Roadheader performance prediction using Machine Learning Methods
Case Study: San Manuel Mine, Arizona

by

Askar Omirzak

THESIS SUPERVISOR

Saffet Yagiz

Thesis submitted to the School of Mining and Geosciences of Nazarbayev University in Partial Fulfillment of the Requirements for the Degree of
Bachelor of Science in Mining Engineering

Nazarbayev University
May 2025

ORIGINALITY STATEMENT

I, Askar Omirzak, hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at Nazarbayev University or any other educational institution, except where due acknowledgement is made in the thesis.

Any contribution made to the research by others, with whom I have worked at NU or elsewhere is explicitly acknowledged in the thesis.

I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.

Signed on 03. 05. 2025

_____

ABSTRACT

This thesis is dedicated to the development and evaluation of predictive models for the performance of roadheader machines (ICR) using machine learning algorithms under conditions of limited data. The scarcity of available datasets is primarily due to high collection costs, issues of commercial confidentiality, and heterogeneous geological conditions, which significantly complicates the application of traditional prediction models. To address this challenge, the study employs data synthesis techniques that expand the training set by generating artificial observations through the addition of Gaussian noise, as well as alternative approaches based on Ridge Regression and Random Forest methods. A comparative analysis of various models is conducted, including linear methods (Ridge, Lasso, ElasticNet), ensemble algorithms (Random Forest, Gradient Boosting, Extra Trees), and nonlinear approaches (SVR, MLP). The results demonstrate that ensemble methods achieve the highest prediction accuracy, as evidenced by high R^2 values and low MSE values, even when using synthetically expanded datasets. However, while data synthesis improves model performance, it does not fully replace real-world observations, necessitating further validation of the developed models under practical conditions. The findings hold practical significance for optimizing planning processes and economic evaluations in the mining and construction industries, and they point to the promising prospects of integrating data synthesis techniques with real-time monitoring systems to enhance the robustness and interpretability of predictive models.

ACKNOWLEDGMENT

I want to express my sincerest gratitude to my thesis supervisor, **Professor Saffet Yagiz**, for his invaluable guidance and support through this research. His expertise, encouragement, and constant feedback have been instrumental in shaping my work.

Also, I would like to thank my friends for their support and mentorship, especially – Ulan Sharipov, a person who giuded me through the world of Machine Learning.

Special thanks to Abylai Salimzhanov, Sultan Turan, Adel Kolesnikov, Alibek Abilgazym, Abylai Kubeyev, Yersultan Tursyn, Zhandos Dauletov, Arsen Kenzhebekov, and many others who was with me during my time at Pochinki and School

Also, Faculty Development Competitive Research Grant program of Nazarbayev University (Grant Number 201223FD8837) for funding this research is acknowledged.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

The mining industry is constantly evolving, adopting advanced technologies to improve efficiency, safety, and environmental sustainability. One of the key technological advances is the use of roadheader machines, versatile excavators equipped with a rotating cutting head that were originally developed for coal mining and are now widely used in various mining and tunneling projects due to their flexibility and accuracy.

The implementation of machine learning methods in predicting the performance of roadheaders is complicated by the limitation of available datasets, which are often small in size and do not fully cover all relevant parameters. Therefore, this study focuses on small dataset methods, aiming to develop robust and accurate forecasting models that can operate effectively even with a limited amount of data.

1.1 Background

Roadheader machines are highly specialized equipment for the precise and efficient crushing of rock, soil and other geological formations. Unlike traditional methods that rely on drilling and blasting, these machines use mechanical cutting heads equipped with carbide picks that continuously crush the material. This technology has revolutionized the fields of underground mining, tunneling and civil engineering, offering a more controlled and safer alternative to traditional mining methods.

The origins of roadheader machine technology date back to the mid-20th century Europe (Kogelmann & Schenck, 1982), when the need for mechanized excavation methods increased due to the limitations and dangers associated with conventional blasting methods. Roadheader machines were initially developed primarily for coal mines, where their excavation efficiency made them indispensable. Over time, improvements in machine design, the use of new materials and the development of automation have expanded the scope of application of roadheader machines, and today they are successfully used even in difficult geological conditions, including work in hard rocks.

Modern roadheader machines are equipped with automated control systems and modern sensors, which allows optimizing the drilling process and significantly increasing operational safety. These improvements ensure stable operation of the equipment, help reduce production costs and improve the quality of mining operations.

Advantages Over Traditional Excavation Methods

Traditional mining methods such as drilling and blasting present a number of operational challenges. Blasting generates significant vibration, noise and dust, which can have a negative impact on worker health and the surrounding rock mass. In addition, the unpredictability of blasting results in re-fracturing of the rock, which requires additional stabilization of the workings and increases the cost of transporting the material.

In contrast, roadheader machines provide a more controlled and continuous mining process. The use of a mechanical cutting head allows the workings to be formed with high accuracy, which minimizes excess material removal and reduces the requirements for working stabilization. As described by Sandbak (1985) - the ability to cut rock without excessive vibration makes roadheaders especially valuable for tunneling projects in urban areas and geologically sensitive areas, where it is critical to minimize structural disturbance. Ozdemir (1997) highlighted another important advantage of roadheaders - their ability to operate in varying geological conditions with minimal downtime. Unlike blasting operations, which require scheduled delays for loading and detonation, roadheader machines allow continuous drilling, which helps to reduce project deadlines and increase overall productivity. The advanced automation technologies implemented in these machines include real-time monitoring and adaptive control systems, which allow for prompt adjustment of operating parameters depending on rock characteristics and equipment specifications. These innovations increase the efficiency of the cutting process and make roadheaders the preferred choice for underground operations where high precision and reliability are required.

Figure 1: Comparison of Roadheader Machine Excavation with Drilling and Blasting in a Mining Environment (Krzystof &Piotr, 2019).

Roadheader Technical Components

Roadheaders are complex machines that consist of several key components that provide efficient excavation, mobility, and material handling. The main systems include the cutting head mechanism, hydraulic and electrical systems, and advanced navigation and control systems.

Cutting Head Mechanism

The cutting head is the primary excavation tool in a roadheader machine, responsible for breaking up rock and soil using mechanical force. It consists of a rotating drum equipped with multiple tungsten carbide picks strategically positioned to maximize cutting efficiency. The cutting process is accomplished through a combination of rotary motion and forward thrust, allowing the machine to effectively penetrate rock masses and break up their structure. Roadheaders are equipped with two main types of cutting heads: transverse and axial (longitudinal). Transverse heads have a horizontally oriented drum, which makes them particularly effective in soft rock, where wide, uniform cuts facilitate faster excavation. In contrast, longitudinal heads with a vertically rotating drum provide deeper penetration and high efficiency in hard rock, concentrating the cutting force on a smaller contact area. The choice of cutting head orientation is determined by the geological conditions of the excavation site and the requirements of the specific project.

The efficiency of a cutting head depends on several factors, including tooth geometry, material composition, and cutting force distribution. Proper tooth spacing is critical to optimizing rock fragmentation and preventing excess energy consumption. Research shows that improper tooth spacing can lead to uneven force distribution, accelerated tool wear, and reduced overall excavation efficiency. In addition, advances in materials science, such as the use of polycrystalline diamond coatings, have improved tool wear resistance and extended tool life.

Modern developments in automation and real-time monitoring have further improved cutting head efficiency. Sensor-based systems are now able to analyze cutting force data in real time, allowing automatic adjustments to the cutting parameters to optimize the process. Furthermore, the integration of water jet cutting technologies has proven effective in reducing cutting resistance, minimizing dust, and reducing the risk of cutting failure.

Hydraulic and Electrical Systems

The hydraulic and electrical systems of a roadheader machine play a critical role in its operational efficiency, providing the necessary power for excavation, maneuverability, and control. The hydraulic system is responsible for actuating the cutting head, controlling boom movement, and driving the machine's crawler tracks. High-pressure hydraulic cylinders adjust the position of the cutting head, providing precise adjustments that improve excavation efficiency. The responsiveness of the hydraulic system directly affects cutting performance, especially in hard rock conditions where greater force is required to penetrate the rock mass.

To optimize energy consumption, modern roadheader machines use variable displacement hydraulic pumps that dynamically adjust fluid flow based on cutting conditions. This allows hydraulic power to be used as efficiently as possible, reducing energy loss and extending the life of the machine. Additional advances in hydraulic drive design have improved machine stability, enabling the machine to maintain consistent cutting performance even in challenging geological conditions.

The machine's electrical system powers key components including the cutter head motor, integrated sensors and lighting. Programmable logic controllers (PLC) and advanced sensors are integrated into modern machine systems to improve accuracy and automate processes. Variable frequency drive systems dynamically change the cutter head speed, enabling the machine to adapt to changing rock conditions in real time. Remote monitoring capabilities provide continuous feedback on machine performance, facilitating predictive maintenance and reducing unplanned downtime.

Navigation and control systems

The integration of modern navigation and control systems has significantly improved the accuracy and efficiency of roadheader. Previously, traditional machines required skilled operators to manually adjust cutting parameters based on visual judgment and experience. However, modern machines are equipped with automated guidance systems that use geotechnical sensors, GPS, and laser scanning to improve cutting accuracy.

Geotechnical sensors integrated into the machine continuously measure rock properties such as uniaxial compressive strength (UCS) and abrasiveness. These sensors provide real-time data, allowing adaptive control systems to dynamically adjust cutting force, speed, and tooth rotation.

The use of 3D mapping technologies further improves cutting accuracy by allowing operators to visualize the process and ensure compliance with tunnel or excavation design requirements.

Remote control capabilities have also become an important feature of modern machines, improving the safety and efficiency of underground mining operations. Wireless communication systems allow operators to control the machine from the safety of operating rooms, reducing direct exposure to hazardous working conditions. In addition, machine learning algorithms are increasingly being used to analyze cutting patterns and optimize cutting parameters based on historical performance data. The future of roadheader machine control systems is moving towards full autonomy, where AI-based decision making and robotic drives have the potential to revolutionize underground mining methods.

According to Yin (2024), the integration of these advanced automation technologies enables modern roadheader machines to operate with higher excavation efficiency, improved safety, and enhanced adaptability to diverse mining and tunneling conditions. Continued development of automated, sensor-based control and monitoring systems is expected to further enhance the accuracy and reliability of these machines in the coming years.

Classification of Roadheaders

Roadheaders are complex equipment used for various purposes and working conditions, which determines the existence of various types and classifications of this equipment. In the literature, several key criteria for classifying roadheaders are distinguished: by the type of cutting head, machine power, movement method and area of application.

Classification by cutting head type

There are two types of roadheaders according to cutting head type:

Transverse head roadheaders: characterized by a cutting head located perpendicular to the axis of the machine (Figure). Such roadheaders are most effective when driving large-diameter tunnels and provide an accurate contour of the workings, which makes them suitable for the construction of transport and utility tunnels. Transverse cutting heads, commonly referred to as the ripping method, are devices adapted from continuous mining machines that rotate perpendicular to the boom axis. This method is particularly effective in soft rock, where high extraction efficiency and greater adaptability to changing geological conditions are achieved.

Transverse machines generate turning forces at right angles to the gripping force, which makes them more stable when cutting rock. The design of these machines allows them to cut rock with strengths up to 100 MPa (15,000 psi), with the most powerful models capable of working at strengths up to 150 MPa (22,000 psi). However, optimum performance is achieved at rock strengths of around 30 MPa (5,000 psi), making them ideal for coal mining, sedimentary work, and soft rock tunneling (Hemphill, 2012).

Axial head roadheaders: have a cutting head located parallel to the longitudinal axis of the machine. This type of roadheader is mainly used for narrow workings, as well as inclined and vertical passages, due to its high stability and lower vibration level. An axial cutting head, also known as an inline drilling head, rotates parallel to the boom axis. This design provides maximum forward cutting force, making it particularly effective in hard rock. Due to the lower cutting speed, axial machines consume fewer picks, which reduces wear and operating costs. They are often equipped with telescopic booms, which provide the necessary force for direct penetration into the rock mass. In hard rock conditions, axial machines are stabilized by hydraulic jacks or support arms, similar to the outriggers on a crane, which increases cutting stability. However, when working in soft rock, support arms may not be effective enough due to the low strength of the rock, and in wide tunnels, the fixed length of the booms limits maneuverability.

Table 1: General Comparison of Axial vs Transverse Roadheaders (Taylor & Francis Group, 2014)

Profile smoothness	**Favorable**	**Unfavorable**
Muck loading efficiency	Unfavorable	Favorable
Application limits	For UCS < 60–80 MPa, non abrasive	Soft to medium-strength rocks (UCS < 100–120 MPa), moderately abrasive
Production rate	Higher for UCS < 40–60 MPa	Higher for UCS > 60–80 MPa

Figure 2: Transverse Roadheader and Axial roadheaders.

Classification by weight

There are several approaches to classifying roadheaders by weight:

According to Tucker's classification (Tucker, 1985):

Light: weight up to 30 t, able to cut rock with a strength of up to 70 MPa. Medium: weight from 34 to 45 t, able to cut rock with a strength of up to 100 MPa. Heavy: weight over 45 t, able to cut rock with a strength of up to 120 MPa.

According to Atlas Copco – Eickhoff classification (Schneider, 1988):

Table 2: Atlas Copco – Eickhoff classification (1988)

Class	Weight (tons)
Class 0	less than 20 t
Class I	20 to 30 t
Class II	30 to 50 t
Class III	50 to 75 t
Class IV	over 75 t

According to Neil et al. (1994) classification:

Table 3: Neil et. Al. (1994) classification

Class	Weight (tons)
Small	less than 20 t
Medium	20 to 30 t
Large	30 to 50 t

Rock cutting mechanics of Roadheaders

The efficiency of roadheader machines is determined by the rock cutting mechanics, which include the interaction between the cutting tools and the rock surface. The ability of the machine to effectively crush rock depends on factors such as rock properties, cutting tool geometry, applied forces, and control systems. The cutting process is based on mechanical fragmentation, where the picks of the cutting head create local stresses in the rock, which leads to its destruction by crushing, shearing, or tensile tearing.

Rock destruction during machine operation occurs primarily through indentation, crack propagation, and chip formation. When the tooth of the cutting tool penetrates the rock surface, it creates local stress concentrations, leading to the formation of an initial crushing zone directly below the tool. As the penetration depth increases, radial and lateral cracks propagate outward, which ultimately leads to fragmentation of the material. The main failure mechanisms are:

- tensile failure, which occurs when tensile stresses exceed the tensile strength of the rock, causing crack propagation and fragmentation;

- shear failure, associated with displacement of material along shear planes, typical of plastic or layered rocks;

- compressive failure, which occurs when the applied forces exceed the compressive strength of the rock, which leads to its fragmentation - this mechanism predominates in hard and dense rock masses.

The machine's cutting performance depends on a number of geomechanical and operational factors. The uniaxial compressive strength (UCS) of the rock is one of the key parameters: soft rocks with a UCS below 50 MPa are easily fragmented, while hard rocks with a UCS above 100 MPa require greater efforts and specialized tooth configurations. Another important indicator is the Cerchar Abrasiveness Index (CAI), which measures the abrasiveness of the rock and its effect on tooth wear – highly abrasive rocks such as quartzite cause rapid tool wear, which increases maintenance costs and reduces cutting efficiency.

The choice of pick geometry and material is critical to cutting efficiency. Picks with a wide tip angle generate higher cutting forces, making them suitable for hard rocks, while picks with a sharp angle are better suited for soft rocks, reducing energy consumption and improving the degree of fragmentation. Modern picks are made of tungsten carbide, which has high wear resistance and impact resistance, and picks with a polycrystalline diamond coating are successfully used for cutting hard rocks, significantly extending tool life and reducing downtime (Huff, 1980).

In addition, performance is affected by operating parameters such as cutting speed, thrust force, torque and power consumption. Higher cutting speeds improve fragmentation but increase tooth wear, and excessive pushing forces can lead to premature tool failure. Higher torque allows deeper cutting, especially in hard rock conditions. Modern machines employ advances in automation, real-time monitoring, and adaptive control systems to optimize cutting efficiency. Machine learning algorithms analyze rock properties on the fly and adjust cutting force, tooth angle, and speed to achieve optimal efficiency. Water-jet-assisted cutting technology has been developed to reduce cutting resistance and minimize wear, and vibration-assisted cutting with ultrasonic vibrations can reduce required cutting forces by up to 30%, improving productivity in hard rock conditions (Grosso, 2014).

1.2 Objectives of the Thesis

1.2.1 To develop a predictive model for estimating the performance of Roadheaders using machine learning methods, taking into account rock properties, machine specifications, and operational parameters.

1.2.2 Conduct a comprehensive analysis of existing performance prediction methods and identify the most effective approaches for use in conditions of limited and small datasets

1.3 Justification of the Research

The mining industry is increasingly challenged by the need to improve operational efficiency and safety, particularly in the context of limited and heterogeneous datasets. Traditional models for predicting the performance of roadheader machines often fail to account for the nonlinear and multifactorial interactions present in complex geological conditions. This shortcoming results in unreliable productivity forecasts and can lead to significant operational delays and increased costs.

Furthermore, existing research was largely focused on large datasets, leaving a gap in methodologies that can effectively handle “small data” scenarios. By integrating advanced machine learning techniques with synthetic data generation, this research aims to bridge that gap. The innovative approach not only leverages the predictive power of ensemble algorithms and neural networks but also enhances the data landscape through carefully generated synthetic observations.

The practical benefits of this study include more accurate productivity forecasts, optimized equipment utilization, and a reduction in the economic risks associated with overestimating machine performance. Scientifically, this research contributes to the field by combining classical mining engineering principles with modern data-driven methods, offering a novel perspective that could be applied to a range of similar challenges in the mining and construction sector

1.4 Scope of Work

The primary objective of this research is to develop and validate a predictive model for estimating the performance of roadheader machines under conditions of limited data. To achieve this, the study will:

2. **Define the Research Objectives:** Focus on the development of machine learning models that incorporate synthetic data generation to overcome the constraints imposed by small sample sizes.

3. **Data Preprocessing and Feature Engineering:** Describe the methods used for cleaning the dataset, normalizing features, and expanding the feature space with polynomial transformations to capture nonlinear relationships.

4. **Synthetic Data Generation:** Evaluate three distinct methods—Ridge Regression-based labeling, Gaussian noise-based labeling, and Random Forest-based labeling—to augment the dataset, and assess their impact on model performance.

5. **Model Development and Validation:** Compare the performance of various predictive models, including linear models, ensemble methods, and neural networks, using standard evaluation metrics such as the coefficient of determination (R^2) and mean square error (MSE).

6. **Analysis and Comparison:** Analyze the robustness of each model and synthetic data approach, providing insights into the conditions under which the models perform optimally.

7. **Limitations and Future Work:** Discuss the inherent limitations of synthetic data generation and the challenges associated with small datasets, as well as propose potential directions for further research.

2. LITERATURE REVIEW

2.1. Introduction to the topic and Relevance of the study

With the development of the mining industry and large-scale construction of underground structures (transport tunnels, mine shafts, workings for laying communications), mechanized mining technologies are becoming especially important. One of the key types of equipment in this area is roadheaders. These machines allow for safer and more efficient work compared to the traditional drilling and blasting method. Due to their mobility and the ability to continuously extract rock, roadheaders reduce downtime, cut costs and increase productivity.

However, despite the obvious advantages, the actual cutting rate (Instantaneous Cutting Rate, ICR) can vary greatly depending on many factors: geological (strength, fracturing, abrasiveness of rocks), technical (type of machine, power, cutting head), as well as organizational and technological (shift schedule, availability of qualified personnel, ventilation scheme, etc.). Incorrect performance assessment can lead to schedule delays, budget overruns and increased risks on the project. Therefore, a reliable ICR forecast is one of the most important elements of planning and economic assessment of future work.

2.1.1 Problem statement: "small data" and complex geology

San Manuel Mine, due to it’s complex geological structure represents the classical problem faced by the researchers trying to develop ML models for mining purposes. Classical approaches to performance assessment or forecasting widely use empirical formulas (Bilgin et al. 1990, Copur et al. 1998, Thuro & Plinninger 1990, etc.), which take into account key rock parameters (UCS, RQD, RMR) and the characteristics of the machine itself. However, the versatility and accuracy of such models are often questioned. The main limitation is that each formula is obtained for specific conditions (a certain type of rock, class of equipment) and can give large errors outside the "native" range of values.

In recent years, machine learning (ML) and data mining methods have been increasingly used to solve engineering problems. For ICR prediction, the most interesting are artificial neural networks (ANN), support vector machine (SVM), decision tree, random forest, ensemble methods (boosting, bagging). These methods are better at “capturing” nonlinear dependencies and, given a sufficient set of examples, are capable of self-training, increasing the accuracy of forecasts.

However, the mining industry often faces the problem of “small data”: there are few real projects with a full range of measurements, mining and geological conditions are often different, and some information may be unavailable due to commercial secrecy and other restrictions. As a result, many ML models begin to overtrain and lose accuracy due to insufficient sampling.

Thus, the problem comes down to the need to create methods that can build reliable forecasts even with a small amount of input data, and also take into account the heterogeneity of geological and operational factors. In this context, methods of synthetic data extension, model regularization, and hybrid approaches (combination of empirics and ML) are of particular interest.

2.1.2 Purpose and objectives of the literature review

The main purpose of the review is to analyze existing approaches to predicting the productivity of roadheaders and identify key trends and “blank spots” in research.

To achieve this goal, it is proposed to solve the following tasks:

- Characterize the role and types of roadheader miners, showing their importance for modern underground construction and mining.
- Analyze modern approaches using machine learning, to evaluate their effectiveness and typical problems (in particular, small data sets).
- Formulate the main gaps in research that require more in-depth study or a new methodology.
- Justify the choice of a specific direction (for example, the use of ML algorithms in conditions of a limited sample) for further research work.

Structure of the further presentation
The literature review is divided into several subsequent sections:

- Section 2 briefly describes the evolution of roadheaders and the key technical parameters that affect their performance.

- Section 3 is devoted to the analysis of traditional empirical and statistical methods for predicting ICR.

- Section 4 considers the use of machine learning methods, including neural networks, SVR and ensemble approaches, as well as the specifics of application in mining.

- Section 5 contains a critical analysis of the state of research, discussing the issues of "small data" and variability of geological conditions.

- Section 6 summarizes the review, highlighting the directions that will form the basis for the methodological part and future experiments.

2.2 Key Parameters Affecting the Performance of Roadheader Machines

In the process of mechanized mining, the speed of rock extraction and the efficiency of work (Instantaneous Cutting Rate, ICR) depend on a combination of mining and geological, technical and organizational factors. In this section, the paper will consider the main parameters that, according to a number of studies (Bilgin et al. 1990, Copur et al. 1998, Ebrahimabadi et al. 2011, Seker & Ocak 2011, etc.), have the most significant impact on the result.

Geological and geomechanical factors

Uniaxial compressive strength (UCS): UCS is considered one of the main parameters for assessing the rigidity and difficulty of rock destruction. The higher the UCS value, the more difficult the cutting process is and the lower the potential productivity of the machine, all other things being equal. Various authors (e.g. Balci et al., 2004) point out that in most empirical models UCS can be included as a linear or power function to predict ICR.

Rock Quality Designation (RQD) and Classification Systems: RQD (Rock Quality Designation): reflects the degree of fracturing of the rock mass. With a high RQD, the rock is relatively solid, which complicates destruction; with a low RQD, the rock mass is more fragmented, and the excavation process can be accelerated.

Rock Mass Rating (RMR) or similar systems: take into account UCS, fracture frequency and orientation, joint condition, etc. The higher the class (good rock mass quality), the more difficult the mechanical cutting is. Therefore, in some models such as Bilgin et al. (1988), Ebrahimabadi et al. (2011), high RMR correlates with lower ICR values.

Analysis of oriented fracturing and soil parameters

The orientation of cracks (angle α between the working axis and the planes of weakening) can significantly change the nature of rock destruction. For example, with a favorable orientation of the layers, the cutting process is facilitated.

Abrasivity: the presence of quartz or other hard minerals in the rock leads to accelerated wear of the machine cutters, reducing efficiency and increasing tool costs.

Humidity and water saturation: If the massif is saturated with water, deterioration in the stability of the roof and side walls is possible. In addition, high water content sometimes reduces the strength characteristics of the rock, but, on the other hand, may require additional drainage and complicate the organization of work

Thus, these parameters - UCS, RQD, RMR, fracture characteristics, water saturation, etc. - have a complex effect on the instantaneous productivity (ICR). The more accurate the values obtained during engineering and geological surveys, the more reliable the forecast for the roadheader.

2.3 Machine learning and its implementation in the mining industry

In the last decade, machine learning (ML) technologies have been rapidly spreading in the mining sector, playing a significant role in the digital transformation processes. One of the key reasons for this has been the development of computing power and data collection tools. Processing large volumes of information coming from numerous sensors and automated monitoring systems has allowed ML analytical algorithms to find hidden nonlinear

dependencies and form accurate predictive models. For example, Mahdevari et al. (2014) studying the relationship between the geomechanical properties of rock and cutting speed, showed that neural networks can outperform classical regression methods in complex geology.

In the mining industry, machine learning is actively used in several main areas. Firstly, to predict the productivity of mining equipment (including roadheaders), where ANN models or support vector machines (SVM) look for patterns between rock characteristics (UCS, RQD, etc.) and the actual rate of advance. Secondly, to analyze the condition of equipment to prevent accidents (Predictive Maintenance). Such approaches use data from vibration diagnostics, acoustic and thermal sensors, which allows for early detection of failures and prevention of downtime. Thirdly, ML technologies are integrated into mining management systems: smart algorithms optimize the routes of road trains, distribute excavators and dump trucks, thereby increasing the overall efficiency of mining transport complexes.

Despite all the obvious advantages, such as the ability to identify deep nonlinear dependencies and quickly adapt to new data, machine learning methods face the problem of “small data”. In the mining sector, large datasets are often unavailable: measurements are taken in unique geological conditions, storage formats and levels of detail are different. This entails the risk of model overfitting and the impossibility of correct validation of the results. To solve such problems, various approaches are proposed in the scientific literature. Seker & Ocak, (2019) include regularization algorithms (dropout, L2-regularization), generation of synthetic data based on the statistical properties of the original samples, as well as combined approaches combining classical mining engineering models with learning algorithms

Thus, machine learning provides significant advantages in the analysis and forecasting of underground and open-pit mining processes. However, to achieve high accuracy and reliability of results, it is necessary to take into account both the technical features of the algorithms and the specifics of the geological data. In particular, to compensate for the lack of observations, ensemble learning (bagging, boosting) or transfer learning mechanisms can be used, when a model trained on one data set is further trained on another, similar in characteristics. Such a flexible approach provides a more universal solution for variable operating conditions of mining equipment (Ebrahimabadi et al., 2011).

2.4 Analysis of existing studies on ICR prediction for roadheaders

The first works by Sandbak (1985), attempted to relate the instantaneous cutting rate (ICR) of a roadheader and rock parameters (usually UCS, fracturing index) to the installed power. The models remained relatively simple and did not reflect the multifactorial nature of the cutting process in complex geology.

Gehring (1989) presented formulas for two types of roadheaders:

For a transverse roadheader (power ≈250 kW):

$$ICR = \left(\frac{719}{\sigma_c^{0.78}}\right) \tag{1}$$

For an axial roadheader (power ≈230 kW):

$$ICR = \left(\frac{1739}{\sigma_c^{1.13}}\right) \tag{2}$$

where σc is the uniaxial compressive strength of the rock (MPa). The higher the σc, the lower the ICR value.

Later, Bilgin et al. (1990) used more detailed empirical relationships. In particular, the study proposed the formula:

$$ICR = 0.28 \times P \times (0.974)^{RMCI} \tag{3}$$

where P is the cutting head power (hp). The RMCI index is calculated as follows:

$$RMCI = \sigma_c \times \left(\frac{RQD}{100}\right)^{2/3} \tag{4}$$

σc is UCS (MPa), RQD is the core integrity index (%). It was assumed that for a fixed power, an increase in RQD often leads to a decrease in ICR, since the rock becomes more integral.

In the works of Copur et al. (1998), the emphasis shifted to the machine weight and specific energy. The authors derived the relationship:

$$ICR = 27.511 \times exp(0.0023 \times RPI),$$

(5)

where RPI = (P × W) / σc, P is the power (kW), W is the weight of the machine (tons), σc is UCS. It was shown that more massive machines (with equal power) better stabilize the cutting process in hard rocks.

In the studies of Thuro & Plinninger (1999) using the example of a 132 kW machine, a trend relationship was found:

$$ICR = 75.7 - 14.3 \times ln(\sigma c)$$

(6)

where σc (MPa) is the uniaxial compressive strength. The model gave acceptable results in the given UCS range (about 30–100 MPa).

Since the early 2000s, ML technologies have become increasingly used (Yagiz et al., 2009; Mahdevari et al., 2014), but empirical methods have also continued to develop. Thus, Balci et al. (2004) consider the cutting depth (d):

For d = 5 mm:

$$ICR = 0.8 \times \frac{P^{0.37}}{\sigma_c^{0.86}}$$

(7)

For d = 9 mm:

$$ICR = 0.8 \times \frac{P^{0.41}}{\sigma_c^{0.67}}$$

(8)

where P is the power (kW), σc is the UCS (MPa). The value of 0.8 was determined empirically.

Also Keles (2005) proposed an expression for the MK2B (milling type) roadheader:

$$ICR = 163.93 \times \sigma_c^{-0.5737}$$

(9)

Among more modern variants, Ebrahimabadi et al. (2011) can be noted. The authors applied the rock mass brittleness index (RMBI) and took into account the orientation of the layers. One of the models looked like this:

$$ICR = 5.56 \times RMBI + 0.60 \times a - 8.17$$

(10)

where a is the bedding angle of the layers, and RMBI is calculated based on UCS, RQD and rock brittleness indices. Another formula relates ICR to the specific energy (SE):

$$ICR = -0.18 \times SE^3 + 28.57 \times SE - 92.82.$$

(2.11)

Earlier, the idea of the SE (specific energy) approach was also developed by Rostami et al. (1994), proposing:

$$ICR = k \times (\frac{P}{SE})$$

(12)

where k is the energy transfer coefficient (0.45–0.55 for roadheader), P is the power (kW), SE is the specific energy (kWh/m³).

Finally, Choudhary et al. (2017) can be mentioned, who reused the cubic dependence on SE, similar to Ebrahimabadi et al. (2011). This emphasizes the trend towards a more “energy” view of the rock cutting process.

Thus, the evolution of ICR models goes from the simplest dependencies such as Gehring (1989), to taking into account an extended set of rock-mechanical indicators – e.g, Bilgin et al., (1990), Copur et al. (1998), and then to complex indices of the massif structure as in work of Ebrahimabadi et al. (2011), and specific energy characteristics as in Rostami et al. (1994). Today, many authors combine these formulas with machine learning methods, which allows them to refine the coefficients and increase the reliability of models when expanding the range of conditions.

Key Works Focusing on Geotechnical parameters

A significant part of the early ICR prediction formulas was built around UCS (uniaxial compressive strength). For example, Gehring (1989), derived an equation for a 250 kW roadheader (transverse) in the form:

$$ICR = \frac{719}{{\sigma_c}^{0.78}} \tag{13}$$

and for 230 kW (axial):

$$ICR = \frac{1739}{{\sigma_c}^{1.13}} \tag{14}$$

where σ_c is UCS (MPa). An increase in σ_c leads to a decrease in ICR.

Another common parameter is RQD (Rock Quality Designation). For example, Bilgin et al. (1990) included in the formula:

$$ICR = 0.28 \times P \times (0.974)^{RMCI} \tag{15}$$

Where P is the cutting head power (hp). The RMCI index is calculated as follows:

$$RMCI = \sigma_c \times (\frac{RQD}{100})^{2/3} \tag{16}$$

At higher RQD (more solid rock), there is a tendency for productivity to decrease at a fixed power P.

The third basic parameter is RMR (Rock Mass Rating), which summarizes UCS, RQD, orientation and state of cracks. An increase in RMR indicates a higher quality rock mass, which complicates mechanical cutting and, accordingly, reduces ICR. The problem is the

ambiguity of RMR calculation methods: different authors can use slightly different scales and weighting factors.

Therefore, UCS, RQD and RMR form a triad, which is most often encountered in predictive models. However, there is no single universal equation that can accurately account for all the diversity of geology. Many researchers have to combine classical mining and mechanical formulas with correction factors or pay attention to the specifics of the equipment, such as the mass and design of the cutting head.

Comparison of Machine Learning Approaches

Since the early 2000s, machine learning (ML) methods have become increasingly popular, which are capable of analyzing the relationships between several dozen parameters at once. The most frequently mentioned are:

ANN (artificial neural networks),
SVR (support vector regression),
Random Forest,
Gradient boosting and other ensemble methods.

Seker & Ocak (2019) compared a range of such algorithms on a dataset that included UCS, RQD, machine weight, cutting head power, and actual ICR values. It turned out that ensemble methods (Random Forest, Gradient Boosted Trees) gave 5–10% higher accuracy than single models (e.g., simple neural network or linear regression).

Ebrahimabadi et al. (2011) used a hybrid approach, combining empirical formulas with the concept of RMBI (Rock Mass Brittleness Index). For example, one of the models:

$$ICR = 5.56\,RMBI + 0.6a - 8.17 \tag{17}$$

where a is the bedding or fracturing angle, and RMBI is calculated using UCS, RQD and brittleness indices. Further optimization of the coefficients was carried out using training algorithms (ANN, SVR), which allowed to increase the accuracy.

The main advantage of the ML approach is its flexibility and the ability to “learn” from data from different deposits. However, a high-quality result requires careful tuning of

hyperparameters and a sufficient number of observations, which is not always achievable in mining.

Experience with small data sets

Data limitations are one of the most pressing issues in building predictive ICR models. Performance measurements are usually taken over narrow intervals and are highly dependent on the unique local geology. This often leads to overfitting and reduced reliability of forecasts.

Salsani et al. (2013) and Seker & Ocak (2019) proposed several strategies to improve the situation:

Synthetic sample expansion: introducing “artificial” points around existing values (based on KNN).

Cross-validation (k-fold or Leave-One-Out): allows each observation to be used as a test one in turn, increasing the objectivity of the accuracy assessment.

Regularization: in neural networks, this can be dropout or L2 regularization, and in trees, a depth limit or a minimum number of observations per leaf.

Such techniques help keep the model from overfitting to noisy data and make it more robust to the variability of mountain conditions. The final results are usually better than those of classical empirical equations, which are designed only for a narrow range of input parameters.

2.5. Critical analysis and identified gaps

This part of the research aims to identify the gaps present in the previous works. Here, a brief analysis of the works aimed at empirical models, and more recent works on ML implication was performed.

Analysis of the accuracy and limitations of various models

Some researchers (Bilgin et al., 1990; Copur et al., 1998) emphasized the importance of a comprehensive accounting of mining and geological factors (UCS, RQD, RMR), but their formulas often show good convergence only in the initial data range. When trying to extend such empirical dependencies to other types of rocks or other types of tunneling machines (differing in weight, power), the accuracy drops significantly. Limitations in the generalizing ability - or, in the language of machine learning, in generalization - arise due to the fact that each model is "tied" to a specific set of conditions and features.

Another obstacle to comparing the results is the variety of initial data processing methods. Some authors may include rock abrasiveness in the calculation (CERCHAR, CAI), while others may not take this parameter into account at all. A similar problem concerns the method of measuring (or calculating) RMR: there are different "branches" of classification (Bieniawski 1973 vs. modifications of 1989, 1993, etc.), leading to variations in the massif assessment. As a result, direct comparison of models or their combination may be incorrect if the feature spaces are not consistent.

Disadvantages and advantages of research for "small data"

In mining, there are often situations when the set of empirical ICR measurement points is limited to tens of observations, and the measurements themselves may contain noise or errors. Under such conditions, even the most advanced machine learning algorithms risk overfitting, i.e. fitting the model to random fluctuations instead of stable patterns. In addition, there is often no independent test sample: all available data are used in training, and it becomes difficult to correctly assess the real accuracy of predictions.

Insufficient variability of geological conditions also plays a role. If, for example, the entire sample is taken from one mine or only one object, the model can "remember" these specific parameters. When transferred to another deposit (with different UCS, RQD, fracture orientation), the forecast accuracy decreases sharply.

Despite all the risks, research in the field of "small data" also has a positive side: it stimulates the search for adaptive or hybrid approaches, when basic empirical models and modern ML methods are combined. This can give a more practical result than universal formulas calculated for thousands of observations, because it is difficult or almost impossible to obtain so many observations in a real project.

The need for an integrated approach

Modern practice shows that one linear (or purely empirical) model is not enough to fully take into account all aspects that determine the instantaneous productivity of a roadheader. Many authors such as Ebrahimabadi et al., (2011), Seker & Ocak, (2019) emphasize the benefits of combining classical geomechanical equations of Bilgin et al. (1990), Copur et al. (1998), Rostami et al. (1994) with machine learning methods. An empirical formula or specific indices (RMBI, SE) provide the supporting “physical logic”, and the ML model refines the coefficients and better adapts to the local features of the massif.

Additionally, expert opinion (soft computing) can be used when working with incomplete or fuzzy data. For example, in conditions of a shortage of UCS, RQD and other key indicators, a fuzzy description is allowed. Expert systems based on the knowledge of geology specialists allow for high-quality model calibration where classical statistical learning is inapplicable due to a lack of numerical data.

Thus, the main problem identified in most studies is the lack of a single "universal" solution capable of producing stable results on any sample. The optimal path of development is a combination of several approaches: from traditional mining-mechanical equations to hybrid ML algorithms taking into account expert assessments. Such a comprehensive path gives a chance not only to improve the accuracy of forecasts, but also to ensure the adaptability of the model to new or poorly studied conditions.

2.6. Future research directions

Integration with additional technologies

Current research tends to connect roadheaders with complex sensor systems and real-time data analysis systems. Li & Gao (2021) showed that the use of laser scanners and 3D face visualization can significantly improve the accuracy of automatic assessment of the working geometry and, accordingly, influence the choice of the optimal cutting mode.

In parallel, the idea of continuous monitoring of machine parameters is being developed: Zhang et al. (2022) implemented a set of sensors to record acoustic signals and vibrations of cutters in order to record the transition from "normal" operation to potential emergency modes. Such data was integrated with machine learning algorithms that quickly update the instantaneous productivity forecast (ICR).

Given a sufficient volume of accumulated observations, a number of authors (Mahdevari et al., 2014; Ghasemi et al., 2020) talk about the prospects of deep learning. For example, LSTM or transformer architectures can use time series recorded during mining. So far, these methods remain relatively rare in mining due to the difficulties in generating "big data", but with the advent of expanded datasets, the result can be very promising.

Strengthening the robustness of ML models

Many studies such as Chang & Peng (2020) emphasize the need for regularization and optimization of models to cope with noisy and heterogeneous mining data. One of the key

techniques is strict cross-validation (k-fold), which allows for a reliable assessment of the generalization ability of the algorithm with a small number of observations.

An important direction is the introduction of hybrid methods. For example, Seker & Ocak (2019) proposed ensembles (bagging, boosting) for working with a small set of training data. Other authors (Salsani et al., 2013; Ghasemi et al., 2020) are experimenting with combining classical neural networks and evolutionary algorithms (genetic algorithms, particle swarm optimization) for selecting hyperparameters. This strategy makes it possible to quickly find optimal settings for network depth, activation function, etc., reducing the risk of overfitting.

Bayesian regression approaches deserve special mention, allowing one to estimate prior distributions of UCS, RQD, or other features when the data is too sparse (Li et al., 2022). Such models not only predict ICR, but also generate a probability interval, which provides a more complete picture of uncertainty.

Model transferability

Transferability (or transfer learning) becomes an important factor when a model trained on one field must quickly adapt to other geolocations. Zhang et al. (2022) experimentally showed that if you have a “base” neural network for a narrow range of UCS, then using an additional 10-15 points from a new field, you can fine-tune the network and achieve accuracy comparable to “pure” training on a large dataset.

Alternatively, Chang & Peng (2020) point to the idea of forming a “universal” set of features (e.g. UCS, RMR, CAI, machine weight, power, cutter head type), standardized by measurement methods. This will allow researchers from different regions to “put” data into a common database and obtain more universal ML models. In the long term, such a step will accelerate the implementation of drilling optimization algorithms already at the design stage, where an engineer will be able to pre-calculate the expected ICR without expensive field tests.

To summarize, further research should cover at least three areas of development: integration with real time via sensors, enhancing the robustness of ML models using regularization and hybridization, and increasing portability through transfer learning and unification of methods for measuring mining and geological parameters.

2.7. Literature Review Summary

Summary of Key Results

The analysis of existing works on forecasting the instantaneous productivity (ICR) of roadheaders revealed that the development of methods went from simple empirical formulas such as Gehring et al. (1989), Bilgin et al. (1990), Copur et al. (1998), taking into account mainly UCS and RQD, to more complex models reflecting many factors: crack orientation, abrasiveness, design features of the miner itself. At the same time, machine learning methods (ANN, SVR, ensemble approaches) are gaining popularity, which are able to better capture nonlinear relationships in data and expand the applicability of models.

At the same time, an important condition for high-quality ICR forecasting remains the presence of a representative observation base. However, in mining practice, the problem of small data volumes is often encountered, which leads to difficulties in training models and in assessing their generalizing ability. To overcome these limitations, synthetic data generation, strict cross-validation and regularization are used.

Justification of the Selected Scientific Direction

Summarizing the results of the review, several critical gaps can be identified. Firstly, most empirical formulas (Bilgin et al. 1990, Copur et al. 1998, Thuro & Plinninger 1990) are designed for a narrow range of geological conditions or for certain types of machines, which means that their accuracy drops sharply when changing a deposit or when working with a new type of roadheader. Secondly, even multifactor models (taking into account RMR, RMBI, SE) remain vulnerable to noise data and fluctuations in geological properties.

On the other hand, machine learning shows promise with sufficient variability and quantity of data. However, “small data” complicates the use of deep neural networks and requires improved techniques to combat overfitting. Therefore, it is logical to move towards an integrated approach, in which nonlinear features are added to classical mining and mechanical parameters, and regularization and synthetic generation mechanisms are incorporated into the model.

Transition to methods and experimentation

Based on the identified gaps in the study, a methodological part will be proposed, where the main focus is on methods for processing limited data sets and expanding the feature space. In particular:

It is planned to use procedures for generating artificial observations (data augmentation) to increase the model's resistance to noise.

It is envisaged to introduce second-order polynomial features reflecting potential nonlinear interactions (e.g., UCS × RQD, RMR^2, etc.).

Various machine learning algorithms (neural networks, SVM, ensembles) will be used to train the models, to which a set of regularization measures and strict cross-validation are added.

Therefore, the next chapter will detail the methodology based on the results of the literature review: what data features are taken into account, how the hyperparameters of ML models are selected, and why the proposed techniques are especially important for predicting ICR in a limited sample.

3. CASE STUDY – SAN MANUEL MINE, ARIZONA USA

The San Manuel Mine, located in Arizona's Lower San Pedro River Basin, was a significant underground copper mine and a cornerstone of the region's mining industry. Operational from the 1950s, the mine utilized various methods, including block caving, open-pit mining, and in-situ leaching, to adapt to changing technological advancements and resource extraction demands. However, due to economic constraints and resource depletion, the mine ceased operations in 1999 (BHP Copper Inc., 2002).

Geologically, the mine is characterized by an ore body hosted in granodiorite porphyry and quartz monzonite, intersected by fault systems such as the San Manuel, East, and West Faults. These structural features provided pathways for mineralization but also posed operational challenges due to zones of weakness and variability in rock strength. The depth of ore deposits, which extended from 700 to 3,000 feet below the surface, further complicated mining operations (Arizona Geological Society, 1987).

In response to these challenges, roadheader technology was tested and utilized at the San Manuel Mine to enhance drift excavation. Roadheaders, such as the DOSCO SL-120, were employed to overcome the limitations of conventional drill-and-blast techniques, particularly in fractured and jointed rock environments. The technology demonstrated its ability to achieve faster excavation rates, reduced overbreak, and improved stability of drifts. Tests conducted in the San Manuel and Kalamazoo ore bodies revealed that roadheaders could be a viable

alternative, particularly in areas where structural complexities and variable rock properties necessitated precision excavation (Sandbak, 1985).

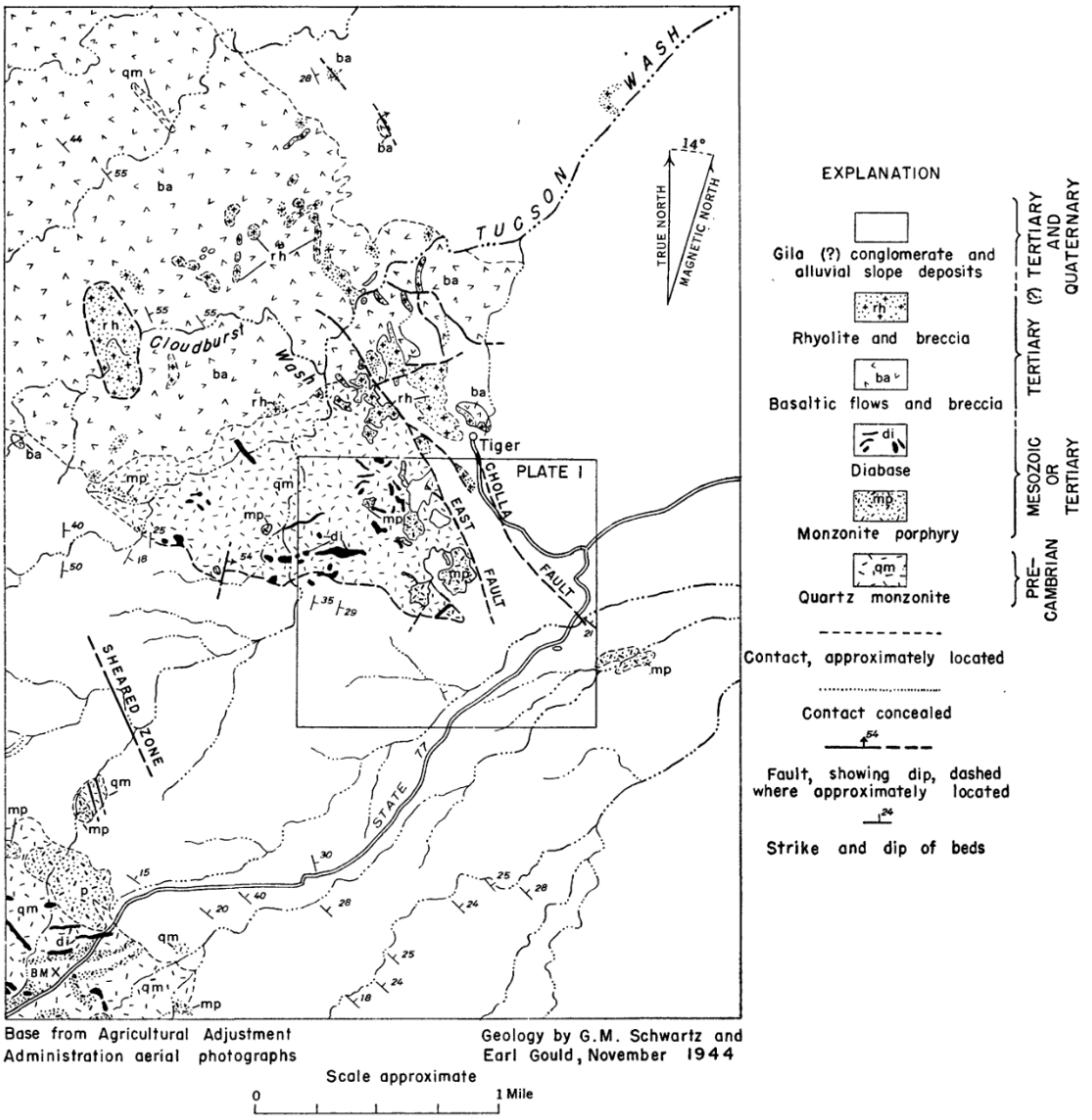


Figure 3. Geologic Map of San Manuel area (Schwartz, 1953).

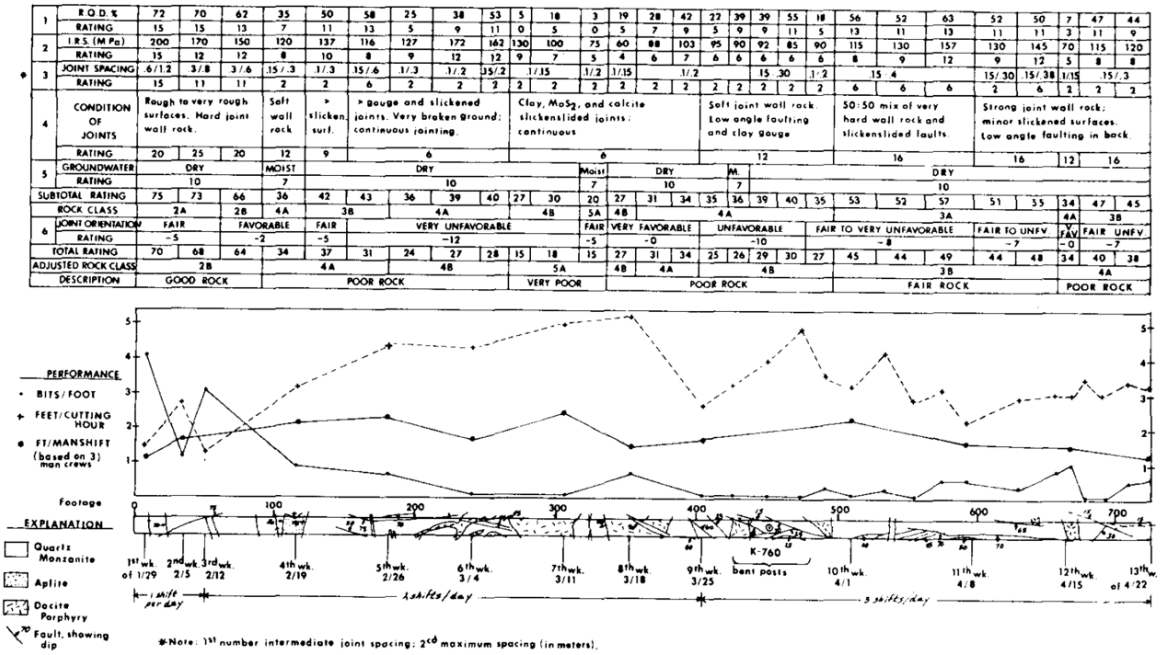


Figure 4. RMR table and Machine Performance Graph

Table 4: Dataset Utilized for Model Establishing

#	Geology Description	RQD (%)	RMR (-.)	UCS (MPa)	ICR (m^3/hr)
1	Good Rock	72	70	200	8,309
2	Good Rock	70	68	170	6,492
3	Good Rock	62	64	150	8,547
4	Poor Rock	35	34	120	12,115
5	Poor Rock	50	37	137	15,016
6	Poor Rock	58	31	116	17,406
7	Poor Rock	25	24	127	18,789
8	Poor Rock	38	27	172	18,491
9	Poor Rock	53	28	162	19,199
10	Very Poor Rock	5	15	130	20,06
11	Very Poor Rock	18	18	100	21,099
12	Very Poor Rock	3	15	75	21,807
13	Poor Rock	19	27	60	22,167
14	Poor Rock	28	31	88	18,997
15	Poor Rock	42	34	103	13,809
16	Poor Rock	22	25	95	12,317
17	Poor Rock	39	26	90	14,357

18	Poor Rock	39	29	92	16,567
19	Poor Rock	55	30	85	19,45
20	Poor Rock	18	27	90	16,446
21	Fair Rock	54	45	115	13,952
22	Fair Rock	52	44	130	14,665
23	Fair Rock	63	49	157	12,182
24	Fair Rock	52	44	130	11,044
25	Fair Rock	50	48	145	12,596
26	Poor Rock	7	34	70	12,79
27	Poor Rock	47	40	115	13,151
28	Poor Rock	44	38	120	14,193

Roadheader Operational Features

The San Manuel mine has served as a testbed for roadheader technology. Field studies, including those conducted by Sandbak, have demonstrated that roadheaders significantly reduce the risk of fracturing and excessive rock failure compared to conventional blasting methods. In the Kalamazoo and San Manuel deposits, roadheading achieved development rates up to 38% higher, especially in weak, highly fractured, and hydrothermally altered rocks.

Detailed geomechanical mapping of test levels (e.g., 2890 and 2375 ft tests) revealed that different rock classes, from weak, highly fractured porphyry to massive quartz monzonite, have a direct impact on cutting rates (in ft/hour) and cutting tool consumption (bits per foot). The results of these tests were used to calibrate the predictive models, confirming that ensemble machine learning methods are particularly effective in describing the complex nonlinear behavior of rock masses during roadheader operations.

The San Manuel case study demonstrates the practical benefits and challenges of applying predictive models to the mining industry. Heterogeneous geological settings, complex structural settings, and associated sedimentation require advanced data synthesis techniques and robust ensemble models. The analysis suspects that ensemble methods significantly outperform simple linear regressions in predicting roadheader performance in such dynamic environments. In addition, geomechanical observations provide an important link between model output and real mining conditions, which is critical for optimizing mine planning, reducing costs, and ensuring mine safety.

4. METHODOLOGY

This chapter describes the methodological approach used to solve the problem of predicting the ICR (roadheader machine performance) using machine learning methods. The stages of data collection and description, their pre-processing and feature formation, methods for generating synthetic data, as well as the selected models and the strategy for their evaluation are considered.

4.1 Data collection and description

The research on tunneling and drift excavation at San Manuel, performed by Louis Sandbak in 1985, contains a complete dataset with the geotechnical parameters for model establishment required. This dataset includes measurements of such parameters as rock quality factor (RQD), RMR value, and Uniaxial compressive strength (UCS), as well as the target variable – ICR, characterizing the equipment productivity in m^3/hour. The section describes the data source, the conditions for collecting measurements, as well as the main characteristics and statistics of the dataset. Such a detailed analysis allows us to understand the original distribution of values and identify potential problems, such as the presence of missing values or outliers.

4.2 Data Preprocessing and Feature Generation

In this study, data preprocessing and feature generation plays a key role, since the success of subsequent modeling depends on the quality of the input data. In the first step, the data is carefully extracted from the original dataset: values from the RMR table and Machine Performance graphs were extracted and digitalized. It’s important to carefully handle data and create a reliable base for further analysis.

The next important step is data normalization. In the original dataset, features are measured in different units (for example, RQD is expressed in percentages, UCS in MPa, and ICR in cubic meters per hour), which can lead to machine learning algorithms paying too much attention to features of a large scale. To address this issue, standard scaling was applied, transforming all features so that their mean is zero and their standard deviation is one. This not only speeds up the convergence process of the algorithms, but also ensures that all variables have an equal influence on the final result.

In addition, to enable the models to capture more complex, nonlinear relationships between features and the target variable, the original feature space was expanded using a polynomial

transformation. Specifically, combinations of second-degree features were generated for each original variable, revealing hidden patterns not reflected in the original data. Although adding polynomial features increases the dimensionality of the dataset, the benefits in terms of the improved ability of the model to describe complex dependencies significantly outweigh the potential computational costs. Thus, the complex preprocessing includes: removing incomplete records to ensure data purity, normalizing features to bring them to a common scale, and creating additional polynomial features to reveal nonlinear relationships. This approach forms a solid foundation for subsequent stages of analysis, synthetic data generation, and model training, contributing to increased accuracy and reliability of predictions.

4.3 Synthetic Data Generation

Due to the limited size of the original data set, synthetic data generation methods are used to improve the learning ability of models. Three methods are described below, each of which is mathematically justified.

1. The Ridge Regression-based Label Method.

The idea of this method is to use a linear model with L2 regularization to predict the target variable on a bootstrapped sample of features. Mathematically, the Ridge regression problem is formulated as the minimization of the loss function:

$$|| X\beta \ - \ y||^2 \ + \ \lambda \, ||\beta||^2 \tag{18}$$

where X is the feature matrix, y is the vector of target values, β is the vector of model coefficients, and λ is the regularization coefficient.

After training the model on the original data, synthetic labels are calculated for the bootstrapped (randomly selected with repetition) subsample X_{boot}:

$$y_{synthetic} == X_{boot} \, \hat{\beta}. \tag{19}$$

To simulate natural variability, a small amount of noise is added to the features, for example,

$$Xboot' = Xboot + \varepsilon,$$

(20)

where $\varepsilon \sim N(0, \sigma^2 I)$. This approach preserves the linear dependencies present in the data and allows for an expanded training sample.

2. Gaussian Noise Generation Method.

This method assumes that synthetic labels are generated by adding random noise to the target values obtained by resampling. Formally, for a resampled target variable $y_{resample}$, the synthetic label is defined as:

$$y_{synthetic} = y_{resample} + \epsilon,$$

(21)

where the noise ϵ\epsilonϵ is normally distributed, $\epsilon \sim N(0, \sigma^2)$. Here, the parameter σ\sigmaσ controls the degree of random fluctuations, allowing the synthetic data to mimic the natural fluctuations found in real measurements.

3. Random Forest Regressor -based method.

Using the Random Forest model, an ensemble of decision trees is trained to predict the target variable. Mathematically, Random Forest estimates a function $f(x)$ such that:

$$y \approx f(X) = \frac{1}{T}\sum_{t=1}^{T} f_t(X),$$

(22)

where T is the number of trees and $f_t(X)$ is the prediction of the t-th tree. After training, for the bootstrapped feature set X_{boot}, the synthetic labels are computed as:

$$y_{synthetic} = f(X_{boot}).$$

(23)

Similarly to the first method, a small noise can be added to the features to increase diversity:

$$X'_{boot} = X_{boot} + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I).$$

(24)

This approach allows us to take into account complex nonlinear dependencies present in the data and generate synthetic labels that are as close as possible to real ones.

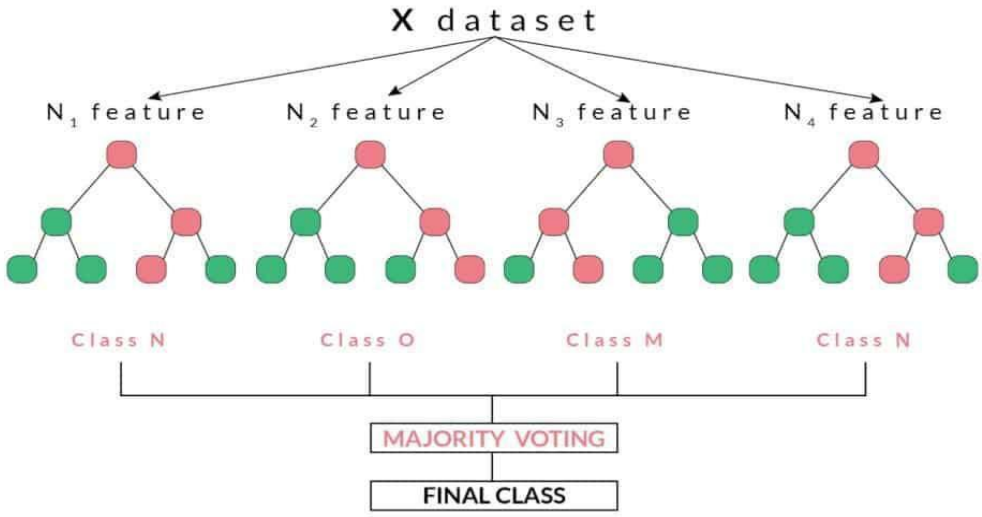


Figure 5: Visualization for Random Forest Regressor Labeling

In this study, a wide range of machine learning models was considered to address the ICR forecasting problem in order to find the optimal balance between accuracy and interpretability. Linear methods were analyzed at first (Ridge, Lasso, and ElasticNet), which are well suited for situations where simplicity and high generalization ability with a large number of features are important. Ridge regression (L2 regularization) helps smooth out possible jumps in coefficients and reduce overfitting, while Lasso (L1 regularization) is able to zero out coefficients of insignificant features and thus perform partial data filtering. ElasticNet represents a compromise, combining the properties of both approaches and often demonstrating more stable results, especially in problems with many correlated variables.

However, given the complexity of the rocks and potential nonlinearities in the data, the analysis was suplemented with ensemble methods: Random Forest, Gradient Boosting, and Extra Trees. They build multiple decision trees and combine their results, which allows them to model complex relationships and deal with noise more effectively. For example, Random Forest averages the results of individual trees trained on bootstrapped samples, which reduces the risk of overfitting and increases robustness to outliers. Gradient Boosting sequentially corrects the errors of previous trees, which helps to better capture subtle patterns in the data, and Extra Trees adds additional randomness to the tree construction process, due to which it can work even

more effectively on heterogeneous datasets. To account for nonlinear relationships, Support Vector Regression (SVR) and a neural network in the form of a multilayer perceptron (MLP) was included. SVR with an appropriate kernel (e.g., RBF) can handle highly nonlinear functions, and a neural network, with the right settings for the architecture and training parameters, often outperforms other models in problems where the ability to generalize to complex distributions is important. Finally, to assess how much better all these approaches are than the trivial option, the DummyRegressor (ZeroR) model, which simply predicts the mean value of the target variable, served as the base model.

The training process for each model consisted of several stages. First, synthetic datasets were generated (using three different methods) to expand the training set and make it more diverse. Next, cross-validation was performed for each model (usually 5-fold), where hyperparameters were selected using Grid Search or alternative optimization algorithms. This helped to refine parameters such as regularization coefficients for linear models, the number of trees and depth for Random Forest and Gradient Boosting, kernel parameters for SVR, and the number of layers and neurons in MLP. After completing the tuning on synthetic data, model trained on the entire expanded dataset and tested on the original one (without adding synthetics) to check the real generalization ability. This comprehensive approach to model selection and training ensures that were considered different classes of algorithms, tested them on a wide range of data, and selected those that actually perform best in predicting the performance of roadheader machines.

Linear models (Ridge, Lasso and ElasticNet)

For linear methods, it is assumed that the target variable y depends linearly on the set of features X. Let X be a matrix of size $n \times d$, where nnn is the number of observations, d is the number of features, and y is a vector of targets of size n. Observation of a vector of coefficients $\beta \in R^d$ that best approximates the dependence $y \approx X\beta$ is performed.

Ridge regression (L2 regularization).

The objective function of Ridge regression is formulated as

$$\min_{\beta} ||X\beta - y||^2 + \lambda||\beta||^2, \tag{25}$$

where $\|\cdot\|$ denotes the Euclidean norm, and $\lambda \geq 0$ is the regularization coefficient. The first term is responsible for minimizing the forecast error, and the second limits the growth of the coefficients, reducing the risk of overfitting.

Lasso regression (L1 regularization).

Unlike Ridge, the L1 norm is used here:

$$\min_{\beta} ||X\beta - y||^2 + \alpha||\beta||_1,$$

(26)

where $||\beta||_1 = \sum_{j=1}^{d} \alpha|\beta_j|$ The parameter α controls the strength of the regularization. Due to the L1 norm, some components of β can be set to zero, which gives a feature selection effect.

ElasticNet (mixed regularization).

ElasticNet combines the properties of Ridge and Lasso by minimizing

$$\min_{\beta} ||X\beta - y||^2 + \alpha(\rho||\beta||_1 + (1-\rho)||\beta||^2),$$

(27)

where α and ρ control the degree of contribution of L1 and L2 to the regularization. This helps to cope with cases where there is both feature sparseness and cross-correlation.

Ensemble methods (Random Forest, Gradient Boosting, Extra Trees)

Ensemble models build several base algorithms (usually decision trees) and combine their predictions.

Random Forest.

Let us have T trees, each of which is denoted as $f_t(x)$. In regression, the final prediction is obtained by averaging:

$$\hat{y}(x) = \frac{1}{T}\sum_{t=1}^{T} f_t(x).$$

(28)

Each tree is trained on a bootstrap sample of the original data, and a random subsample of features is used to select features for splitting at each node. This approach reduces the correlation of trees and improves generalization ability.

Gradient Boosting.

The idea of gradient boosting is that at each step a new tree is built that approximates the antigradient of the loss function. Let $F_m(x)$ be the model after the mmm-th step, then at step m+1 a tree $h_{m+1}(x)$ is built, which approximates the residuals or errors of the previous model. The final model takes the form:

$$F_{m+1}(x) = F_m(x) + \nu \cdot h_{m+1}(x),$$

(29)

where ν is the learning rate. This iterative scheme allows us to gradually “refine” the predictions, reducing the error.

Extra Trees.

The Extra Trees (Extremely Randomized Trees) model is similar to Random Forest in many ways, but it selects partition thresholds even more randomly. Formally, when constructing a tree, a subset of features and random thresholds are randomly selected for each node, rather than a detailed enumeration. This additional "loosening" of the tree structure can improve the robustness to overfitting on heterogeneous datasets.

Nonlinear methods and neural networks (SVR, MLP)

Support Vector Regression (SVR).

The SVR model searches for a function $f(x) = \langle w, x\rangle + b$ that fits into the ε -neighborhood of the target value and has the minimum norm of the vector w. The problem is formulated as follows:

$$\min_{w\,\beta} \frac{1}{2}||w||^2 \text{ under constraints } \{y_i - \langle w, x_i\rangle - b \le \epsilon, \langle w, x_i\rangle + b\ - y_i \le \epsilon\ \},$$

(30)

where ε defines the error tolerance. When using kernels (e.g., RBF), the regression goes into a nonlinear feature space, which helps to capture complex dependencies.

Multilayer Perceptron (MLP).

An MLP neural network consists of several layers: an input layer, one or more hidden layers, and an output layer. Let there be HHH neurons in the hidden layer, then the output of each neuron is calculated as

$$z_j = \sigma\left(\sum_{i+1}^{n} w_{ji}^{(1)} x_i + b_j^{(1)}\right), \tag{31}$$

where σ(·) is the activation function (e.g. ReLU or sigmoid), $w_{ji}^{(1)}$ and $b_j^{(1)}$ are the weights and biases for the first layer. The output layer undergoes a similar transformation, and the final output of the network gives an estimate of the target variable. The network is trained by minimizing the loss function (e.g. MSE) using backpropagation, where the weights are adjusted using gradient methods.

Baseline Model (DummyRegressor, ZeroR)

The baseline model simply predicts the mean (or median) of the target variable, ignoring the feature values. Let $\underline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Then for any x, the model predicts $\hat{y}(x) = \underline{y}$. This is the simplest approach, which serves as a lower bound on the quality: all other models should perform better than the trivial averaging.

4.4 Cross-validation and evaluation metrics

For an objective assessment of the quality of the constructed models, our study uses 5-fold cross-validation. This method allows dividing the synthetic sample into five equal parts, sequentially training the model on four parts and testing it on the remaining one. This approach helps to obtain a reliable estimate of the generalization ability of the model, minimizing the impact of random data partitioning and preventing overfitting. Cross-validation ensures the stability of the results and makes it possible to optimize the hyperparameters of each model, which is especially important when working with a small amount of initial data supplemented synthetically.

As metrics for assessing the effectiveness of the models, three main indicators are used - the determination coefficient (R^2), the mean square error (MSE) and variance accounted for (VAF). The determination coefficient R^2 and VAF are inter-related, and show what proportion of the variance of the target variable is explained by the model. An R^2 value close to 1 indicates that the model is able to adequately describe the data, while low values indicate insufficient explanatory power of the model. The mean squared error (MSE) measures the average squared difference between predicted and actual values. The smaller the MSE, the more accurate the model. Using these metrics allows for a comprehensive assessment of both the relative ability of the model to explain data variability and the absolute accuracy of its forecasts. The metrics are calculated in two modes: during cross-validation on synthetic data, which allows for selecting optimal hyperparameters and assessing the stability of the model, and when testing the trained model on the original data set. This approach ensures that the achievements obtained on synthetic data are actually reflected in the quality of forecasts on the real sample. As a result, comparing R^2 and MSE for different models and methods of generating synthetic data allows for choosing the most optimal algorithm for ICR forecasting, ensuring a high level of accuracy and reliability of forecasts.

5. DATA ANALYSIS

Data Analysis

At this stage of the study, a detailed analysis of the original data set containing the geotechincal and roadheader performance parameters from San Manuel mine was carried out. The dataset that was obtained contains 28 datapoints, including the RMR, RQD, UCS, and ICR values. Initial data is given in the RMR table and Machine performance graph in the figure 3.

Data preprocessing and expansion of the feature space

First of all, features were normalized to bring them to a single scale, which is critical for correct model training. After that, second-degree polynomial features were generated. This allowed us to identify hidden nonlinear relationships between the original characteristics and the target variable (ICR). Expansion of the feature space made it possible to detect additional dependencies that were not obvious when analyzing the original data.

Descriptive statistics

To assess the quality and distribution of data after the preprocessing stages, a descriptive statistics table was compiled. It presents the following indicators for each feature (both original and generated):

Count — number of observations (in this case, 28 records).
Mean — average value of the feature.
Std — standard deviation, reflecting the spread of values around the mean.
Min, 25%, 50%, 75%, Max — minimum value, first quartile, median, third quartile and maximum value, allowing to estimate the distribution and range of values.

Examples of statistics interpretation

For example, for the original RQD feature, the mean value is close to zero, and the standard deviation is approximately one. This indicates the correctness of the data scaling. In the case of polynomial features, such as RQD^2 or RQD × RMR, the values of the mean and standard deviations also confirm the precision of the scaling.

Table 5: EDA summary for the Original Dataset

	count	mean	std	min	25%	50%	75%	max
RQD	28.0	40.0	19.459	3.0	24.25	43.0	53.25	72.0
RMR	28.0	35.786	14.312	15.0	27.0	32.5	44.0	70.0
UCS	28.0	119.429	34.009	60.0	91.5	118.0	139.0	200.0
ICR	28.0	15.215	4.151	6.492	12.526	14.511	18.841	22.167

Table 6: EDA summary for the Polynomial features of the Original Dataset

Feature	count	mean	std	min	25%	50%	75%	max
RQD	28	0	1,00	-1,936	-0,824	0,157	0,693	1,675
RMR	28	0	1,00	-1,479	-0,625	-0,234	0,584	2,434
UCS	28	0	1,00	-1,779	-0,836	-0,043	0,586	2,413
RQD²	28	1	1,096	0,003	0,239	0,576	1,326	3,749
RQD RMR	28	0,798	1,204	-0,377	0,036	0,367	0,825	4,077
RQD UCS	28	0,674	1,116	-0,809	-0,049	0,237	1,024	4,04
RMR²	28	1	1,571	0,007	0,116	0,391	0,787	5,926
RMR UCS	28	0,636	1,348	-0,984	0,001	0,254	0,683	5,873
UCS²	28	1	1,286	0	0,1	0,63	1,355	5,821

Second-degree polynomial features were created to capture complex nonlinear relationships between original features. Generating new features, such as squared original features and their interactions, produced data with diverse values and large scatter. For example, squared features, such as RQD^2, RMR^2, UCS^2 have a mean value 1 with a high standard deviation, indicating significant variations in the data. These features can be useful for models as they provide additional features that can help in better prediction.

Interactions between features, such as RQD RMR, RMR UCS, have both positive and negative values, reflecting complex relationships between geological characteristics. These interactions can help models uncover hidden dependencies and improve predictive ability, especially for nonlinear models such as ensemble methods.

The high scatter and large ranges of polynomial features highlight their importance for machine learning models, where such features can improve prediction accuracy. These polynomial features enable models to capture more complex relationships, which is especially useful for predicting drilling machine performance, where the interaction of different characteristics is critical.

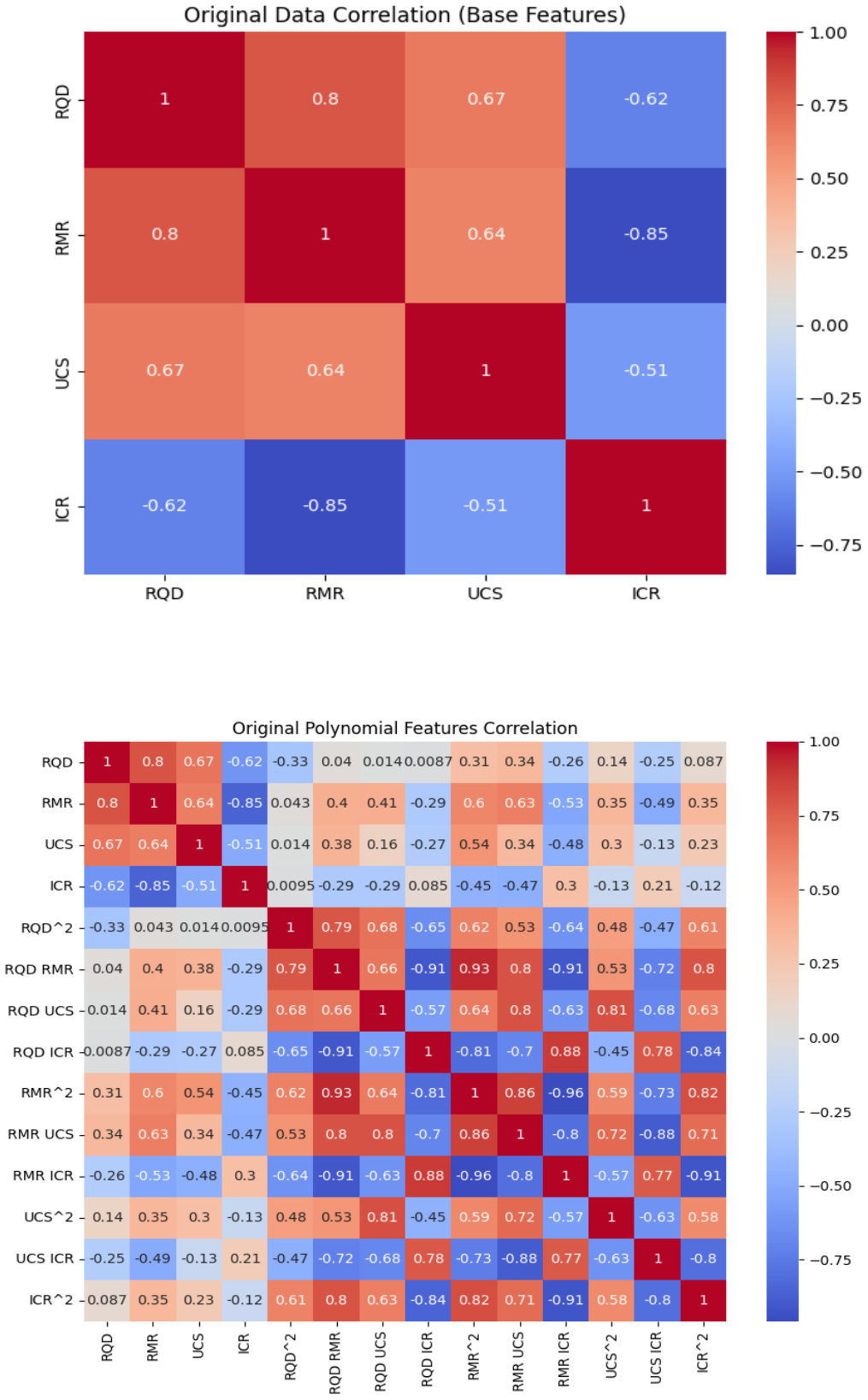


Figure 6: Correlation Heatmaps of the Original Data. (Figure A - Heatmap of the Base Features, B - Polynomial Features)

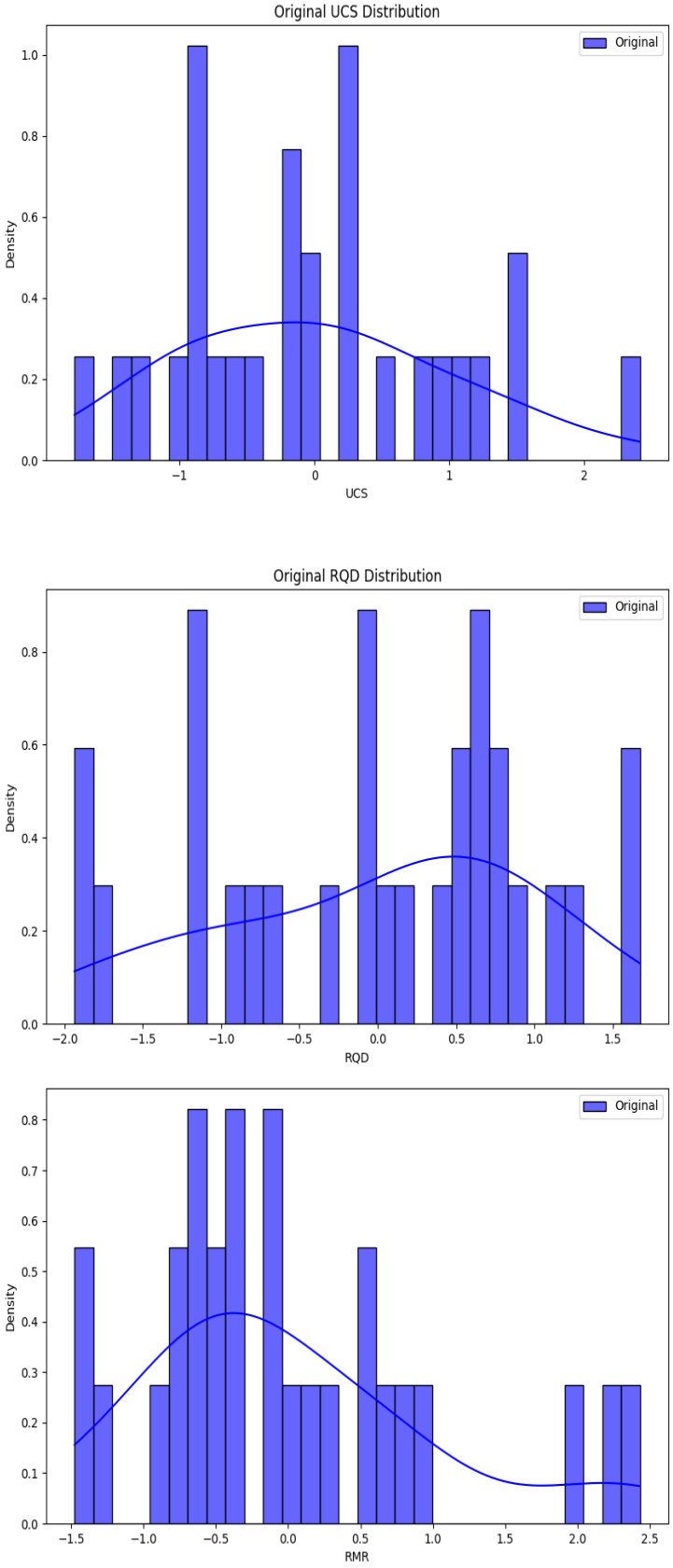


Figure 7: Histograms of Standardized Original Data Distributions (Histogram A - UCS, B - RQD, C - RMR value).

5.2 Model training and evaluation

This section provides a comparative analysis of the performance of models trained on a synthetically expanded sample, followed by an evaluation of their performance on the original data. A detailed description of the training process is presented in the methodological section, so here the emphasis is on comparing the results and interpreting the obtained metrics.

Based on synthetically generated data using three different methods (Ridge Regression Labeling, Gaussian Noise Labeling and Random Forest Regression Labeling), 5-fold cross-validation was performed for each model. The main evaluation metrics are the determination coefficient (R^2) and the mean square error (MSE). These indicators allow us to evaluate the extent to which models trained on synthetic data are able to generalize information when moving to the original data set.

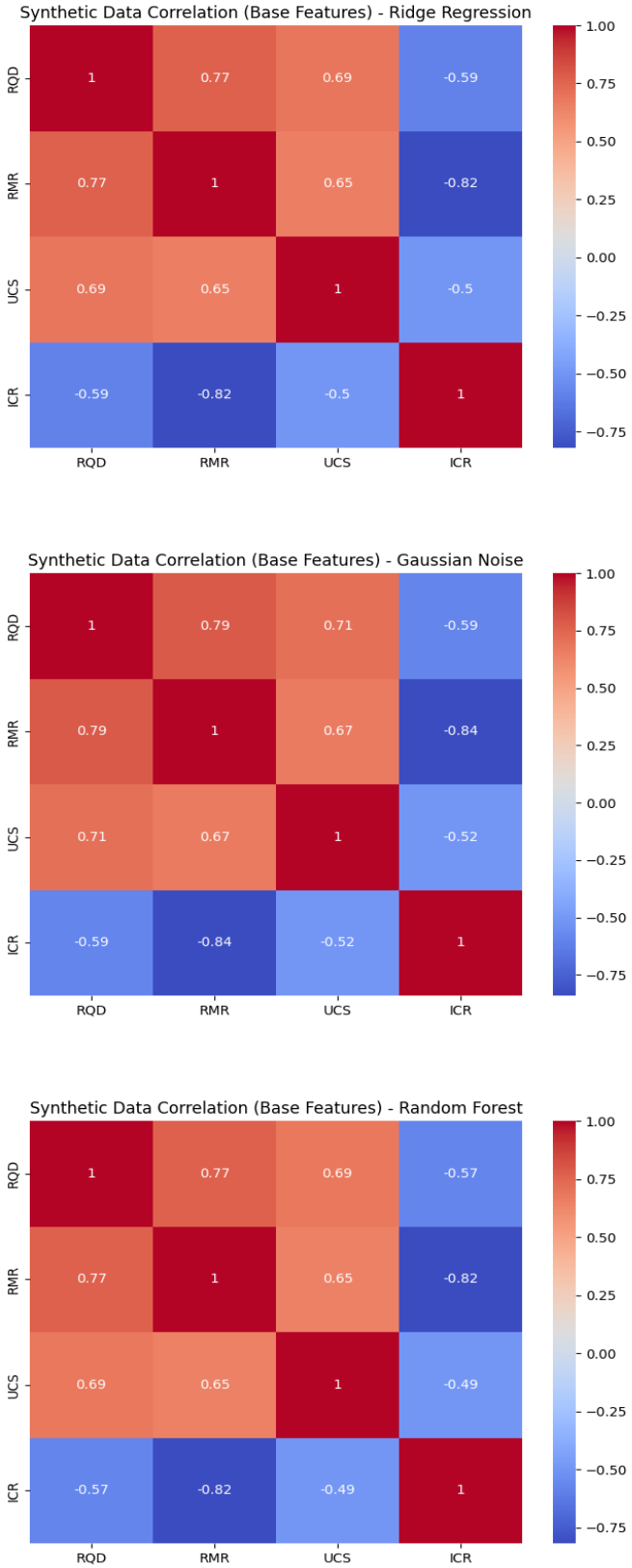


Figure 8: Base Features Correlation Heatmaps of the Synthetic Data. (Figure A - Ridge , B - Gaussian, C - Random Forest)

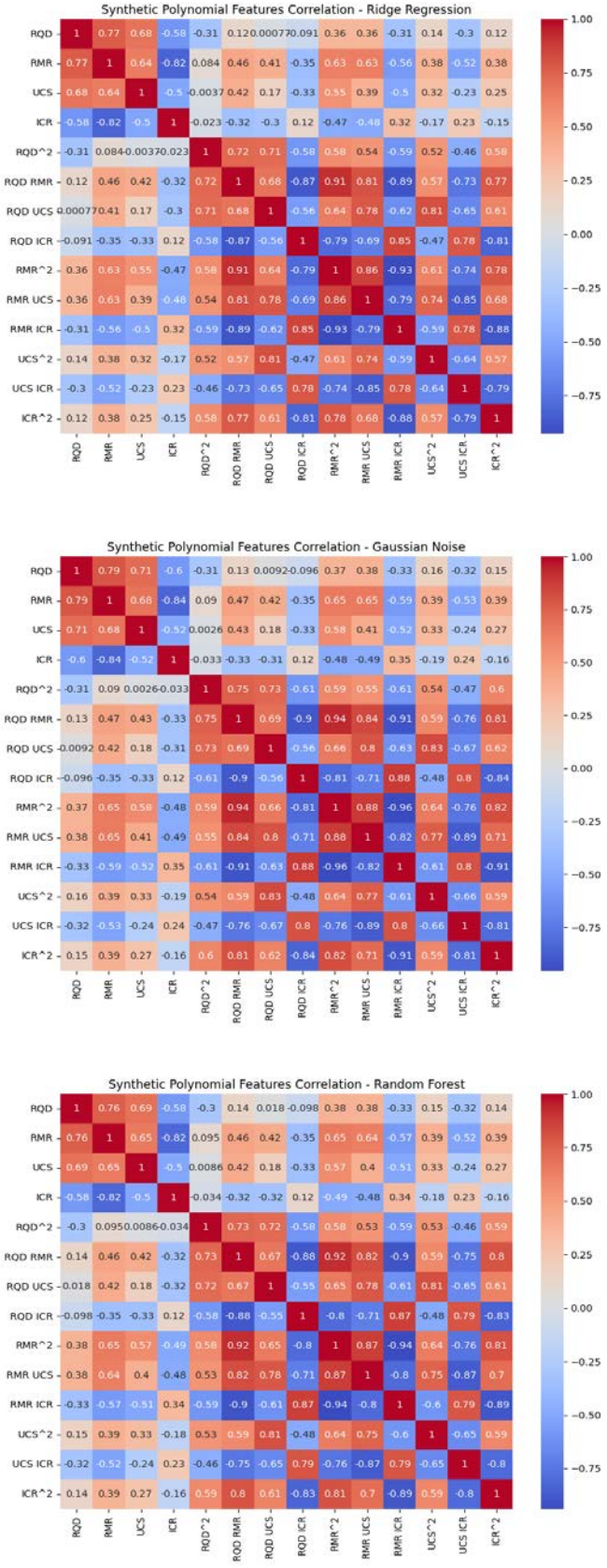


Figure 9: Polynomial Features Correlation Heatmaps of the Synthetic Data. (Figure A - Ridge , B - Gaussian, C - Random Forest)

Following the correlation heatmaps for base and polynomial features, the histograms for the synthetic data distribution were plotted:

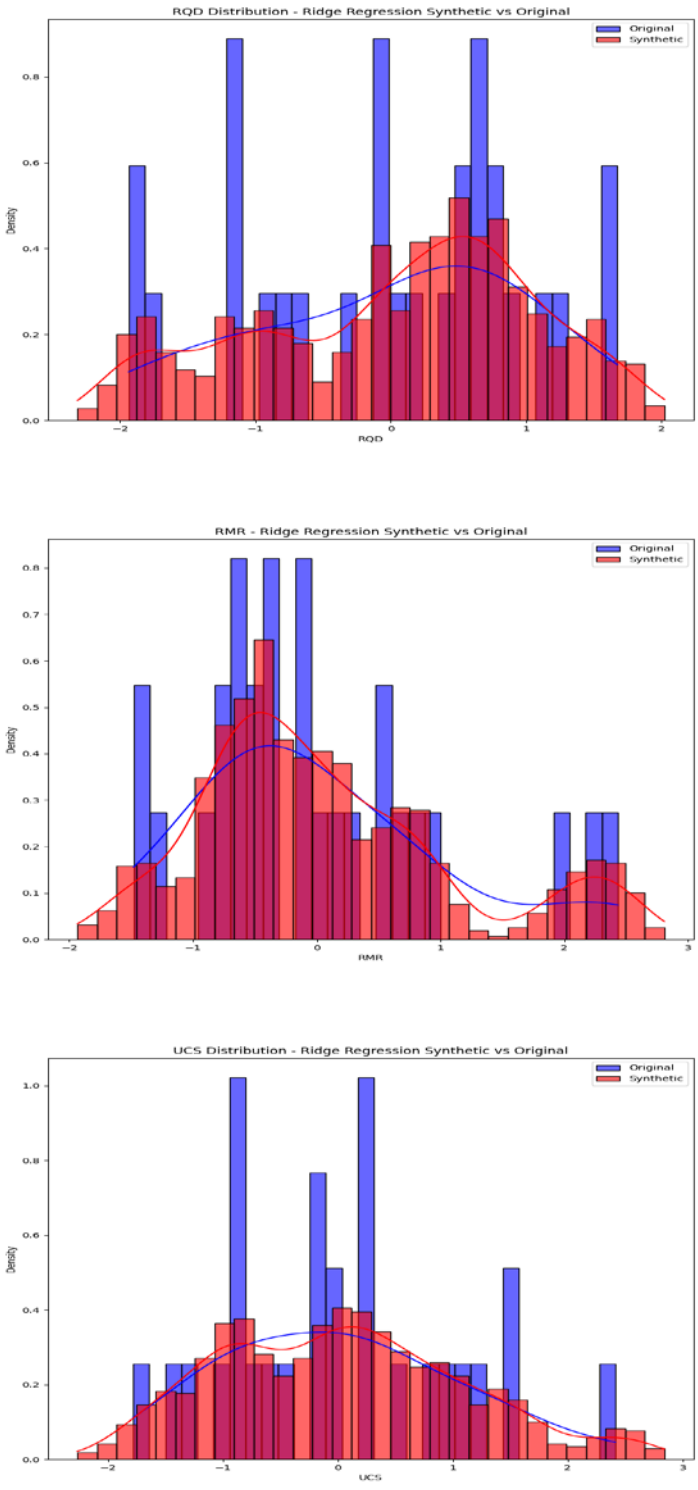


Figure 10: Synthetic Data Distribution for Ridge Regression Labeling (Figure A - RQD , B - RMR, C - UCS).

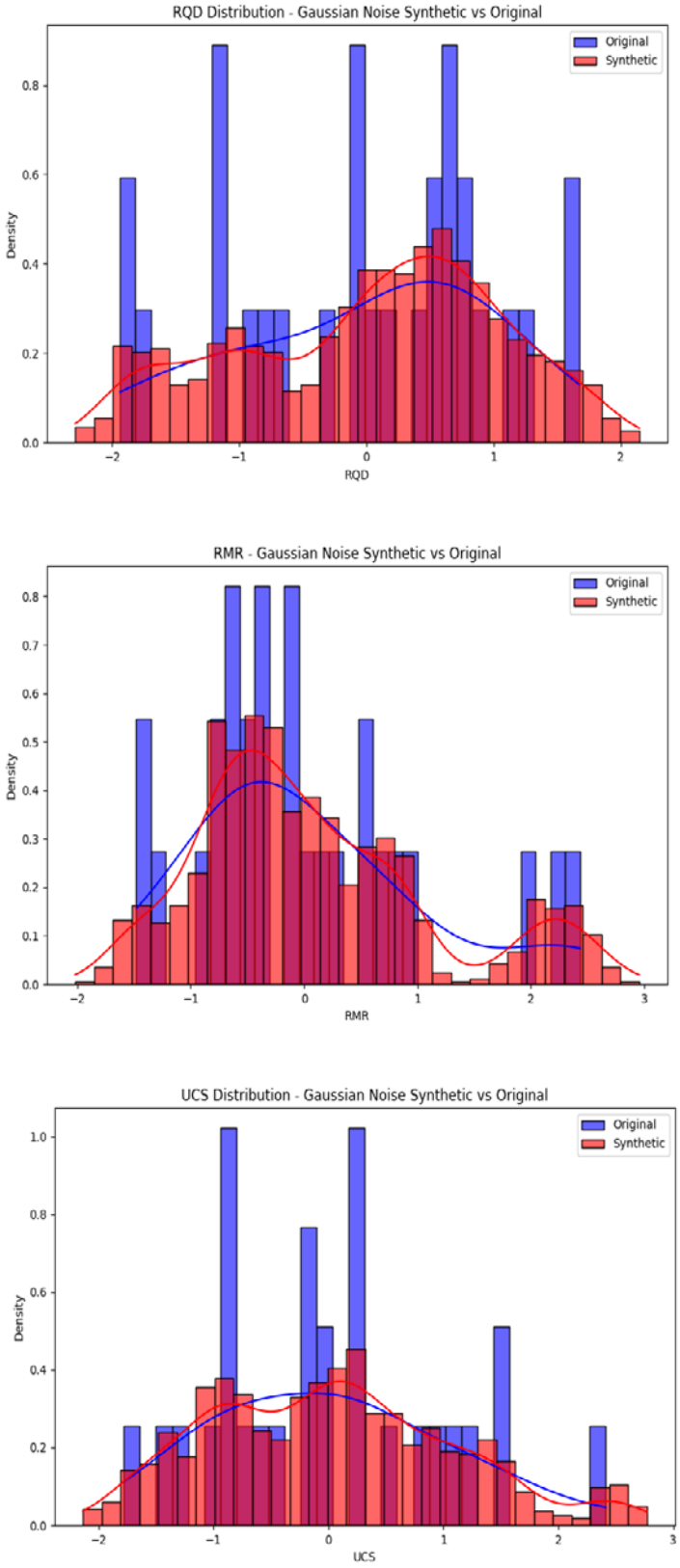


Figure 11: Synthetic Data Distribution for Gaussian Noise Labeling (Figure A - RQD , B - RMR, C - UCS).

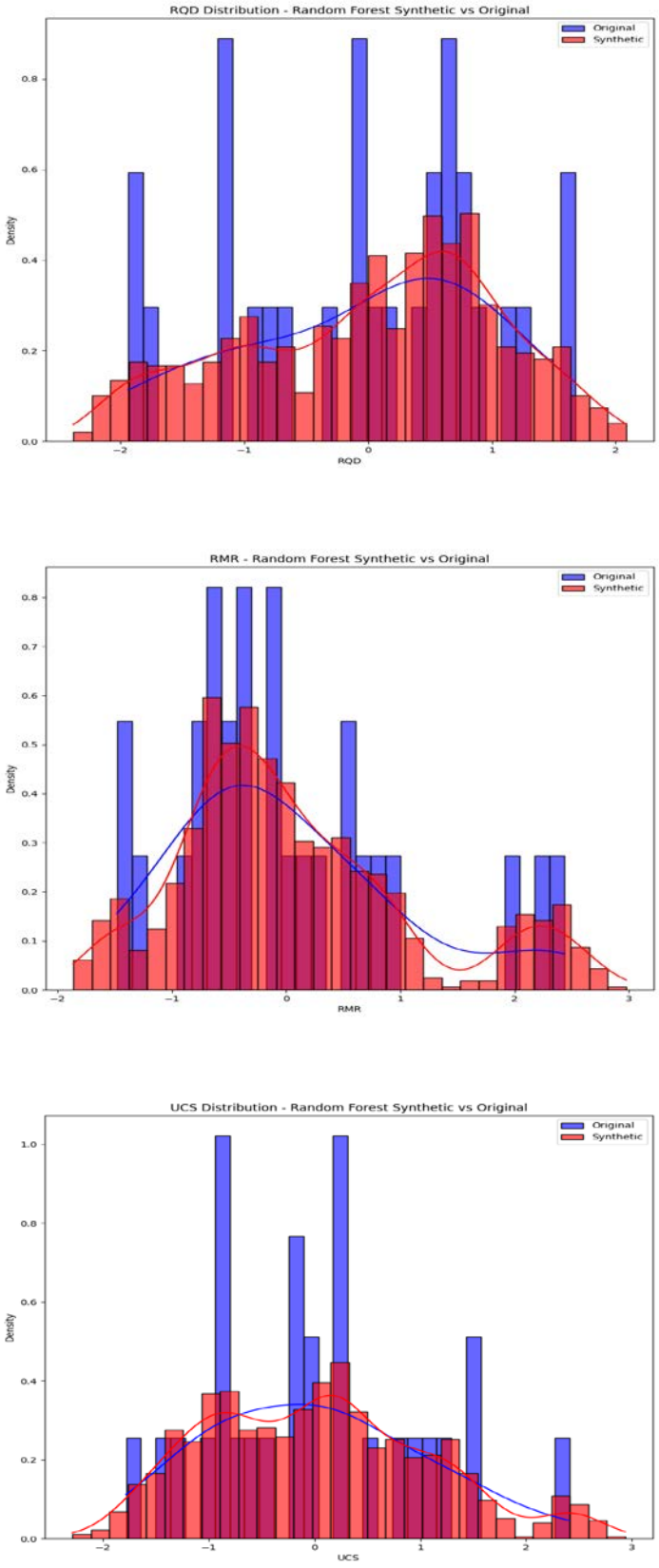


Figure 12: Synthetic Data Distribution for Random Forest Regression Labeling (Figure A - RQD , B - RMR, C - UCS)**.**

Following the synthetic data generation, the models were trained on the augmented data. After the training was performed, models were evaluated on the original dataset which serves the purpose of the test set.

The scatterplots were plotted with the purpose of representing the output:

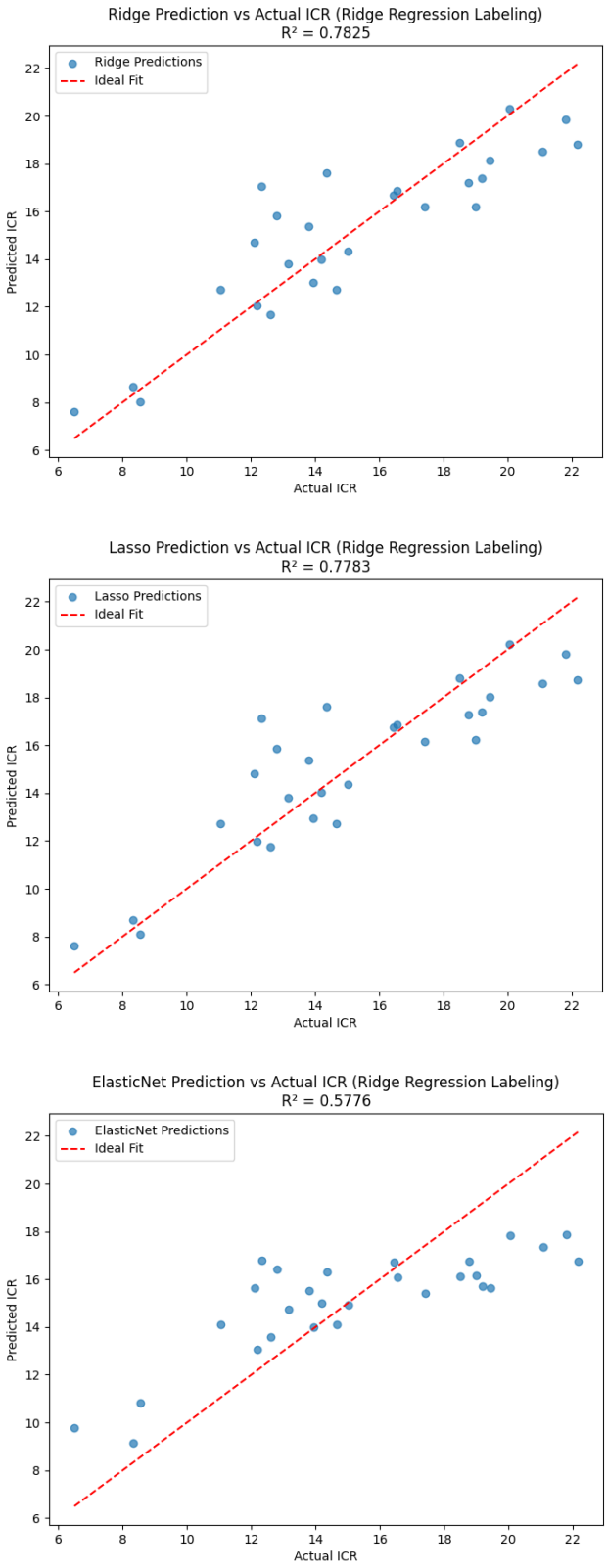


***Figure 13*:** ICR plots for Linear Models trained on Ridge Regression Labeling data (Scatterplot A - Ridge regression, B - Lasso, C - ElasticNet).

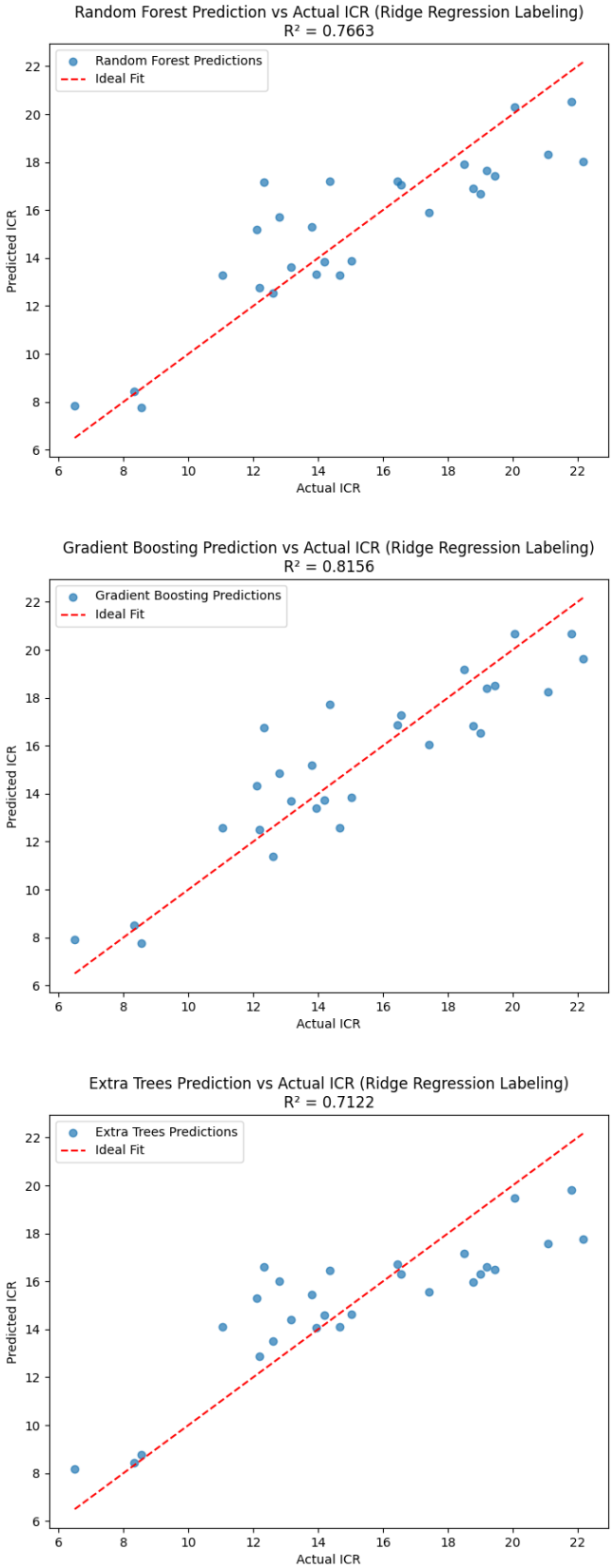
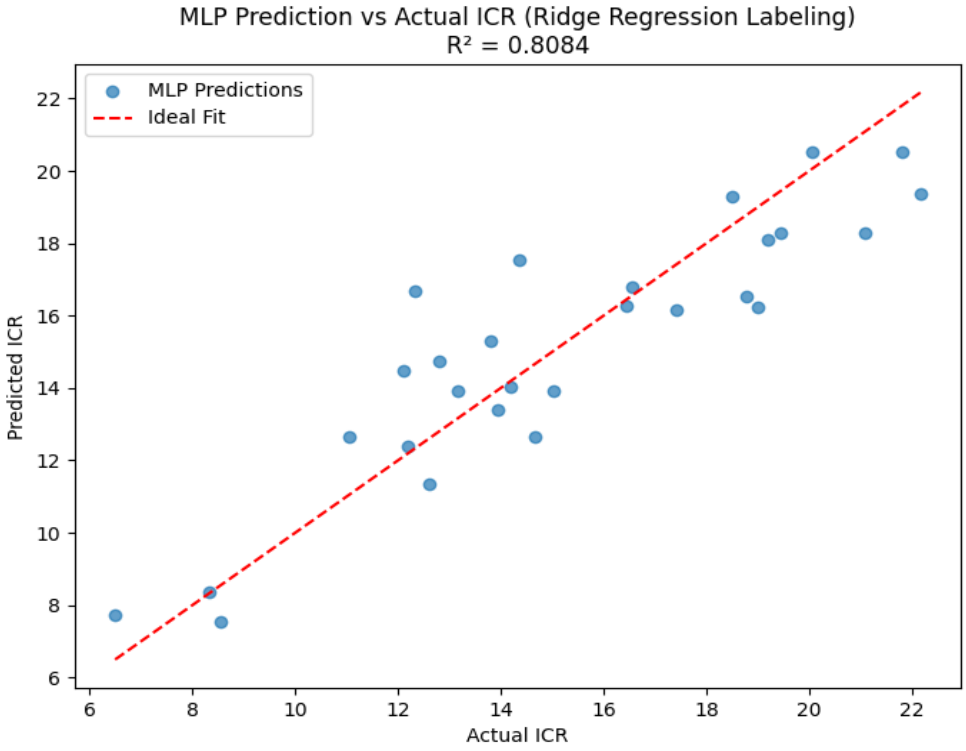


Figure 14: ICR plots for Ensemble Models trained on Ridge Regression Labeling data (Scatterplot A - Random Forest, B - Gradient Boosting, C - ExtraTrees).



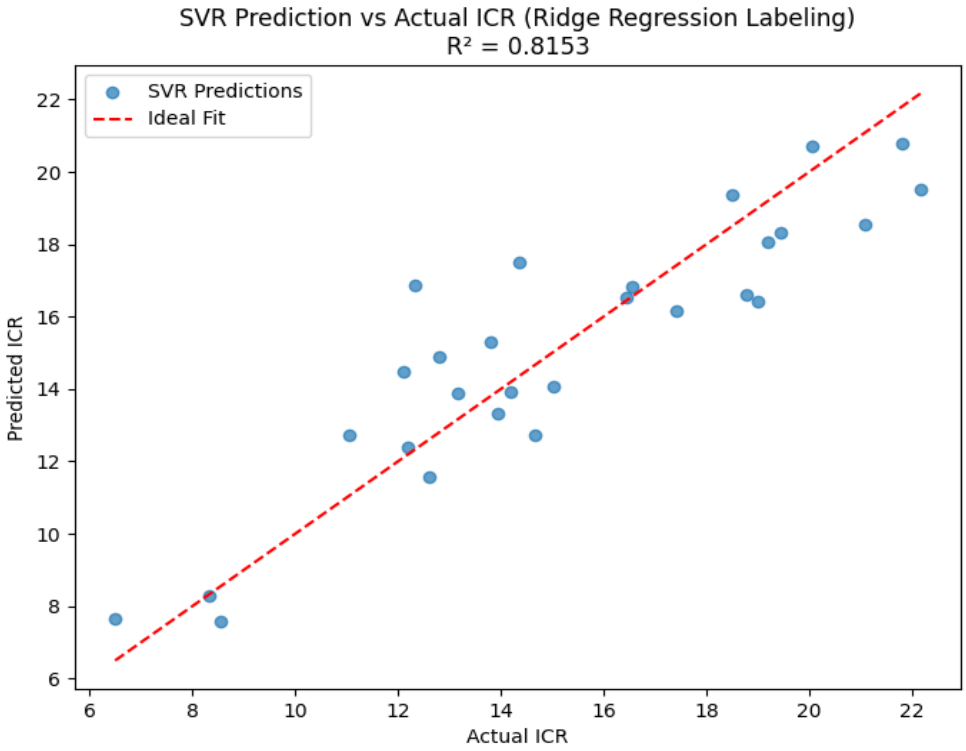


***Figure 15*:** ICR plots for Non-Linear Models trained on Ridge Regression Labeling data (Scatterplot A - MLP, B - SVR).

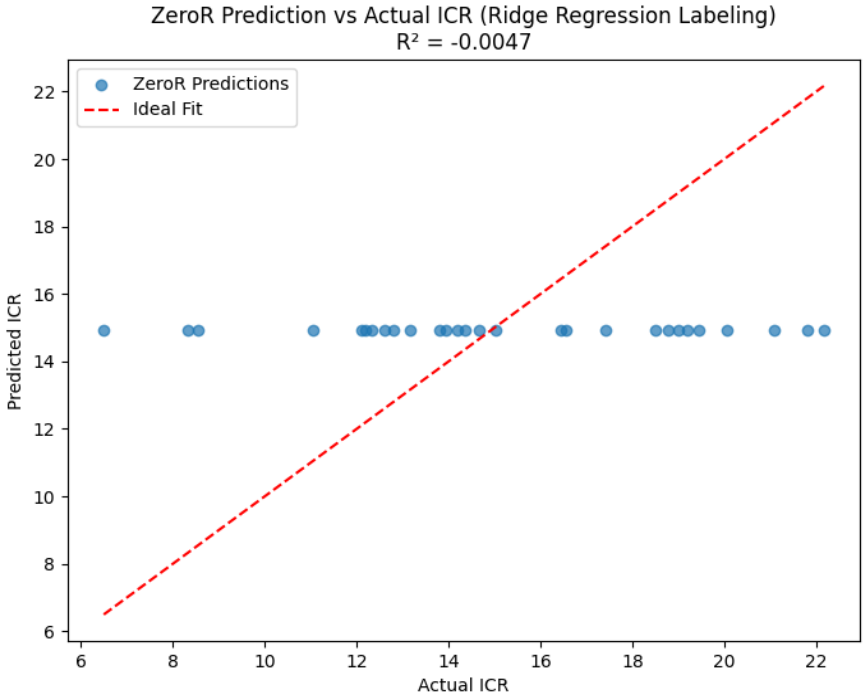


Figure 16: ICR plot for Base Model trained on Ridge Regression Labeling data(ZeroR).

The same plots are made for models trained on **Gaussian Noise Labeling** method.

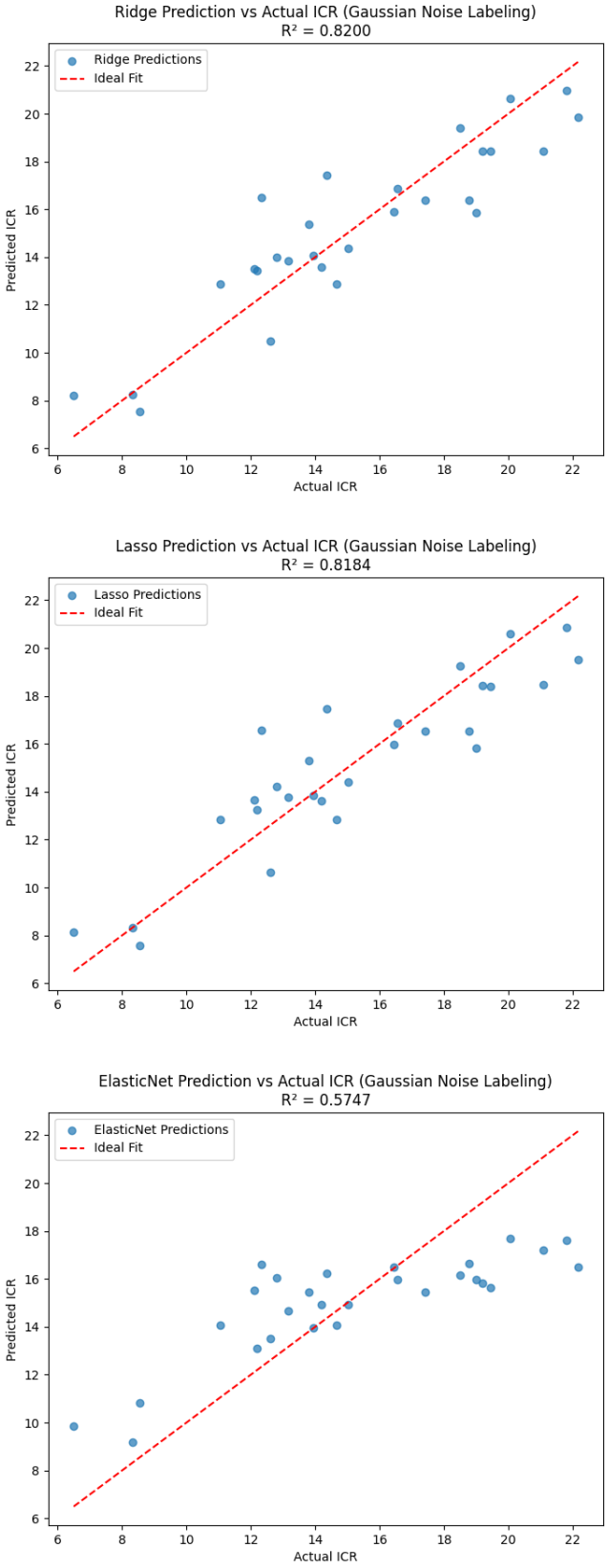


Figure 17: ICR plots for Linear Models trained on Gaussian Noise Labeling method (Scatterplot A - Ridge regression, B - Lasso, C - ElasticNet).

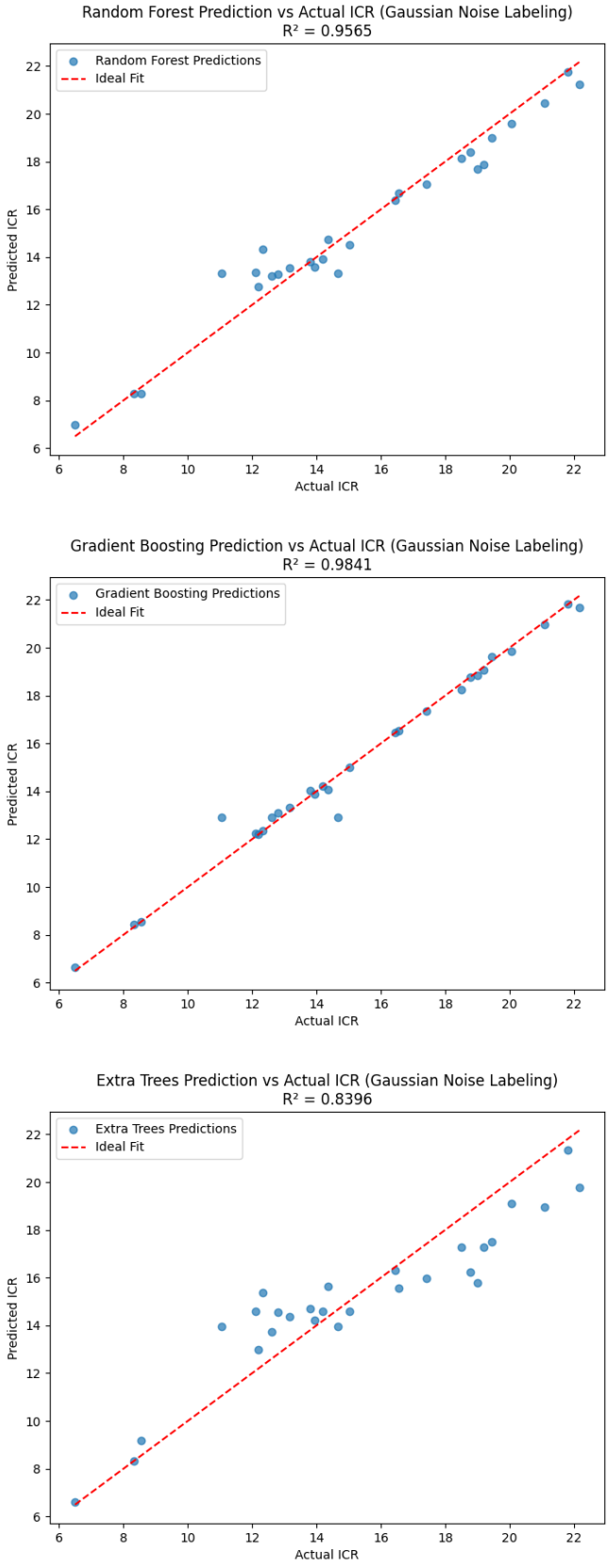
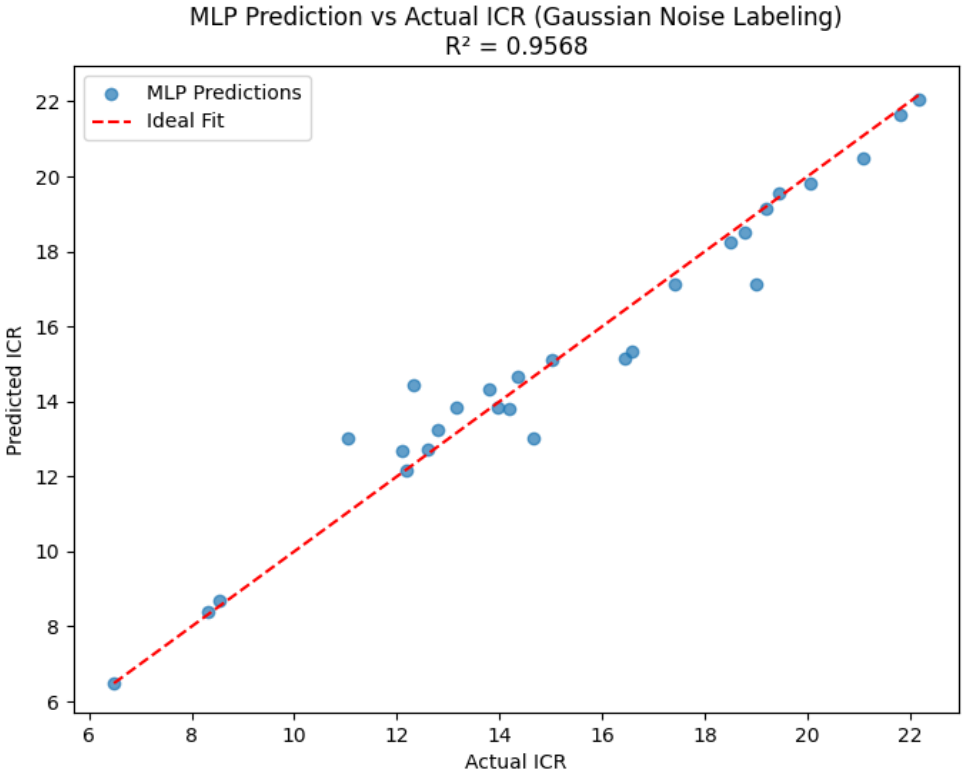


***Figure 18**:* ICR plots for Ensemble Models trained on Gaussian Noise Labeling method (Random Forest, Gradient Boosting, ExtraTrees).



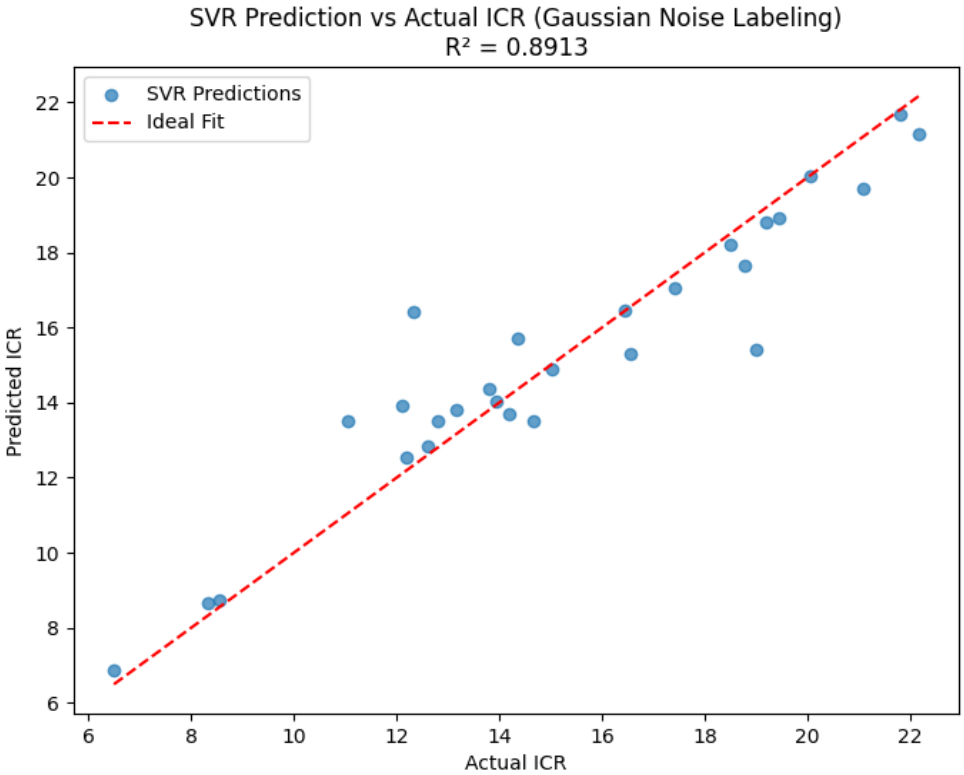


***Figure 19*:** ICR plots for Non-Linear Models trained on Gaussian Noise Labeling data (Scatterplot A - MLP, B - SVR).

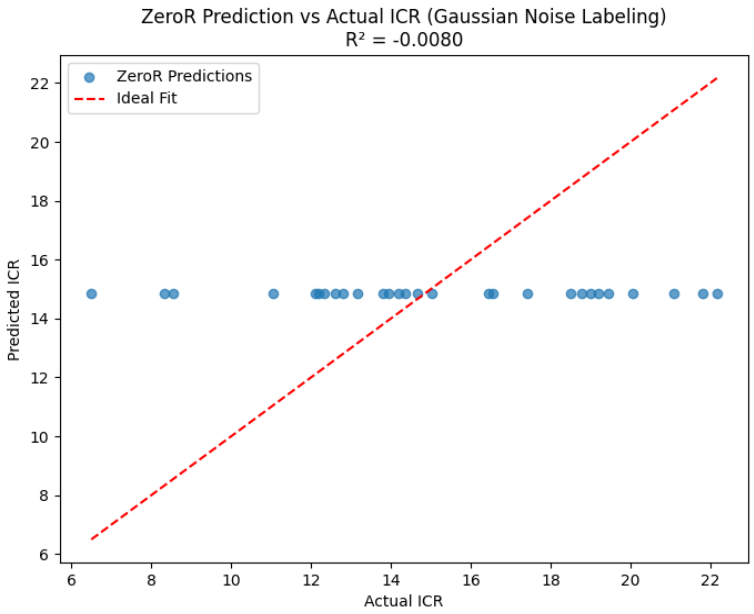


***Figure 20*:** ICR plot for Base Model trained on Ridge Regression Labeling data(ZeroR).

Scatterplots of ICR values obtained on the models trained on Random Forest Regression Labeling.

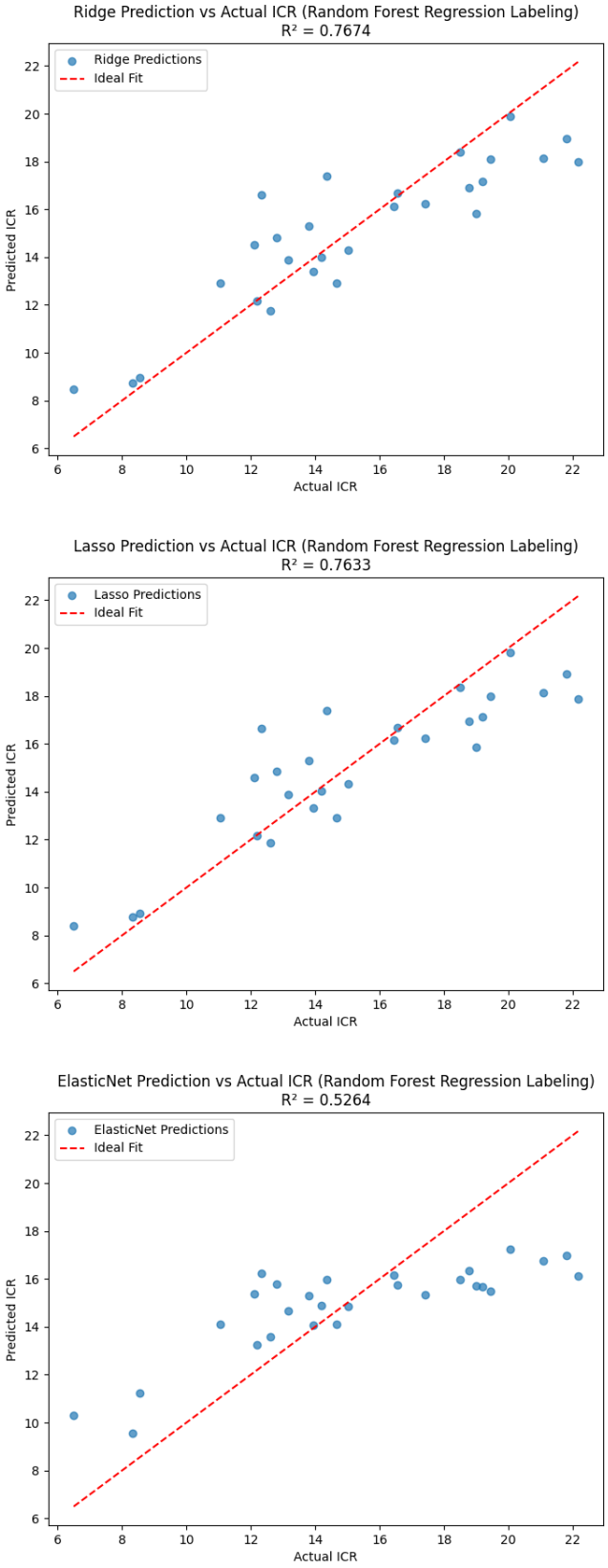


***Figure 21*:** ICR plots for Linear Models trained on Random Forest Regression Labeling data (Scatterplot A - Ridge Regression, B - Lasso, C - ElasticNet).

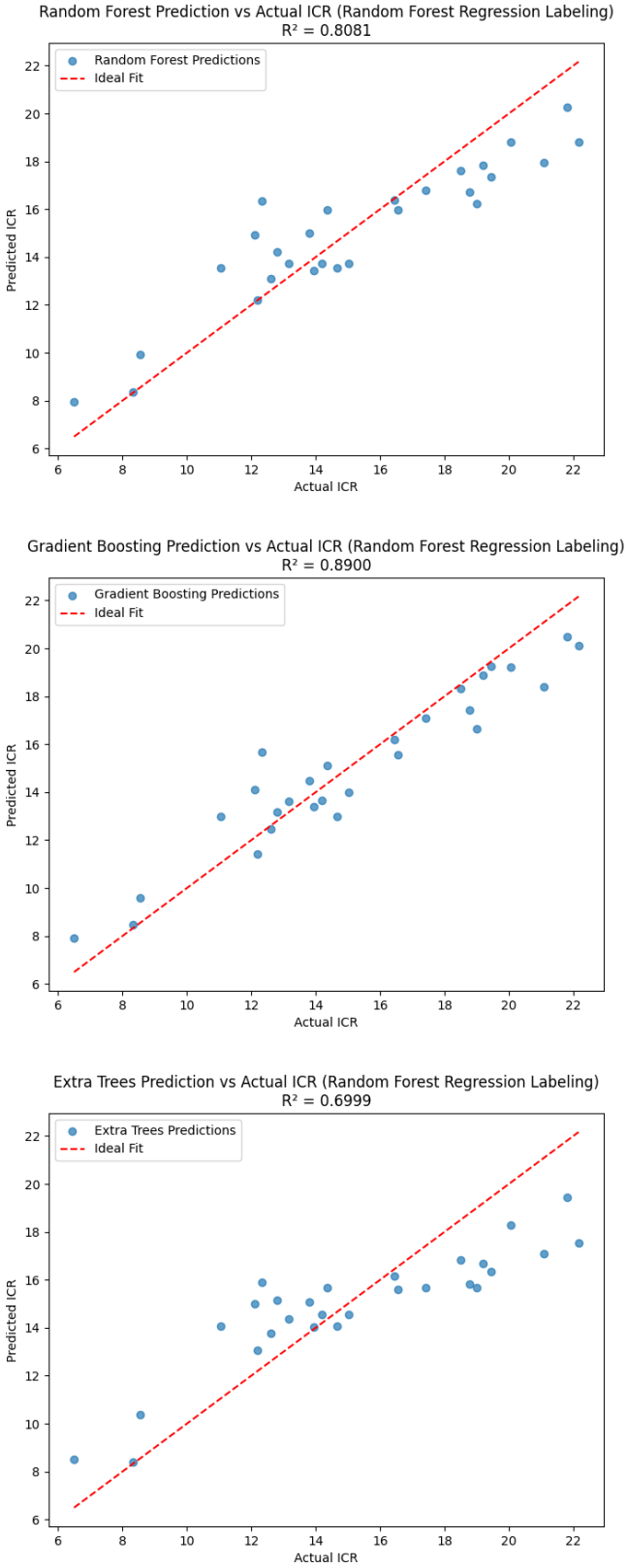


***Figure 22*:** ICR plots for Ensemble Models trained on Ridge Regression Labeling data (Scatterplot A - Random Forest, B - Gradient Boosting, C - ExtraTrees).

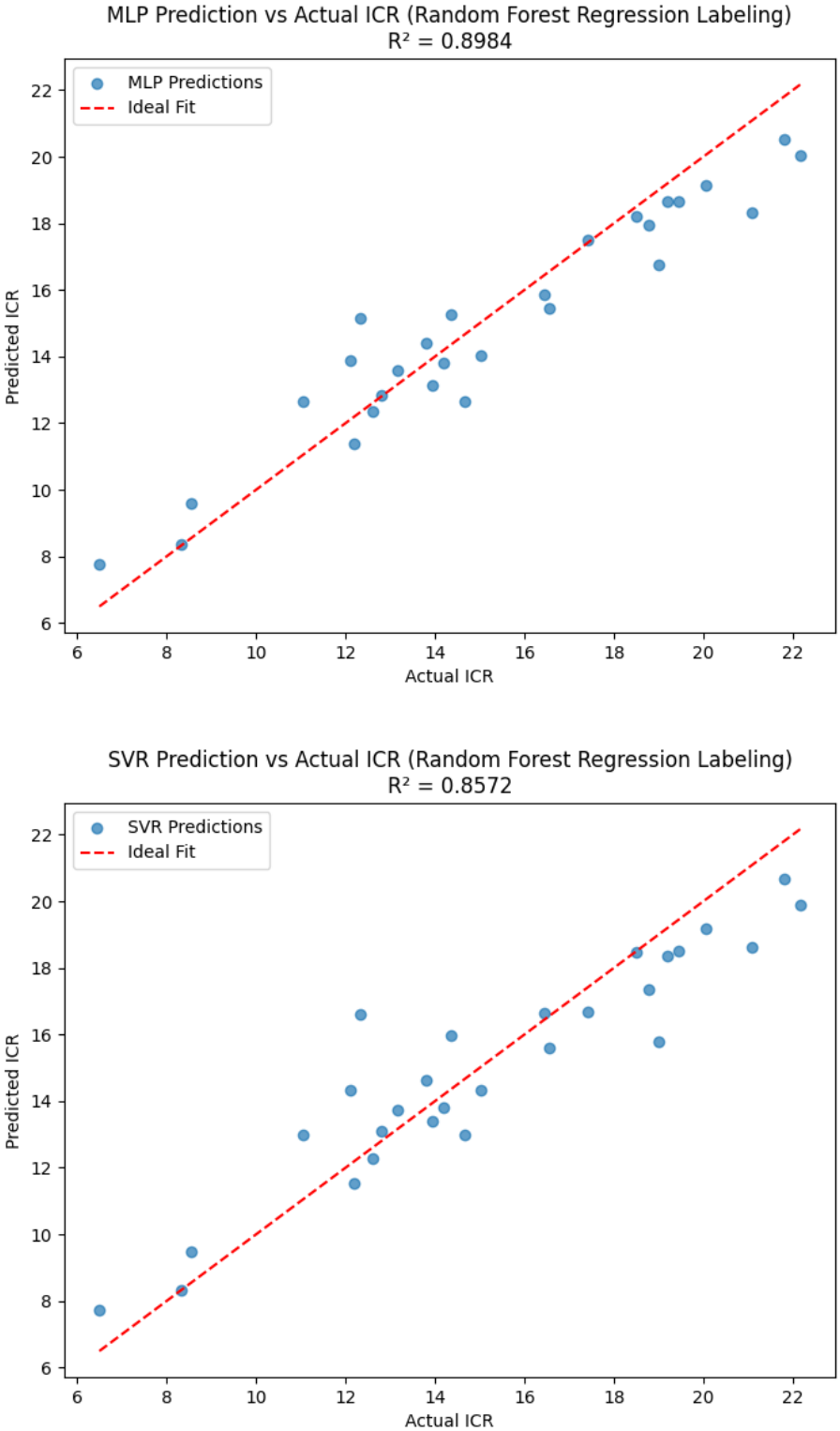


***Figure 23*:** ICR plots for Non-Linear Models trained on Random Forest Regression Labeling data (Scatterplot A - MLP, B - SVR).

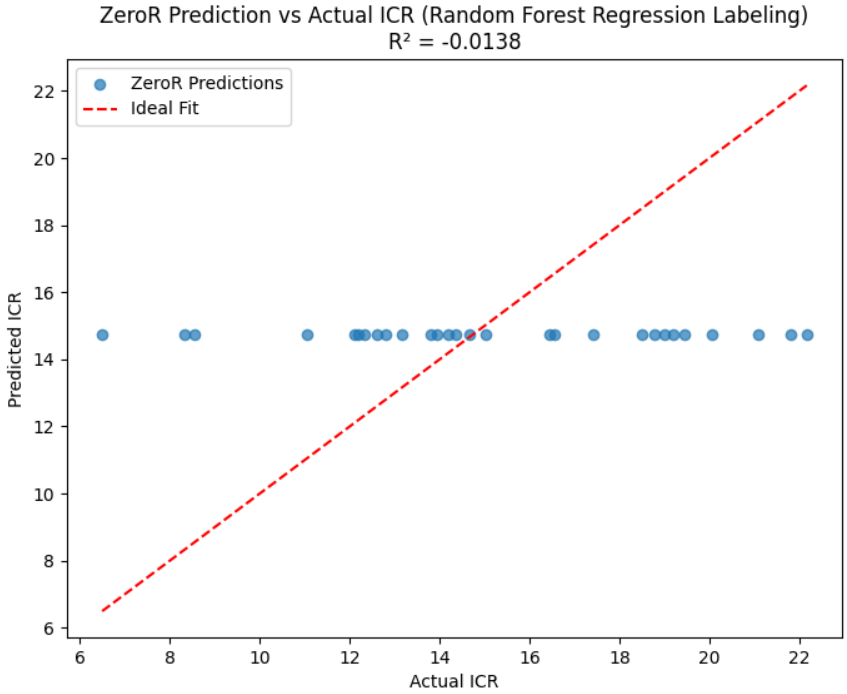


***Figure 24*:** ICR plot for Base Model (ZeroR) trained on Random Forest Regression Labeling**.**

The results show that ensemble methods (e.g. Random Forest, Gradient Boosting, Extra Trees) demonstrate the best performance, providing high prediction accuracy and resistance to overfitting. At the same time, linear models (Ridge, Lasso, ElasticNet) are inferior in forecasting quality, which indicates the difficulty of identifying nonlinear relationships in the original data.

Table 7: Example Outputs of Predicted ICR values

Model	ICR actual	Ridge Regression	Gaussian Noise	R.Forest Regression
Ridge	8,309	8,662	8,120	8,737
Lasso	8,309	8,684	8,216	8,783
ElasticNet	8,309	9,077	9,093	9,539
RandomForest	8,309	8,451	8,209	8,339
XGboost	8,309	8,475	8,400	8,438
ExtraTrees	8,309	8,451	8,257	8,390
MLP	8,309	8,359	8,447	8,263
SVR	8,309	8,322	8,421	8,345
ZeroR	8,309	14,935	14,910	14,736

By examining the table, results confirm the effectiveness of using a synthetically expanded sample for training models. The tables containing the results for all predicted ICR values are attached in Appendices 1-9.

Comparative analysis demonstrated that the optimal balance between prediction accuracy and model stability is achieved when using ensemble algorithms. These findings serve as an important basis for further practical application of the developed ICR forecasting technique.

5.3 Performance Comparison

This section provides a detailed comparison of the models' performance based on the key metrics - the coefficient of determination (R^2) and the mean square error (MSE). The experimental results are obtained for three synthetic data generation methods, which allows us to evaluate how the selected models cope with ICR forecasting under different approaches to sample expansion.

Table 8: Performance Metrics – Ridge Regression Labeling

Model	**CV R^2 Mean (Synthetic)**	**CV R^2 Std (Synthetic)**	**CV MSE Mean (Synthetic)**	**Test R^2 (Original)**	**Test MSE (Original)**	**Test VAF (Original)**
Ridge	0.9352	0.0092	0.8340	0.7825	3.6142	0.7827
Lasso	0.9347	0.0096	0.8401	0.7783	3.6826	0.7786
ElasticNet	0.7484	0.0248	3.2639	0.5776	7.0171	0.5787
Random Forest	0.9392	0.0102	0.7833	0.7663	3.8832	0.7663
Gradient Boosting	0.9749	0.0042	0.3227	0.8156	3.0629	0.8157
Extra Trees	0.8649	0.0195	1.7434	0.7122	4.7816	0.7128
MLP	0.9676	0.0082	0.4179	0.8084	3.1831	0.8091
SVR	0.9771	0.0054	0.2934	0.8153	3.0689	0.8154
ZeroR	-0.0058	0.0028	13.0774	-0.0047	16.6920	0.0000

Table 9: Performance Metrics – Gaussian Noise Labeling

Model	R^2 Mean (Train)	R^2 Std (Train)	MSE Mean (Train)	Test R^2 (Test)	Test MSE (Test)	Test VAF (Test)
Ridge	0.8182	0.0152	2.9544	0.8200	2.9899	0.8204
Lasso	0.8170	0.0155	2.9743	0.8184	3.0166	0.8188
ElasticNet	0.5831	0.0383	6.7863	0.5747	7.0662	0.5774
Random Forest	0.9505	0.0085	0.8039	0.9565	0.7222	0.9566
Gradient Boosting	0.9831	0.0039	0.2746	0.9841	0.2634	0.9841
Extra Trees	0.8358	0.0268	2.6580	0.8396	2.6644	0.8406
MLP	0.9676	0.0056	0.5248	0.9568	0.7177	0.9569
SVR	0.8915	0.0159	1.7601	0.8913	1.8056	0.8915
ZeroR	-0.0103	0.0076	16.4882	-0.0080	16.7470	0.0000

Table 10: Performance Metrics – Random Forest Regression Labeling

Model	R^2 Mean (Train)	R^2 Std (Train)	MSE Mean (Train)	Test R^2 (Test)	Test MSE (Test)	Test VAF (Test)
Ridge	0.8261	0.0103	2.0391	0.7674	3.8640	0.7698
Lasso	0.8255	0.0111	2.0454	0.7633	3.9329	0.7657
ElasticNet	0.6355	0.0321	4.2635	0.5264	7.8689	0.5329
Random Forest	0.8743	0.0177	1.4620	0.8081	3.1885	0.8105
Gradient Boosting	0.9386	0.0126	0.7118	0.8900	1.8272	0.8922
Extra Trees	0.7877	0.0290	2.4714	0.6999	4.9853	0.7052
MLP	0.9330	0.0207	0.7737	0.8984	1.6885	0.9033
SVR	0.9245	0.0105	0.8816	0.8572	2.3724	0.8592
ZeroR	-0.0103	0.0061	11.8555	-0.0138	16.8432	0.0000

In this work, three different synthetic generation methods were used to evaluate the quality of the forecasting models: Ridge Regression Labeling, Gaussian Noise Labeling, and Random Forest Regression Labeling. Each of these strategies generated synthetic training data, which was cross-validated (with R^2 and MSE estimates), and then the ability of the models to generalize to the original dataset was assessed. In addition to the main metrics such as R^2 and MSE, the Test VAF (Variance Accounted For) metric was also calculated, which allows us to

judge how well the model explains the variance of the target variable, which actually correlates with the R^2 value.

When using the Ridge Regression Labeling method, linear models such as Ridge and Lasso showed high results: the cross-validation R^2 was about 0.935, and the test R^2 was about 0.782. Test VAF was comparable to the test R^2, indicating robustness of the explained variance. ElasticNet performed slightly worse, with a test R^2 of around 0.578, indicating weaker generalization ability in this configuration. Among the nonlinear models and ensemble methods, Gradient Boosting performed outstandingly with a test R^2 of around 0.816, while SVR and MLP also showed comparable values (around 0.815 and 0.808, respectively). Interestingly, Random Forest and Extra Trees performed more modestly in this method, while the mean-predicting ZeroR model did not produce any significant results.

There synthetic data generated for gaussian noise labeling demonstrates the best quality so far, as reflected in the cross-validation R^2 values for the models (CV and Test values are almost similar). Due to this, the models achieve high performance on the test set: the test R^2 for Ridge and Lasso reaches around 0.820, and ensemble models such as Random Forest, Gradient Boosting, and even MLP demonstrate test R^2 of around 0.956–0.984. Since the Test VAF is almost identical to the Test R^2, it can be concluded that the data quality allows the models to generalize well on the original dataset.

The Random Forest Regression Labeling method shows intermediate results. The cross-validation metrics for the Ridge and Lasso models are around 0.826, but their test R^2 values are slightly lower (around 0.767 and 0.763, respectively). ElasticNet still demonstrates poor generalization ability with a test R^2 of around 0.526. Among the ensemble methods, Gradient Boosting shows good results (test R^2 is about 0.890), and MLP and SVR models also demonstrate satisfactory performance. It is worth noting that for this synthetic generation method, Extra Trees turned out to be less robust, which is reflected in a decrease in test scores to 0.6999.

Now, the ranking of the models' performance is conducted. The ranking was done for all three methods of data synthesis, however the attention will be focused on Gaussian noise labeling since generally the models show higher performance for this method.

The ranking is done based on the following criteria:

R^2 value - The Higher the better;

MSE - The Lower the better;

VAF - The Higher the better.

Now, the ranks for each model are computed:

Table 11: Ranking table for models trained on Ridge Regression Labeling data.

Model	Rank
Gradient Boosting	**1**
SVR	2
MLP	3
Ridge	4
Lasso	5
Random Forest	6
Extra Trees	7
ElasticNet	8

Strong inter-linked correlation between the performance metrics is seen during the comparison. Since the metrics have are mathematically inter-related to each other, they tend to change in unity with each other. This may seem like usage of these metrics may not display the independence in the computation sources, however, all of these metrics are strong standards for assessment of ML models. The same ranking is performed for Gaussian Regression Labeling, where the models display the overall best performance.

Table 12: Ranking table for models trained on Gaussian Noise Labeling data.

Model	Rank
Gradient Boosting	**1**
MLP	2
Random Forest	3
SVR	4
Extra Trees	5
Ridge	6
Lasso	7
ElasticNet	8

A change in pattern can be observed when ranking the performance of models trained on the data synthesized by Random Forest Regression labeling. When utilizing this synthesis method, MLP (Multilayer Perceptron) algorithm outperforms the favorite among the models - GBR (Gradient Boosting regressor).

Table 13: Ranking table for models trained on Random Forest Regression Labeling data.

Model	Rank
MLP	**1**
Gradient Boosting	2
SVR	3
Random Forest	4
Ridge	5
Lasso	6
Extra Trees	7
ElasticNet	8

It should be noted that the choice of the synthetic label generation method significantly affects the results of the models. Gaussian Noise Labeling achieves the best test performance for flexible and ensemble models. It's also important to note that Gaussian Noise Labeling method, generates synthetic data with the same noise levels (noise_std was set as 0.2) when replicating the patterns of original data. Ridge Regression Labeling provides stable performance for linear models, although the results of ensemble models may vary. Random Forest Regression Labeling provides moderate performance, remaining an intermediate option between the other two approaches. Such a difference in results highlights that choosing an appropriate synthetic label generation strategy is critical for optimizing forecasting models and the overall generalization quality on real data.

6. DISCUSSION

This study analyzed the data collected from the San Manuel Mine, which provided detailed measurements of rock geomechanical properties and machine performance metrics. The main focus of this study is on comparing the performance of different models in predicting the productivity of roadheader machines (ICR). The models were evaluated using several metrics, such as the coefficient of determination (R^2) and the mean squared error (MSE), and Variance accounted for (VAF). Models were also evaluated and trained by using three synthetic data generation methods: Ridge Regression Labeling, Gaussian Noise Labeling, and Random Forest Regression Labeling.

The obtained results demonstrate that the use of ensemble methods, such as Random Forest, Gradient Boosting and Extra Trees, can effectively model complex nonlinear relationships between rock geomechanical properties and ICR. The results obtained in this study correlate with the ones obtained by the previous researchers. Performance predictions from the work of Ebrahimabadi et al. in 2019 also displayed that Ensemble methods demonstrated highest forecasting ability.

In this study, Ensemble methods, such as Random Forest and Gradient Boosting showed the best results in terms of R^2, VAF and MSE metrics. Models trained using these methods demonstrated the highest R^2 values, indicating high prediction accuracy. Among all models, highest R^2 and lowest MSE were obtained using the Gradient Boosting model, confirming its superior performance on this task. At the same time, linear models such as Ridge and Lasso performed noteably worse, with R^2 around 0.8 and MSE above 3, indicating that they were unable to account for the complex nonlinear dependencies between the features and the target variable, compared to ensemble methods. One possible reason is the penalty factor utilized in the Ridge and Lasso models. Consequently, since the ElasticNet is a combination of Ridge and Lasso algorithms, it’s poor performance can be a result in even higher penalties employed by this model.

Synthetic data generation methods also had varying impacts on the performance of the models. Gaussian Noise Labeling showed the best test results, with models trained on data generated by this method demonstrating near-perfect R^2 and minimal MSE values. This indicates better generalization ability of the models and prevents overfitting. Random Forest Regression

Labeling also showed good results, with high R^2 values and moderate MSE values. However, the Ridge Regression Labeling method yielded worse results, especially for linear models.

Particular attention is paid to the issue of availability of large datasets. As stated in the study, collecting extensive datasets is associated with high costs, as well as difficulties in exchanging information between companies, since the data is often a commercial secret. This circumstance forces us to resort to data synthesis methods, which allows us to expand the training sample and improve the generalization ability of models. The method of generating synthetic data using Gaussian noise showed the best results, significantly increasing the accuracy of forecasts, but it should be noted that synthesis cannot guarantee 100% compliance with real operating conditions. In the future, it will be useful to continue working on improving synthetic data generation methods so that they more accurately reflect real-world conditions.

In addition, the use of synthetic data is accompanied by a number of limitations. Despite the improvement in metrics, models trained on a synthetically expanded sample may not fully reflect all the features of real production conditions. Also, as it was mentioned in the Sandbak's study (Sandbak, 1985) the geology of San Manuel is exceptionally complex, and obviously the conditions will vary for other mine sites. Taking this into account, the models will require further fine tuning when applied for other tunneling related projects.

Despite the existing limitations, the obtained results show a high potential for applying machine learning to predict the performance of roadheader machines. The application of ensemble methods such as Gradient Boosting proved to be the most effective, and these models can be used to improve optimization processes in the mining and construction industries. Further improvements can be achieved by more careful tuning of the hyperparameters of the models, as well as by applying more sophisticated data generation methods and improving the interpretability of the models.

7. CONCLUSIONS AND RECOMMENDATIONS

The results obtained from applying machine learning techniques to the data obtained from San Manuel mine confirm the high potential of using machine learning methods for predicting the performance of roadheader machines (ICR) in conditions of limited data. In conditions where obtaining extensive data sets is expensive and data is rarely provided by companies for open use, data synthesis becomes a necessary tool for improving the quality of models. The use of synthetic sample expansion using Gaussian noise significantly improved the R^2 and MSE indicators, which indicates the high efficiency of this approach for training models.

However, it should be taken into account that data synthesis cannot completely replace real observations, and the obtained models, despite high accuracy on the test sample, require additional verification in real conditions. The limited availability of input data and heterogeneity of geological conditions remain key issues requiring further research. Further work can be aimed at improving the methods of generating synthetic data, integrating with real data flows through monitoring systems, and increasing the interpretability of models using decision explanation methods (e.g., SHAP).

The practical significance of the study lies in the possibility of using the developed models to optimize the planning and economic assessment of work in the mining and civil construction industries. Highly accurate ICR forecasting allows reducing the risks associated with overestimation of productivity and making more informed decisions when choosing equipment and mining technologies. In the long term, further integration of data from real production, improvement of synthesis methods, and adaptation of algorithms to specific field conditions can lead to the creation of universal solutions capable of providing stable and reliable forecasting in a wide range of geological conditions.

REFERENCES

1. **Balci, C., Demircin, M. A., Copur, H., Tuncdemir, H. (2004).** *Estimation of optimum specific energy based on rock properties for assessment of roadheader performance (567BK)*. https://hdl.handle.net/10520/AJA0038223X_2978

2. **Bilgin, N., Seyrek, T., Erding, E., & Shahriar, K. (1990).** Roadheaders glean valuable tips for Istanbul Metro. *Tunnels & Tunnelling International*, *22*(10).

3. **Bilgin, N., Seyrek, T., & Shahriar, K. (1988).** Roadheader performance in Istanbul. Golden Horn clean-up contributes valuable data. *Tunnels & tunnelling*, *20*(6), 41–44.

4. **Bilgin, N., Kuzu, C., Eskikaya, S., & Özdemir, L. (1997).** Cutting performance of jack hammers and roadheaders in Istanbul Metro drivages. In *World tunnel congress* (Vol. 97, No. 1, pp. 455–60).

5. **Choudhary, R. & H. K. Gianey. (2017).** Comprehensive Review On Supervised Machine Learning Algorithms. In *2017 International Conference on Machine Learning and Data Science (MLDS)*, Noida, India, pp. 37–43, doi: 10.1109/MLDS.2017.11.

6. **Chang, W., & Peng, X. (2020).** Hybrid evolutionary approaches for rock cutting optimization in small data scenarios. *International Journal of Rock Mechanics & Mining Sciences*, *134*, 104456.

7. **Ciccu, R., & Grosso, B. (2014).** Improvement of disc cutter performance by water jet assistance. *Rock Mechanics and Rock Engineering*, *47*(2), [Page Range]. https://doi.org/10.1007/s00603-013-0433-4

8. **Copur, H., Ozdemir, L., & Rostami, J. (1998).** Roadheader applications in mining and tunneling. *Mining Engineering*, *50*.

9. **Ebrahimabadi, A., Goshtasbi, K., Shahriar, K., & Cheraghi Seifabad, M. (2011).** A model to predict the performance of roadheaders based on the Rock Mass Brittleness Index. *Journal of The Southern African Institute of Mining and Metallurgy*, *111*(5), 355–364.

10. **Ebrahimabadi, A., Goshtasbi, K., Shahriar, K., & Cheraghi Seifabad, M. (2011).** Predictive models for roadheaders' cutting performance in coal measure rocks. *Yerbilimleri, 32*(2), 89–104.

11. **Ebrahimabadi, A., Goshtasbi, K., Shahriar, K., & Cheraghi Seifabad, M. (2012).** A universal model to predict roadheaders' cutting performance. *Archives of Mining Sciences, 57*(4), 1015–1026.

12. **Gehring, K. H. (1989).** A cutting comparison. *Tunnels and Tunnelling;(UK), 21*(11).

13. **Ghasemi, A., Azadeh, A., & Akbari, R. (2020).** Application of deep learning methods for performance prediction of mechanical miners. *International Journal of Mining Science and Technology, 30*(2), 145–156.

14. **Huff, C. F., & Varnado, S. G. (1980).** Recent Developments in Polycrystalline Diamond-Drill-Bit Design.

15. **Keleş, S. (2005).** Cutting performance assessment of a medium weight roadheader at Cayırhan coal mine (Master's thesis, Middle East Technical University (Turkey)).

16. **Kogelmann, W. J., & Schenck, G. K. (1983).** Recent North American advances in boom-type tunnelling machines. *Transactions of the Institution of Mining and Metallurgy, Section A, 92*, A155–A165.

17. **Krzysztof, K., & Piotr, M. (2019).** Methods of Mechanical Mining of Compact-Rock—A comparison of efficiency and energy consumption. *Energies, 12*(18), 3562. https://doi.org/10.3390/en12183562

18. **Li, M., & Gao, Y. (2021).** Laser scanning and real-time data fusion for automated roadheader cutting. *Mining Technology, 45*(3), 215–228.

19. **Li, Y., Li, M., & Gao, Y. (2022).** Bayesian regression for uncertain geological conditions in roadheader performance modeling. *Tunneling and Underground Space Technology, 123*, 103575.

20. **Mahdevari, S., Shahriar, K., Yagiz, S., & Akbarpour Shirazi, M. (2014).** A support vector regression model for predicting tunnel boring machine penetration rates. *International Journal of Rock Mechanics and Mining Sciences, 72*, 214–229.

21. **Neil, D. M., Rostami, J., Ozdemir, L., & Gertsch, R. (1994).** Production estimating techniques for underground mining using roadheaders. *Preprints-Society of Mining Engineers of Aime.*

22. **Ozdemir, L. (1977).** Development of theoretical equations for predicting tunnel boreability.

23. **Salsani, A., Daneshian, J., Shariati, S., Yazdani-Chamzini, A., & Taheri, M. (2014).** Predicting roadheader performance by using artificial neural network. *Neural Computing & Applications*, *24*(7–8), 1823–1831.

24. **Sandbak, L. A. (1985).** Roadheader drift excavation and geomechanical rock classification at San Manuel, Arizona. In *Rapid Excavation and Tunnelling Conference, New York* (Vol. 2, pp. 902–916).

25. **Sandvik. (n.d.).** Sandvik MR341/MR361 specification sheet. *Sandvik Mining and Rock Technology.*

26. **Schneider, H. J., Özgür, N., & Palacios, C. M. (1988).** Relationship between alteration, rare earth element distribution, and mineralization of the Murgul copper deposit, northeastern Turkey. *Economic Geology*, *83*(6), 1238–1246.

27. **Seker, S. E., & Ocak, I. (2019).** Performance prediction of roadheaders using ensemble machine learning techniques. *Neural Computing and Applications*, *31*(2), 1103–1116.

28. **Thuro, K. P. R. J., & Plinninger, R. J. (1999).** Roadheader excavation performance-geological and geotechnical influences. In *ISRM Congress* (pp. ISRM-9CONGRESS). ISRM.

29. **Tucker, R. H. (1985).** Improvement of potential in the mining development and tunnelling systems in the National Coal Board. *Min. Eng.(London);(United Kingdom)*, *144*(285).

30. **VMT GmbH. (n.d.).** TUnIS Navigation Roadheader: Product info. Retrieved August 2024, from vmt.global/tunnelling/TUnIS_Navigation

31. **Yagiz, S., Gokceoglu, C., Sezer, E. A., & Iplikci, S. (2009).** Application of two non-linear prediction tools to the estimation of tunnel boring machine performance. *Engineering Applications of Artificial Intelligence*, *22*(4–5), 808–814.

32. **Yagiz, S., Sezer, E. A., & Gokceoglu, C. (2011).** Artificial neural networks and nonlinear regression techniques to assess the influence of slake durability cycles on the prediction of uniaxial compressive strength and modulus of elasticity for carbonate rocks. *Numerical Analysis and Geomechanics, 36*(14), 1636–1650. https://doi.org/10.1002/nag.1066

33. **Zhang, K., Wang, B., & Li, T. (2022).** Transfer learning approach to roadheader performance analysis using multi-sensor data. *Journal of Mining & Environment, 43*(1), 37–52.

34. **Zhou, J., Qiu, Y., Zhu, S., Armaghani, D. J., Li, C., Nguyen, H., & Yagiz, S. (2021).** Optimization of support vector machine through the use of metaheuristic algorithms in forecasting TBM advance rate. *Engineering Applications of Artificial Intelligence, 97*, 104015.

APPENDICES

Github Link to access the Python code.
https://github.com/Asekebaseke/BSc-Thesis/blob/main/README.md

Table 1: Ridge Regression ICR values

ICR actual	ICR Ridge Regression Labeling	ICR Gaussian Noise Labeling	ICR Random Forest Regression Labeling
8,309	8,662	8,120	8,737
6,492	7,708	8,201	8,683
8,547	8,167	7,643	9,182
12,115	14,765	13,475	14,390
15,016	14,351	14,290	14,261
17,406	16,184	16,515	16,112
18,789	17,272	16,428	16,784
18,491	18,870	19,572	18,599
19,199	17,379	18,525	17,501
20,060	20,304	20,797	19,707
21,099	18,577	18,551	18,048
21,807	19,788	21,063	18,924
22,167	18,771	20,164	18,155
18,997	16,240	15,973	15,827
13,809	15,432	15,456	15,272
12,317	17,094	16,569	16,505
14,357	17,677	17,680	17,269
16,567	16,939	17,074	16,609
19,450	18,152	18,829	17,813
16,446	16,692	15,947	16,025
13,952	13,084	14,143	13,479
14,665	12,774	12,817	12,916
12,182	12,017	13,222	12,246
11,044	12,774	12,817	12,916
12,596	11,724	10,376	11,682
12,790	15,662	14,133	14,984
13,151	13,846	13,867	13,892
14,193	14,058	13,553	13,948

Table 2: Lasso Regression ICR values

ICR actual	ICR Ridge Regression Labeling	ICR Gaussian Noise Labeling	ICR Random Forest Regression Labeling
8,309	8,684	8,216	8,783
6,492	7,684	8,142	8,618
8,547	8,199	7,695	9,131
12,115	14,850	13,627	14,481
15,016	14,370	14,321	14,303
17,406	16,146	16,676	16,095
18,789	17,348	16,568	16,839
18,491	18,839	19,432	18,527
19,199	17,389	18,542	17,444
20,06	20,207	20,759	19,643
21,099	18,643	18,618	18,057
21,807	19,748	20,939	18,892
22,167	18,689	19,817	18,042
18,997	16,263	15,929	15,831
13,809	15,431	15,410	15,264
12,317	17,155	16,633	16,549
14,357	17,695	17,713	17,247
16,567	16,939	17,038	16,584
19,45	18,011	18,821	17,706
16,446	16,743	16,024	16,083
13,952	12,985	13,907	13,398
14,665	12,760	12,766	12,929
12,182	11,933	13,049	12,241
11,044	12,760	12,766	12,929
12,596	11,790	10,518	11,775
12,79	15,701	14,363	15,036
13,151	13,834	13,804	13,890
14,193	14,089	13,584	13,990

Table 3: ElasticNet ICR values

ICR actual	ICR Ridge Regression Labeling	ICR Gaussian Noise Labeling	ICR Random Forest Regression Labeling
8,309	9,077	9,093	9,539
6,492	9,815	9,817	10,381
8,547	10,871	10,836	11,352
12,115	15,618	15,577	15,323
15,016	14,927	14,946	14,805
17,406	15,452	15,540	15,325
18,789	16,732	16,734	16,329
18,491	16,116	16,186	15,946
19,199	15,729	15,853	15,657
20,06	17,822	17,827	17,312
21,099	17,345	17,351	16,784
21,807	17,829	17,773	17,017
22,167	16,728	16,623	16,091
18,997	16,149	16,075	15,709
13,809	15,534	15,515	15,278
12,317	16,758	16,711	16,228
14,357	16,338	16,358	15,947
16,567	16,086	16,083	15,727
19,45	15,689	15,748	15,464
16,446	16,689	16,602	16,125
13,952	14,029	14,006	14,059
14,665	14,105	14,086	14,086
12,182	13,060	13,081	13,181
11,044	14,105	14,086	14,086
12,596	13,543	13,502	13,556
12,79	16,369	16,163	15,768
13,151	14,754	14,726	14,634
14,193	15,013	14,987	14,840

Table 4: Random Forest ICR values

ICR actual	ICR Ridge Regression Labeling	ICR Gaussian Noise Labeling	ICR Random Forest Regression Labeling
8,309	8,451	8,209	8,339
6,492	7,837	6,884	7,951
8,547	7,765	8,493	9,918
12,115	15,008	13,430	14,575
15,016	13,995	14,478	13,786
17,406	16,158	17,470	16,776
18,789	17,012	18,338	16,542
18,491	18,077	18,248	17,305
19,199	17,552	18,148	17,605
20,06	20,265	19,900	18,662
21,099	18,483	20,502	17,712
21,807	20,457	21,815	20,246
22,167	18,532	21,759	18,774
18,997	16,545	17,695	16,315
13,809	15,489	13,720	14,624
12,317	17,151	14,366	16,261
14,357	17,250	14,993	15,818
16,567	17,214	16,999	15,793
19,45	17,355	18,993	17,510
16,446	17,266	16,713	16,274
13,952	13,268	13,700	13,502
14,665	13,264	13,367	13,467
12,182	12,771	12,279	12,140
11,044	13,264	13,367	13,467
12,596	12,352	12,960	13,138
12,79	15,483	13,379	14,105
13,151	13,609	13,650	13,761
14,193	13,773	13,718	13,802

Table 5: Gradient Boosting ICR values

ICR actual	ICR predicted Ridge Regression Labeling	ICR Gaussian Noise Labeling	ICR Random Forest Regression Labeling
8,309	8,475	8,400	8,438
6,492	7,889	6,495	7,887
8,547	7,780	8,688	9,627
12,115	14,364	12,269	14,034
15,016	14,022	15,006	14,014
17,406	16,221	17,868	17,495
18,789	16,762	18,819	17,106
18,491	19,306	18,299	18,018
19,199	18,252	19,101	18,765
20,06	20,650	20,237	19,465
21,099	18,730	20,913	18,356
21,807	20,660	21,745	20,481
22,167	19,525	22,220	20,068
18,997	16,070	18,778	17,237
13,809	15,641	13,748	14,029
12,317	16,867	12,399	15,506
14,357	17,557	14,532	15,328
16,567	17,385	16,816	15,641
19,45	18,616	19,670	19,142
16,446	16,765	16,557	16,350
13,952	13,200	13,949	13,351
14,665	12,888	12,942	12,887
12,182	12,310	11,903	11,514
11,044	12,888	12,942	12,887
12,596	11,139	12,588	12,585
12,79	14,788	13,229	13,014
13,151	13,407	13,703	13,639
14,193	13,707	14,077	13,696

Table 6: ExtraTrees ICR values

ICR actual	ICR Ridge Regression Labeling	ICR Gaussian Noise Labeling	ICR Random Forest Regression Labeling
8,309	8,451	8,257	8,390
6,492	8,367	6,559	8,494
8,547	9,021	9,251	10,226
12,115	15,384	14,621	15,218
15,016	14,642	14,655	14,808
17,406	15,527	16,155	15,710
18,789	15,971	16,300	15,807
18,491	17,152	17,357	16,881
19,199	16,064	17,553	16,606
20,06	19,490	19,562	18,532
21,099	17,399	18,752	17,007
21,807	19,800	21,163	19,473
22,167	17,749	20,301	17,466
18,997	16,193	15,862	15,784
13,809	15,546	14,715	15,158
12,317	16,514	15,538	15,999
14,357	16,344	15,775	15,694
16,567	16,325	15,689	15,614
19,45	16,682	17,713	16,221
16,446	16,615	16,390	16,153
13,952	14,165	14,204	14,125
14,665	14,139	13,886	14,109
12,182	13,210	12,667	12,873
11,044	14,139	13,886	14,109
12,596	13,493	13,481	13,878
12,79	16,136	14,758	15,245
13,151	14,571	14,425	14,448
14,193	14,617	14,646	14,694

Table 7: MLP ICR values

ICR actual	ICR Ridge Regression Labeling	ICR Gaussian Noise Labeling	ICR Random Forest Regression Labeling
8,309	8,359	8,447	8,263
6,492	7,691	6,524	7,629
8,547	7,725	8,750	9,514
12,115	14,520	12,395	13,700
15,016	14,087	15,211	14,311
17,406	16,041	17,846	17,485
18,789	16,521	18,829	17,940
18,491	19,391	18,353	18,019
19,199	17,958	19,156	18,899
20,06	20,231	20,370	19,068
21,099	18,652	20,670	18,101
21,807	20,546	21,813	20,613
22,167	19,414	22,514	19,921
18,997	16,057	17,813	17,203
13,809	15,300	14,318	14,243
12,317	16,776	13,873	15,151
14,357	17,641	14,867	15,384
16,567	16,709	15,789	15,550
19,45	18,347	19,774	18,817
16,446	16,212	15,589	16,004
13,952	13,492	13,978	13,301
14,665	12,854	12,906	12,812
12,182	12,176	11,960	11,460
11,044	12,854	12,906	12,812
12,596	11,419	12,355	12,251
12,79	14,747	13,400	12,834
13,151	13,966	14,068	13,596
14,193	14,049	13,881	13,917

Table 8: Ridge Regression ICR values

ICR actual	ICR Ridge Regression Labeling	ICR Gaussian Noise Labeling	ICR Random Forest Regression Labeling
8,309	8,322	8,421	8,345
6,492	7,688	6,689	7,714
8,547	7,533	8,594	9,546
12,115	14,540	13,928	14,096
15,016	13,960	14,896	14,439
17,406	15,882	17,287	16,785
18,789	16,624	17,637	17,147
18,491	19,369	18,681	18,349
19,199	17,962	18,978	18,452
20,06	20,649	20,255	19,204
21,099	18,602	19,340	18,615
21,807	20,825	21,533	20,642
22,167	19,483	21,595	19,849
18,997	16,464	15,497	15,831
13,809	15,374	14,441	14,464
12,317	16,881	16,292	16,532
14,357	17,559	15,760	15,919
16,567	16,905	15,379	15,527
19,45	18,364	18,967	18,575
16,446	16,555	16,358	16,560
13,952	13,283	14,150	13,506
14,665	12,705	13,470	13,119
12,182	12,234	12,129	11,644
11,044	12,705	13,470	13,119
12,596	11,625	12,499	12,539
12,79	14,959	13,575	13,186
13,151	13,922	13,877	13,663
14,193	13,979	13,740	13,674

Table 9: ZeroR ICR values

ICR actual	ICR Ridge Regression Labeling	ICR Gaussian Noise Labeling	ICR Random Forest Regression Labeling
8,309	14,935	14,910	14,736
6,492	14,935	14,910	14,736
8,547	14,935	14,910	14,736
12,115	14,935	14,910	14,736
15,016	14,935	14,910	14,736
17,406	14,935	14,910	14,736
18,789	14,935	14,910	14,736
18,491	14,935	14,910	14,736
19,199	14,935	14,910	14,736
20,06	14,935	14,910	14,736
21,099	14,935	14,910	14,736
21,807	14,935	14,910	14,736
22,167	14,935	14,910	14,736
18,997	14,935	14,910	14,736
13,809	14,935	14,910	14,736
12,317	14,935	14,910	14,736
14,357	14,935	14,910	14,736
16,567	14,935	14,910	14,736
19,45	14,935	14,910	14,736
16,446	14,935	14,910	14,736
13,952	14,935	14,910	14,736
14,665	14,935	14,910	14,736
12,182	14,935	14,910	14,736
11,044	14,935	14,910	14,736
12,596	14,935	14,910	14,736
12,79	14,935	14,910	14,736
13,151	14,935	14,910	14,736
14,193	14,935	14,910	14,736

Blank page