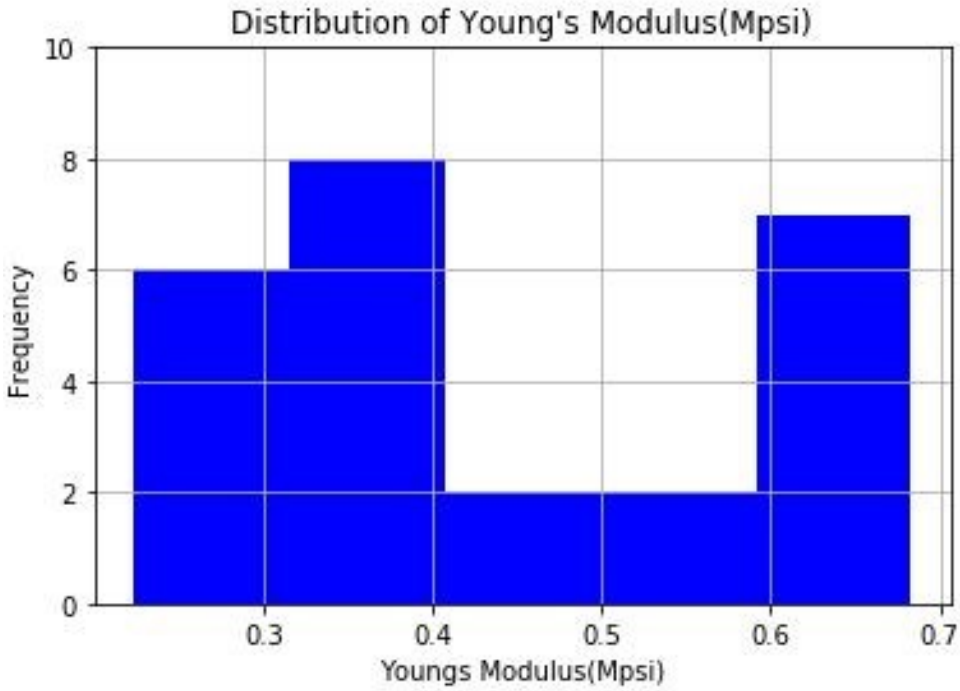


Figure 3.5i: Distribution of Young's Modulus

Figure 3.5i shows the distribution of Young's Modulus which is seen to be positively skewed. This means that more of the data is to the left hand side of the distribution with a few large ones on the right. The mean value of the Young's Modulus is 0.432Mpsi.

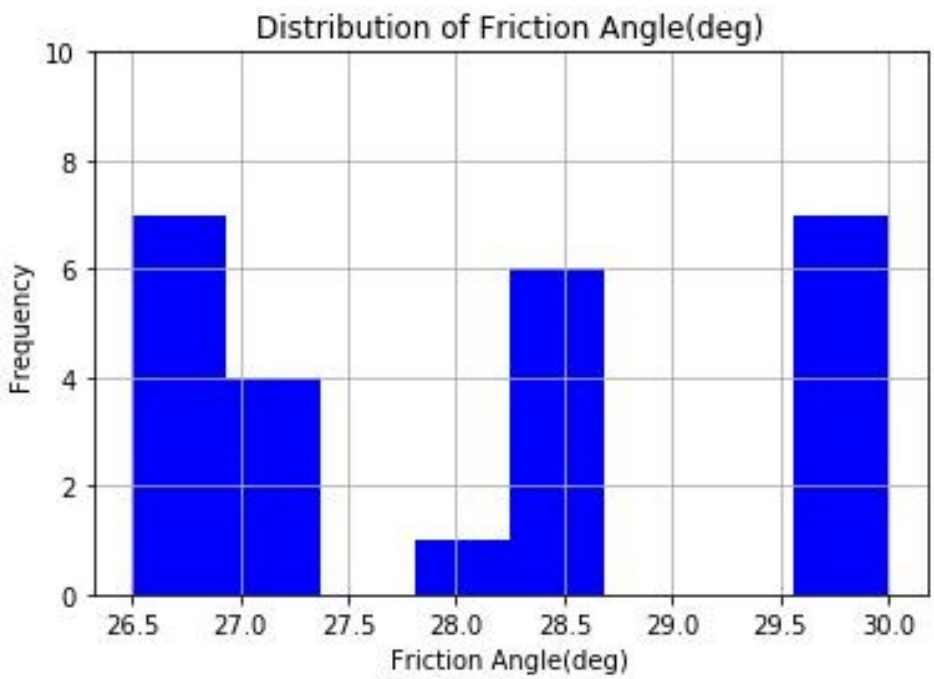


Figure 3.5j: Distribution of Friction Angle

Figure 3.5j shows the distribution of the friction angle, and it is positively skewed. This means that more of the data is to the left hand side of the distribution with a few large ones on the right. The mean value of the friction angle is seen to be 28.15 degrees.

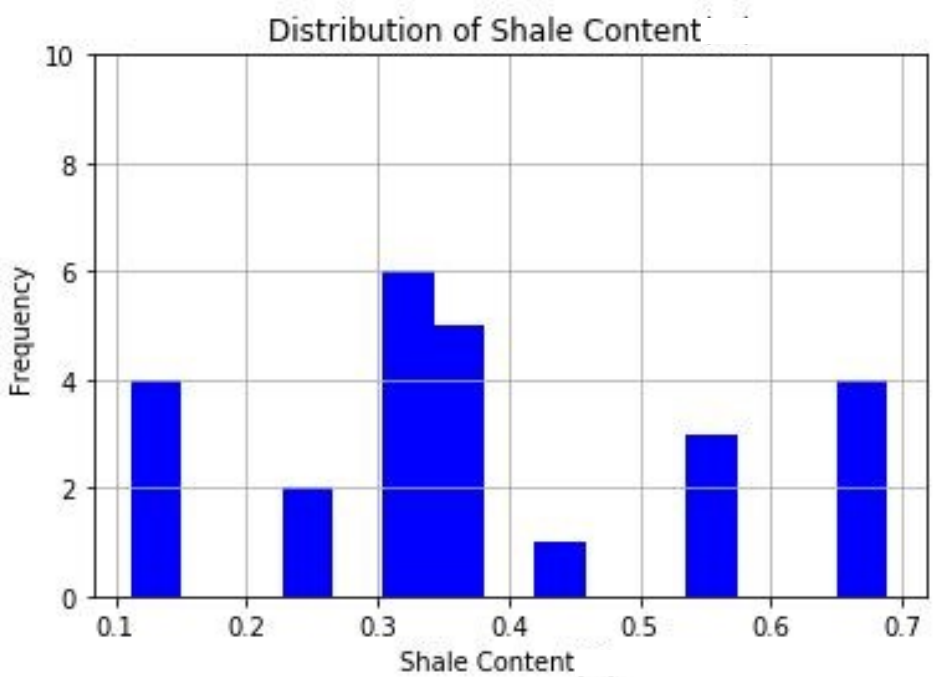


Figure 3.5k: Distribution of Shale Content

Figure 3.5k shows the distribution of the shale content, which is seen to be negatively skewed. The mean value of the shale content is 0.379.

3.4 Algorithm for training ANN with stochastic gradient descent.

The major steps in training the ANN algorithm used in this study are:

1. Initialize weights at random to small numbers near zero but not zero.
2. Using the features (x_i) for each input node, record the observations.
3. Forward propagation from left to right: The neurons are activated using the rectifier function. The more the neurons are activated, the easier the signal is passed across the neuron. The activation function (ϕ) is applied to the sum of each weight (w_i) and its corresponding feature (x_i). That is:

$$Activation\ function = \varphi(\sum_{i=1}^{m} w_i x_i) \qquad [3.1]$$

Apply the sigmoid function to the output layer (y_i). This allocates the probabilities for the different classes (sand or no-sand). That is:

$$P(y = 1) = \varphi(\sum_{i=1}^{m} w_i x_i) \quad [3.2]$$

4. Compare the actual results to the predicted and measure the error.
5. Based on the error, the weights are adjusted according to how each affects the error. The weights are then back-propagated from right to left. Also, the weights are adjusted based on the learning rate of the neural network.
6. After each batch of observations, repeat steps 1 to 5.
7. Repeat more epochs (e.g., 200 epochs). An epoch is when the whole training set is passed through the ANN.

4 Application of artificial neural network in this study

The preprocessed data is used for training and testing the ANN with 11 input variables, 2 hidden layers, and a binary output layer to predict whether a well will produce sand (0) or not (1). A rectifier activation function is used for developing the two hidden layers and the sigmoid activation function for the output layer. This is illustrated in the figure 3.6 below.

5 Application of Support Vector Machine (SVM) algorithm in this study

The steps used in applying the SVM to the Niger Delta input data are:

1. Data pre-processing (same as for ANN).
2. Feature scaling or data normalization (Same as for ANN).

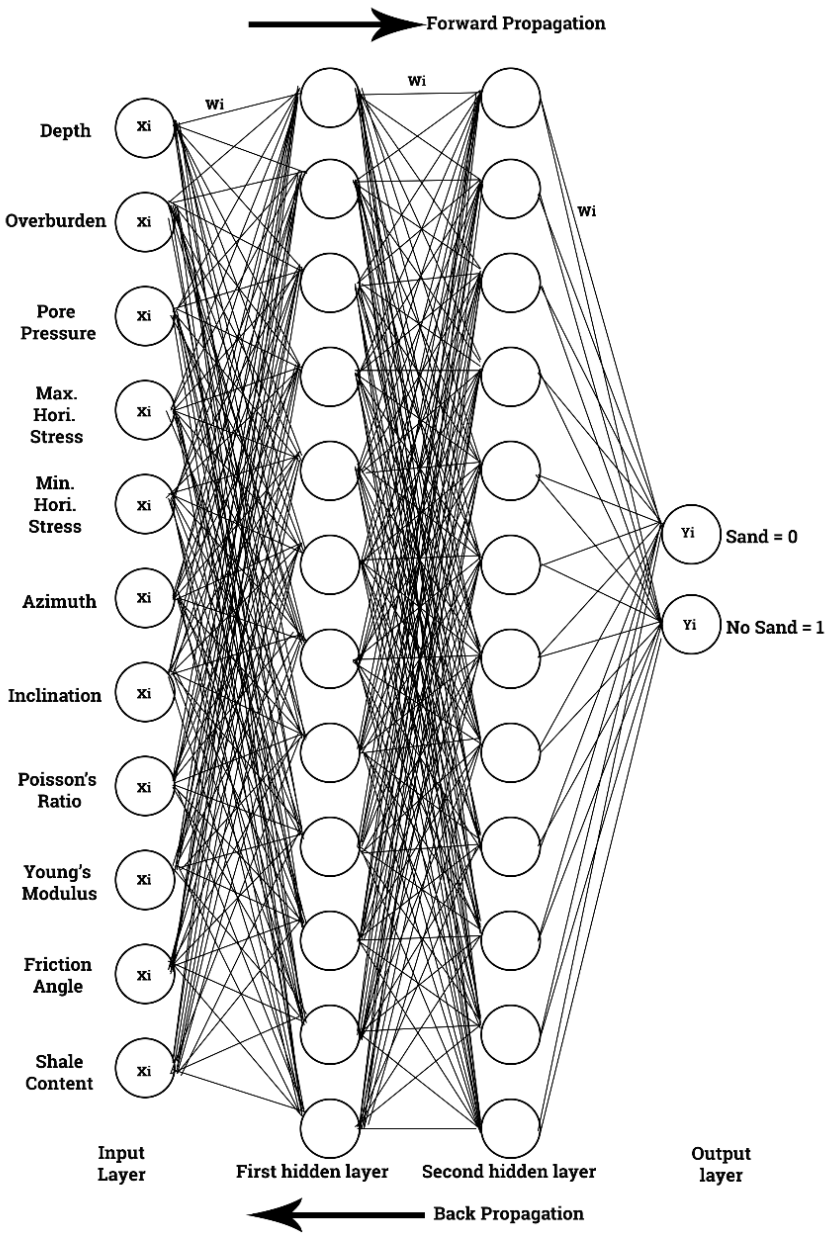
3. Fitting the SVM classifier to the training set by using the linear kernel given by:

$$Kernel = x^T y + c \qquad [3.3]$$

   Where c controls the tradeoff between the marginal and testing error, y are the outputs and x are the input parameters.

4. Predicting the test results.
5. Creating the confusion matrix.
6. Visualize training and test set results.

The performance of the ANN and SVM algorithms is evaluated in the following section. This would provide a quantitative measure for recommending the better algorithm for predicting sand production in the Niger Delta.



x_i – features

w_i – weights

y_i – outputs

Activation functions

Rectifier function:

$$\phi(x) = \max(x, 0)$$

Sigmoid function:

$$\phi(x) = \frac{1}{1 + e^{-x}}$$

Figure 3.6: ANN Architecture

6 Algorithm Evaluation Criteria

To evaluate the performance of the two algorithms, we used the following criteria: classification accuracy, confusion matrix, precision, recall, F1-Score, and the Cohen Kappa statistical measure. (Shin, 2020). These criteria are defined in the following section and explained in detail by Shin (2020).

1. **Classification Accuracy**: For a given dataset, we define,

$$Classification\ Accuracy = \frac{Total\ Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \quad [3.4]$$

2. **Confusion Matrix:** Confusion matrix not only tells us about the performance of the model, but which classes have been predicted incorrectly, correctly, as well as the errors made in the prediction. The Table 3.2 below is a confusion matrix of a two-class classification problem as used in this project.

Table 3.2: Confusion Matrix of two-class classification

Class	Positive Prediction	Negative Prediction
Positive Class	True Positive (TP)	False Negative (FN)
Negative Class	False Positive (FP)	True Negative (TN)

3. **Precision:** For a binary classification problem with two classes, we define,

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positves} \quad [3.5]$$

Results for precision are between 0.0 for no precision and 1.0 for full or perfect precision.

4. **Recall:** For a binary classification problem with two classes, we define,

$$Recall = \frac{True\ Positives}{True\ Positves + False\ Negatives} \quad [3.6]$$

The results for recall are between 0.0 for no recall and 1.0 for full or perfect recall.

5. **F1–Score:** Also known as F-Measure or F-Score is a combination of precision and recall folded into a single score. This is a harmonic mean of both precision and recall. Traditionally, the F-Measure is defined as

$$F_{Measure} = \frac{2 * Precision * Recall}{Precision + Recall} \quad [3.7]$$

6. **Cohen Kappa:** Cohen Kappa is a statistical measure that is used to measure the reliability of two raters who are rating the same quantity, and it identifies how frequently the raters agree. If there are N items that needed to be classified into C mutually exclusive categories, the work of Cohen Kappa is to measure the agreement between the two raters in order to classify N items to C.

7. **Loss function:** Also referred to as the mean squared error (MSE) for regression problems or Logarithmic Loss for binary and multi-class classification problems (Brownlee, 2017). This is calculated by:

$$MSE = (Actual - Predicted)^2 \quad [3.8]$$

Based on the proposed methodology, the model was validated using the processed data from a Niger Delta field and the following results were obtained.

Chapter 4: Results and Discussions

This chapter contains the results obtained from validating the models (ANN and SVM) discussed in the previous chapter. These results are discussed in detail and conclusions drawn. The discussion is presented in subsections. Firstly, the results obtained from ANN and SVM are displayed and discussed in the first two sections. The third section is a comparative study whereby the performances of both algorithms are evaluated based on the same evaluation criteria. From these, conclusions are drawn in the next chapter.

4.1 Results from ANN

To evaluate the algorithm performance, the input data was split into training and test data sets. For example, test size of 20% indicates that 80% of the data is used for training the ANN and 20% is used for testing. For this model validation the test size was varied from 20% to 80%.

4.1.1 Results of performance parameters for ANN.

Figure 4.1 shows a graphical representation of the results obtained from the ANN Algorithm. The performance parameters include: loss, accuracy, precision, recall, F1-score, and Cohen Kappa score. The results are also shown in Table 4.1.

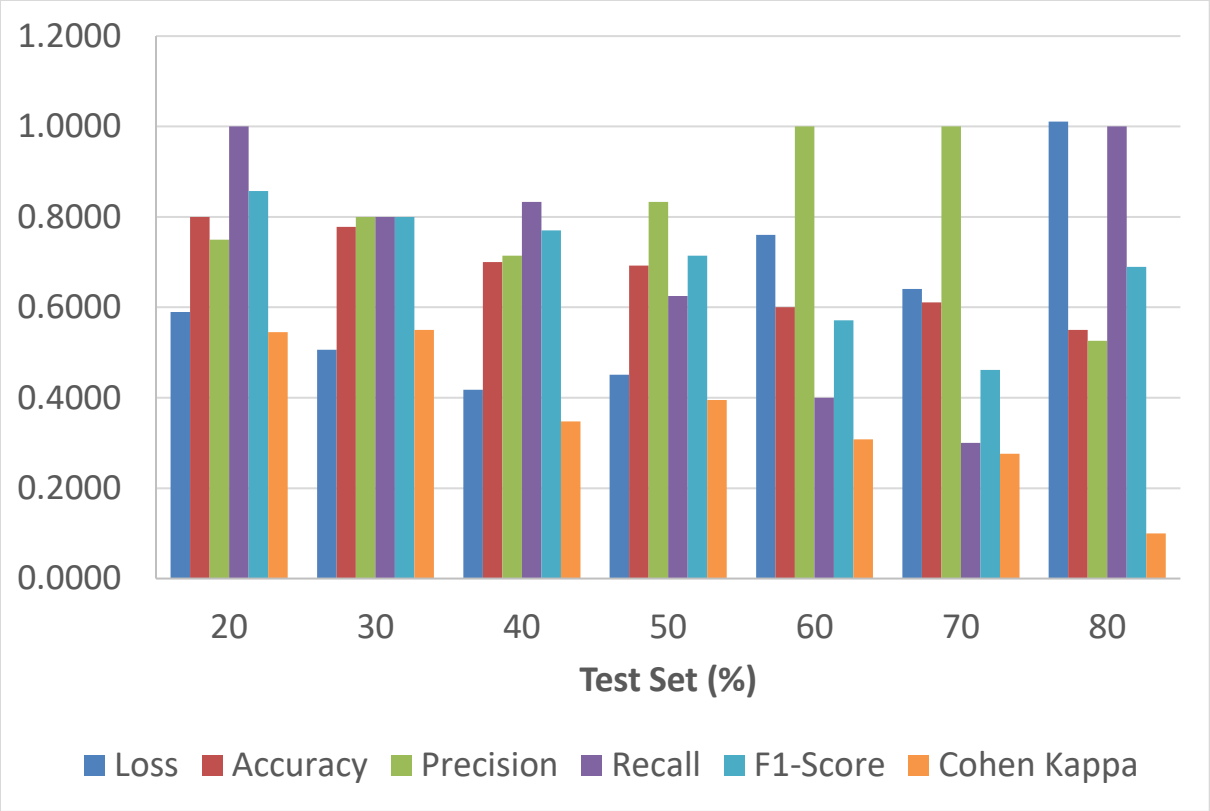


Figure 4.1: ANN performance at various test set percentages

Table 4.1: ANN performance results

Criteria	Test Size						
	20%	30%	40%	50%	60%	70%	80%
Loss	0.5900	0.5063	0.4179	0.4507	0.7601	0.6408	1.0111
Accuracy	0.8000	0.7778	0.7000	0.6923	0.6000	0.6111	0.5500
Precision	0.7500	0.8000	0.7143	0.8333	1.0000	1.0000	0.5263
Recall	1.0000	0.8000	0.8333	0.6250	0.4000	0.3000	1.0000
F1-Score	0.8571	0.8000	0.7699	0.7143	0.5714	0.4615	0.6897
Cohen Kappa	0.5455	0.5500	0.3478	0.3953	0.3077	0.2759	0.1000

From the results presented in Table 4.1, the following performance evaluation criteria are analyzed as follows:

The loss fluctuates from 41.79 to 101.11; this implies that the algorithm loss increases when we increase the percentage of data used for testing. This is a consequence of using small amount of the data for learning. The algorithm gets stuck at 80% test set as we observe a loss of above 100%. The accuracy measure decreases as the test size percentage increases in Table 4.1. Even though accuracy is a performance measurement tool, it can be observed that the classification between the wells that produce sand and those that do not is not equal. So, we observed an accuracy of 80% at the 20% test set percentage, whereas the "optimum percentage (that is, 30% test size), we have a lower percentage. Using the other evaluation tools such as precision, recall, f1-score, and Cohen Kappa score, it is observed that the highest values of these parameters conform with results reported in the literature (i.e., at optimal test set percentage of 30%).

The results displayed in Table 4.1 show that as the test size increases from 20% to 80%, the ANN performance becomes degraded and reduces in both accuracy and recall.

The graphical representations of the ANN results (for accuracy and loss) for various test set percentages with respect to the number of epochs can be found in the appendix. Recall, epoch is when the whole training set is passed through the ANN.

4.1.2 Results of the Confusion Matrix for ANN.

The confusion matrix of the ANN results is presented in Table 4.2. Note that the confusion matrix represents the results of the classification of a two-class problem. The interpretation of the results of the confusion matrix (True or False positive; True or False negative) for the test set percentage varying from 20% to 80% is displayed in the table.

Table 4.2:Confusion matrix and interpretation of results from ANN application case study

Test Set (%)	Confusion Matrix	Interpretation
20	[1 1] [0 3]	• One well was predicted to produce sand which actually produced sand (True Positive). • No well was predicted to produce sand which did not actually produce sand (False Positive). • One well was predicted not to produce sand which actually did produce sand (False Negative). • Three wells were predicted not to produce sand which actually did not produce sand (True Negative).
30	[3 1] [1 4]	• Three wells were predicted to produce sand which actually produced sand (True Positive). • One well was predicted to produce sand which did not actually produce sand (False Positive). • One well was predicted not to produce sand which actually did produce sand (False Negative). • Four wells were predicted not to produce sand which actually did not produce sand (True Negative).
40	[2 2] [1 5]	• Two wells were predicted to produce sand which actually produced sand (True Positive). • One well was predicted to produce sand which did not actually produce sand (False Positive).

		• Two wells were predicted not to produce sand which actually did produce sand (False Negative). • Five wells were predicted not to produce sand which actually did not produce sand (True Negative).
50	[4 1] [3 5]	• Four wells were predicted to produce sand which actually produced sand (True Positive). • Three wells were predicted to produce sand which did not actually produce sand (False Positive). • One well was predicted not to produce sand which actually did produce sand (False Negative). • Five wells were predicted not to produce sand which actually did not produce sand (True Negative).
60	[5 0] [6 4]	• Five wells were predicted to produce sand which actually produced sand (True Positive). • Six wells were predicted to produce sand which did not actually produce sand (False Positive). • No well was predicted not to produce sand which actually did produce sand (False Negative). • Four wells were predicted not to produce sand which actually did not produce sand (True Negative).

70	[8 0] [7 3]	• Eight wells were predicted to produce sand which actually produced sand (True Positive). • Seven wells were predicted to produce sand which did not actually produce sand (False Positive). • No well was predicted not to produce sand which actually did produce sand (False Negative). • Three wells were predicted not to produce sand which actually did not produce sand (True Negative).
80	[1 9] [0 10]	• One well was predicted to produce sand which actually produced sand (True Positive). • No well was predicted to produce sand which did not actually produce sand (False Positive). • Nine wells were predicted not to produce sand which actually did produce sand (False Negative). • Ten wells were predicted not to produce sand which actually did not produce sand (True Negative).

In summary, evaluating the confusion matrix of the ANN at the optimal test set percentage (30%), the ANN algorithm showed that; three wells were predicted to produce sand which actually produced sand (True Positive), one well was predicted to produce sand which did not actually produce sand (False Positive), one well was predicted not to produce sand which actually did produce sand (False Negative), and four wells were predicted not to produce sand which actually

did not produce sand (True Negative). Using a test size of 80% (i.e., 20% training set), the algorithm is observed to indicate that the ANN training has not been properly done to completion, and its difficulty in classifying the test set into the correct class is inevitable. Furthermore, the results show that the ANN seems to be working perfectly when we use a 70% to 30% of the data set as the training to testing ratio. This is a good fit as has been indicated by several researchers in the literature (Khamehchi et al., 2014; Shi ,2020; Brownlee, 2017; and Azad et al., 2011).

4.2 Results from SVM

4.2.1 Results of performance parameters for SVM.

The results of the SVM performance parameters include: loss, accuracy, precision, recall, F1-score, and Cohen Kappa score. Tables 4.3 shows the performance results obtained from the SVM. In this case, there was little or no degradation in the SVM performance for all test sizes.

Table 4.3:SVM Performance Results

Criteria	Test Size						
	20%	30%	40%	50%	60%	70%	80%
Accuracy	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9474
Precision	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Recall	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9091
F1-Score	1.0000	1.0000	1.0000	1.0000	1.0000	0.9500	0.9500

From Table 4.3, the performance results obtained by using SVM algorithm is 100% for all test set percentages except for the test sizes exceeding 70%, whereby the accuracy, recall, and f1-score drop by a few percentage below 100%. Despite this minor drop in efficiency, it is observed that the SVM is a very powerful tool for performing binary classifications as illustrated in this case study—“either the well will produce sand or it will not”

4.4.2 Results of the Confusion Matrix for SVM.

Table 4.4 shows the confusion matrix and the interpretation of results from the application of the SVM to the Niger Delta data set for a range of test set percentages.

Table 4.4:Confusion matrix and interpretation of results from SVM Application to case study

Test Set (%)	Confusion Matrix	Interpretation
20	[4 0] [0 1]	• Four wells were predicted to produce sand which actually produced sand (True Positive). • No well was predicted to produce sand which did not actually produce sand (False Positive). • No well was predicted not to produce sand which actually did produce sand (False Negative). • One well was predicted not to produce sand which actually did not produce sand (True Negative).
30	[5 0] [0 3]	• Five wells were predicted to produce sand which actually produced sand (True Positive). • No well was predicted to produce sand which did not actually produce sand (False Positive). • No well was predicted not to produce sand which actually did produce sand (False Negative). • Three wells were predicted not to produce sand which actually did not produce sand (True Negative).

40	[5 0] [0 5]	• Five wells were predicted to produce sand which actually produced sand (True Positive). • No well was predicted to produce sand which did not actually produce sand (False Positive). • No well was predicted not to produce sand which actually did produce sand (False Negative). • Five wells were predicted not to produce sand which actually did not produce sand (True Negative).
50	[5 0] [0 8]	• Five wells were predicted to produce sand which actually produced sand (True Positive). • No well was predicted to produce sand which did not actually produce sand (False Positive). • No well was predicted not to produce sand which actually did produce sand (False Negative). • Eight wells were predicted not to produce sand which actually did not produce sand (True Negative).
60	[7 0] [0 8]	• Seven wells were predicted to produce sand which actually produced sand (True Positive). • No well was predicted to produce sand which did not actually produce sand (False Positive). • No well was predicted not to produce sand which actually did produce sand (False Negative).

		• Eight wells were predicted not to produce sand which actually did not produce sand (True Negative).
70	[7 0] [0 11]	• Seven wells were predicted to produce sand which actually produced sand (True Positive). • No well was predicted to produce sand which did not actually produce sand (False Positive). • No well was predicted not to produce sand which actually did produce sand (False Negative). • Eleven wells were predicted not to produce sand which actually did not produce sand (True Negative).
80	[9 0] [1 10]	• Nine wells were predicted to produce sand which actually produced sand (True Positive). • One well was predicted to produce sand which did not actually produce sand (False Positive). • No well was predicted not to produce sand which actually did produce sand (False Negative). • Ten wells were predicted not to produce sand which actually did not produce sand (True Negative).

In summary, the results from the SVM showed that at an optimal test set percentage (30%, in Table 4.4) the algorithm yielded the correct classification, i.e., five wells were predicted to produce sand which actually produced sand (True Positive); no well was predicted to produce sand which did not actually produce sand (False Positive); no well was predicted not to produce sand which actually did produce sand (False Negative); and three wells were predicted not to produce sand which actually did not produce sand (True Negative).

4.3 Comparative Analysis of Performance of ANN vs SVM

The results of a comparative study between ANN and SVM algorithms are shown in Table 4.5. Table 4.6 shows the graphical representations of the results. At the various test set percentages, it is observed that SVM proves to be more effective than ANN for the cases examined in this study. The SVM yielded good results for all the performance evaluation metrics used in the validation.

Table 4.5: Comparative analysis of Performance of ANN and SVM

Criteria	Test Size									
	20%		30%		40%		50%		60%	
	ANN	SVM	ANN	SVM	ANN	SVM	ANN	SVM	ANN	SVM
Accuracy	0.8000	1.0000	0.7778	1.0000	0.7000	1.0000	0.6923	1.0000	0.6000	1.0000
Recall	0.7500	1.0000	0.8000	1.0000	0.7143	1.0000	0.8333	1.0000	1.0000	1.0000
Precision	1.0000	1.0000	0.8000	1.0000	0.8333	1.0000	0.6250	1.0000	0.4000	1.0000
F1-Score	0.8571	1.0000	0.8000	1.0000	0.7699	1.0000	0.7143	1.0000	0.5714	1.0000

Table 4.5 continued

Criteria	Test Size			
	70%		80%	
	ANN	SVM	ANN	SVM
Accuracy	0.6111	1.0000	0.5500	0.9474
Recall	1.0000	1.0000	0.5263	1.0000
Precision	0.3000	1.0000	1.0000	0.9091
F1-Score	0.4615	0.9500	0.6897	0.9500

Table 4.6: Comparison between ANN and SVM based on performance criteria.

The results of the comparative analysis of the performance of the two algorithms, indicate that the SVM is superior to the ANN especially for binary classification with sparse training data sets. This is a major finding from this study.

Chapter 5: Conclusion and Recommendations

The conclusions derived from this study as well as the proposed recommendations for future studies to improve upon this work are presented in this chapter.

5.1 Summary and Conclusions

In this study, two machine learning algorithms—ANN and SVM--are evaluated for their suitability for predicting sand production. The algorithms were implemented using 11 parameters identified to be impactful in predicting the onset of sanding in sandstone reservoirs. A comparative study to evaluate the performance of the algorithms was conducted using data from the Niger Delta. Based on the methodology and results of this study, the following conclusions are drawn:

1. The learning speed and accuracy for ANN depends on the number of input parameters and the number of hidden layers.
2. Both ANN and SVM algorithms show promising performance as tools for predicting sand production. The choice of the algorithm to use depends on the area of application.
3. The criteria of Precision, Recall, F1-score, and confusion matrix provided better performance measurement tools for evaluating machine learning algorithms.
4. The SVM algorithm is faster, takes less computing power, and less storage capacity compared to the ANN algorithm.
5. The SVM algorithm is quite effective for binary classification even for sparse training data set. The results indicated that SVM works relatively well even when we used 20% training set and 80% test set. For the case study, even at 80% test set, there was minimal reduction in

performance for the SVM algorithm, while the ANN algorithm showed a significant reduction in performance for the same 80% test set.

6. For the ANN algorithm, a loss of 1.0111 (that is over 100%) is observed at 80% test set, this suggest that the ANN algorithm malfunctioned. It can be concluded from this case that majority of the data set needs to be used for training and not for testing.
7. For the ANN algorithm, the Recall performance criterion can be improved by increasing the size of the training set (reducing test set size). In contrast, Precision can be improved by increasing the size of the test set data.
8. The SVM algorithm outperformed the ANN in terms of performance efficiency, and it should be the choice algorithm for developing ML tools for predicting sand production.
9. Based on data used for validation, the proposed machine learning tools are remarkably effective in predicting sanding as shown by the degree of accuracy, precision, and F1-score values obtained in this study.

5.2 Recommendations

The following recommendations are suggested to highlight areas of additional research to improve the validity of the work:

- Additional datasets from various depo-belts in the Niger Delta should be used in validating the developed algorithms.
- Other machine learning tools can be used to derive a more in-depth knowledge on the applications of this work. Such ML tools would include Particle Swam Optimization (PSO), Least Squares Support Vector machines (LSSVMs), Gene Expression Programming (GEP), and a host of other recent machine learning tools.

References

1. Agatonovic-Kustrin, S., & Beresford, R. (2000). Basic Concepts of Artificcial Neural Network (ANN) Modeling and its Application in Pharmaceutical Research. Journal of Pharmaceutical and Biomedical Analysis, 22(5), 717-727. Retrieved from https://doi.org/10.1016/S0731-7085(99)00272-1
2. Alghazal, M. (2015). Development and Testing of Artificial Neural Network Based Models for Water Flooding and Polymer Gel Flooding in Naturally Fractured Reservoirs. Pennsylvania State University , John and Willie Leone Family Department of Energy and Mineral Engineering. Pennsylvania : Pennsylvania State University . Retrieved February 24, 2021, from https://etda.libraries.psu.edu/files/final_submissions/11210
3. Brownlee, J. (2017). Machine Learning Mastery with Python. Retrieved from http://www.MachineLearningMastery.com
4. Coates, A., Lee, H., Ng, A.Y. (2011). An analysis of single-layer networks in unsupervised feature learning, in: Journal of Machine Learning Research.
5. Cortes, C., Vapnik, V., 1995. Support-Vector Networks. Mach. Learn. https://doi.org/10.1023/A:1022627411411
6. Geron, A. (2019). Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media.
7. Hamid Rahmanifard, T. P. (2018, April). Applicatioin of Fast Analytical Approach and AI Optimization Techniques to hydraulic Fracture Stage Pplacement in Shale Gas Reservoirs. Journal of Natural Gas Science and Engineering, 52, 367-378. doi:10.1016
8. Inui, M., Shoaff, W., Fausett, L., Schneider, M. (1994). The recognition of imperfect strings

generated by fuzzy context sensitive grammars. Fuzzy Sets Syst. https://doi.org/10.1016/0165-0114(94)90069-8

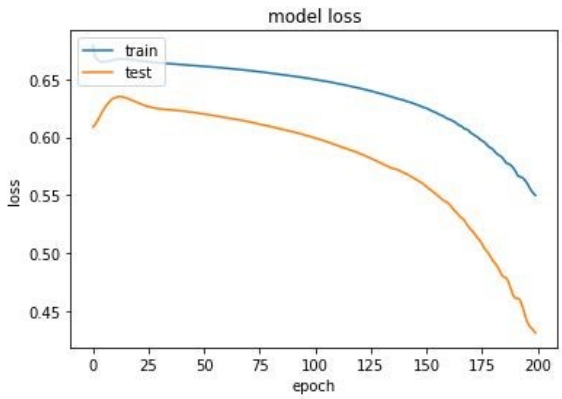
9. Mitchell, T.M. (1997). Machine Learning. Burr Ridge, McGraw Hill.
10. Ng, A. (2016). Machine Learning Yearning-Draft. Open Draft Report.
11. Ng, A. (2007). Advice for applying Machine Learning. I Can
12. Nielsen, M. (2015). Neural Networks and Deep Learning. Retrieved 11/6/2020, from http://neuralnetworksanddeeplearning.com
13. Olatunji, O.O., Micheal, O. (2017). Prediction of sand production from oil and gas reservoirs in the Niger Delta using support vector machines SVMs: A binary classification approach, in: Society of Petroleum Engineers - Nigeria Annual International Conference and Exhibition 2017, NAICE, 1247–1256. https://doi.org/10.2118/189118-ms
14. Osisanya, Samuel O. (2010). Practical guidelines for predicting sand production, in: Society of Petroleum Engineers - Nigeria Annual International Conference and Exhibition 2010, NAICE. https://doi.org/10.2118/136980-ms
15. Samuel, A.L. (1959). Some Studies in Machine Learning. IBM J. Res. Dev.
16. Sayadi, A.R., Tavassoli, S.M.M., Monjezi, M., Rezaei, M. (2014). Application of neural networks to predict net present value in mining projects. Arab. J. Geosci. https://doi.org/10.1007/s12517-012-0750-z
17. Shin, T. (2020, 05 01). Towards Data Science. Retrieved from Meduim: https://towardsdatascience.com/understanding-the-confusion-matrix-and-how-to-implement-it-in-python-319202e0fe4d
18. Smith, T. C., & Frank, E. (2016). Statisitical Genomics. In Introducing Machine Learning Concepts with Weka (pp. 353-378). Springer.

19. Udebhulu, D. O. (2014). Mechanistic Models for Predicting Sand Production: A Case Study of Niger Delta Wells. MSc Thesis, African University of Science and Technology, Abuja, Nigeria.
20. Udebhulu, D. O., & Ogbe, D. (2015). Mechanistic Models for Predicting Sand Production: A Case Study of Niger Delta Wells. Nigeria Annual International Conference and Exhibition. Lagos: Society of Petroleum Engineers. Retrieved April 6th, 2020.

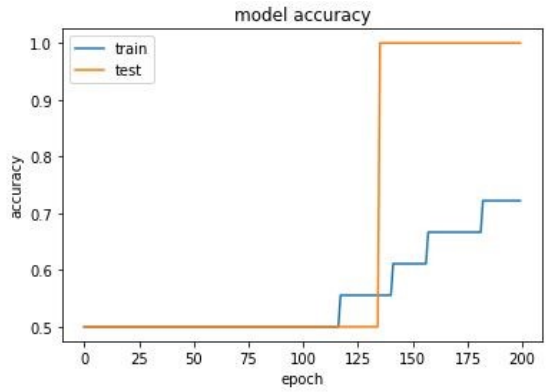
APPENDIX

Additional results on the accuracy and loss for application of the ANN algorithm are presented in this appendix. The graphs presented below show the various accuracies and losses plotted against the number of epochs for the training and test set data. Note, the term Epoch is when the whole training set is passed through the ANN.

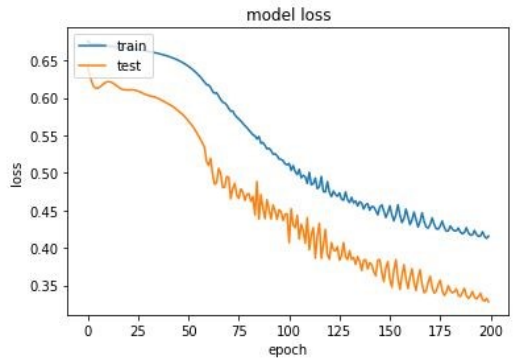
Appendix A



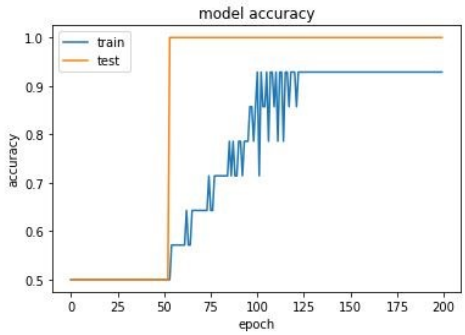
Appendix 1:Model Loss for 20% Test Set Using ANN



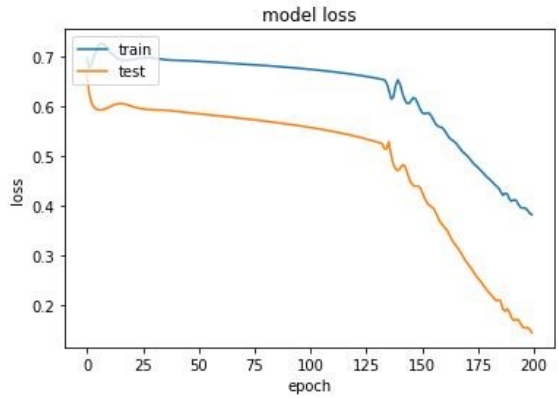
Appendix 2:Model Accuracy for 20% Test Set Using ANN



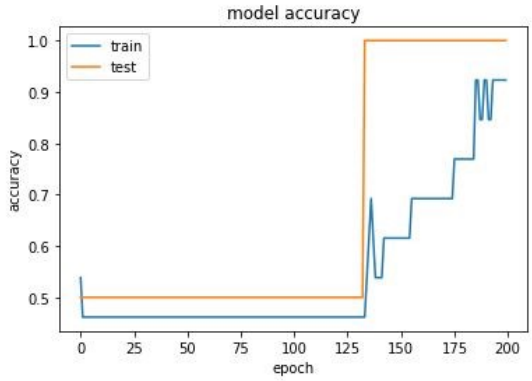
Appendix 3:Model loss for 30% Test Set Using ANN



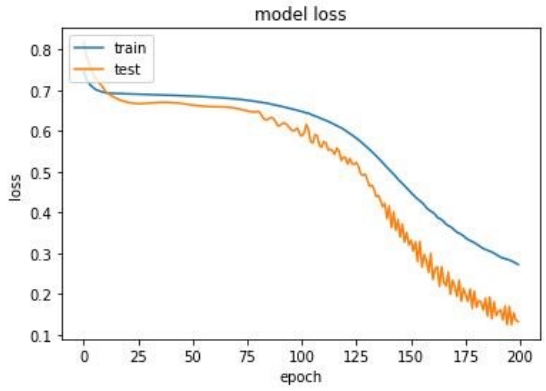
Appendix 4:Model Accuracy for 30% Test Set Using ANN



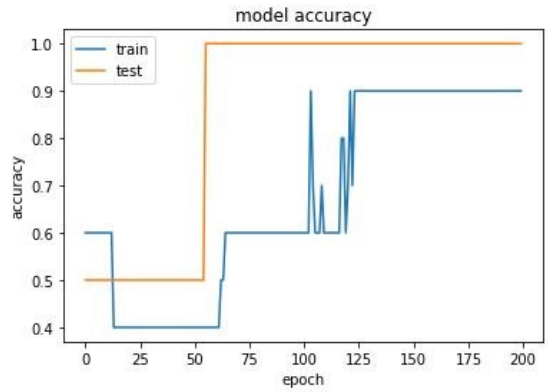
Appendix 5:Model Loss for 40% Test Set Using ANN



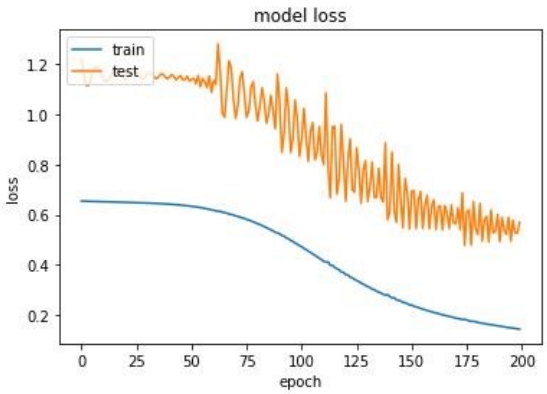
Appendix 6:Model Accuracy for 40% Test Set Using ANN



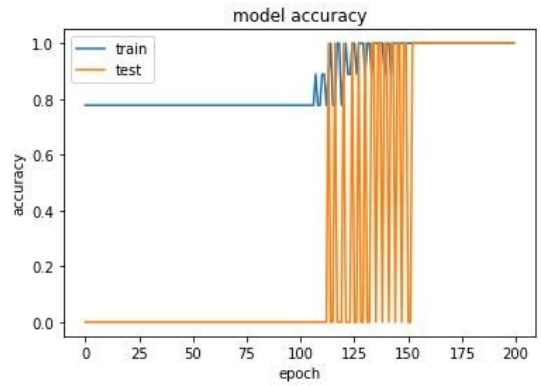
Appendix 7:Model Loss for 50% Test Set Using ANN



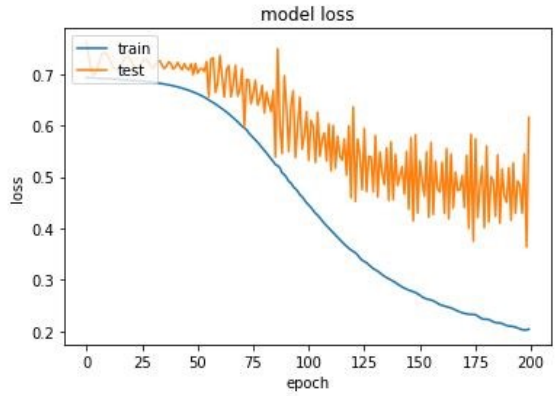
Appendix 8:Model Accuracy for 50% Test Set using ANN



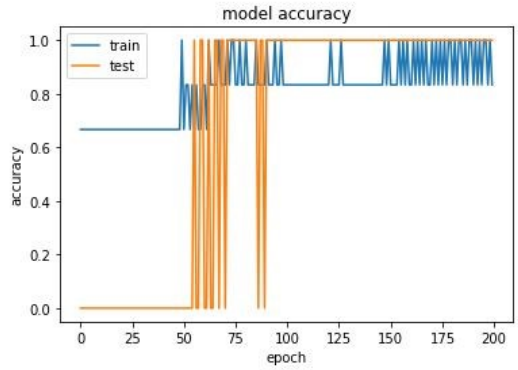
Appendix 9:Model Loss for 60% Test Set Using ANN



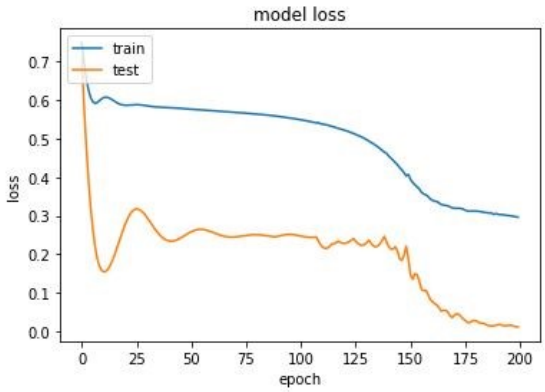
Appendix 10:Model Accuracy for 60% Test Set Using ANN



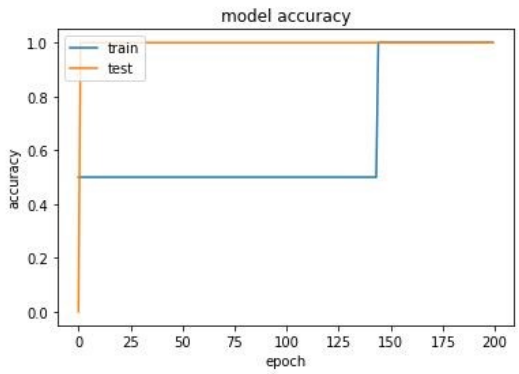
Appendix 11:Model Loss for 70% Test Set Using ANN



Appendix 12:Model Accuracy for 70% Test Set Using ANN



Appendix 13:Model Loss for 80% Test Set Using ANN



Appendix 14:Model Accuracy for 80% Test Set Using ANN