REPUBLIC OF TURKEY
YILDIZ TECHNICAL UNIVERSITY
DEPARTMENT OF COMPUTER ENGINEERING

SOURCE LOCALIZATION OF THE SOUND OF EARTHQUAKE VICTIMS

18011004 — Umut GÜZEL
18011070 — Mehmet ÇALOĞLU

SENIOR PROJECT

Advisor
Assoc. Prof. Ali Can KARACA

Aralık, 2024

© All rights reserved to Yıldız Technical University, Computer Engineering Department.

ACKNOWLEDGEMENTS

We would like to express our appreciation to Yildiz Techical University and its individuals. Firstly we are grateful to the our project advisor Assoc. Prof. Ali Can KARACA for his guidance and encouragements during the project. We also grateful to Research Assistant Idris Demir for the informations and feedbacks that he provided us.

We would like to thank to our managers and collagues in SANLAB Simulation for their patiance and encouragement during the project.

Lastly we are thankful to our friends and families for their support.

Umut GÜZEL
Mehmet ÇALOĞLU

TABLE OF CONTENTS

LIST OF SYMBOLS

km	Kilometer
ϵ	Epsilon-Tolerance
min	Minimum
w	Weight Vector
C	Regularization Parameter
ϕ	Mapping Function
ζ	Slack Variable of Support Vector Regression
n	Number of Samples
b	Bias Term
$\sum$	Summary
y	Target Value Vector

LIST OF ABBREVIATIONS

SSL	Sound Source Localization
ILD	Interaural Level Difference
ITD	Interaural Time Difference
TDOA	Time Difference of Arrival
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
DGL	Deep Reinforcement Learning
MFCC	Mel-frequency Cepstral Coefficients
STFT	Short-Term Fourier Transform
ICA	Independent Component Analysis
2D	2-Dimenional
3D	3-Dimensional
kHz	kiloHertz
Ch	Channel
SSD	Solid State Disc
RAM	Random Access Memory
ML	Machine Learning
DL	Deep Learning
Corr	Correlation
SVM	Support Vector Machine
SVR	Support Vector Regression
ReLU	Rectified Linear Unit
MSE	Mean Square Error

MAE	Mean Absolute Error
GCC-PHAT	Generalized Cross-Correlation with Phase Transform
LTSM	Long Short-Term Memory

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

SOURCE LOCALIZATION OF THE SOUND OF EARTHQUAKE VICTIMS

Umut GÜZEL
Mehmet ÇALOĞLU

Department of Computer Engineering
Senior Project

Advisor: Assoc. Prof. Ali Can KARACA

Natural disasters, especially earthquakes, can be devastating for the countries that do not have proper infrastructures and precautions. When this disasters occur, stranded people should be rescued from the disaster area. Finding victims in wreckage is challenge for rescue missions.

In this project, our goal is that localization of the earthquake victims sounds. 8-channel sound database will be used which is recorded by developed circular microphone array mechanism in the same medium at different locations. Our goal is to use different approaches to solve problem known as "Sound Source Localization(SSL)". Machine learning and Deep Learning methods will be applied and their accuracies and performances will be compared and evaluated.

Based on the natural ability of ears of human and animals, various SSL methods developed and implemented to interactive applications. Accuracy of SSL process is related to binaural features that based on phase and amplitude. This features are ILD and ITD information. While ILD information represents amplitude differences, ITD represents phase difference.

ITD feature is used in this project. The input of 8 channel microphone preprocessed and phase difference array obtained for every 2 channel by using similar method that traditional algorithms used. With this information, the machine learning methods that requires complex features such as Support Vector Regression and Random Forest

Regression implemented succesfully. Also CNN method is used because of ability to handle multidimensional data and robustness. The other advantage of CNN is that it requires less complex Network structure for the features that are used in this project.

In conclusion all of out proposed solutions for this problem has better accuracy for elevation polar axis. But in azimuth, GCC-PHAT method has better accuracy.

Keywords: Localization, binaural, SSL, amplitude, phase, ILD, ITD, Deep Learning, CNN, machine learning, SVR, Random Forest Regression

ÖZET

DEPREMZEDELERİN SESLERİNDEN KONUMLARININ TESPİTİ

Umut GÜZEL
Mehmet ÇALOĞLU

Bilgisayar Mühendisliği Bölümü
Bitirme Projesi

Danışman: Doç. Dr. Ali Can KARACA

Doğal afetler, özellikle de depremler, uygun altyapıya ve önlemlere sahip olmayan ülkeler için yıkıcı olabilir. Bu felaketler meydana geldiğinde, mahsur kalan insanların felaket bölgesinden kurtarılması gerekir. Enkazdaki kurbanları bulmak kurtarma misyonları için zorlu bir görevdir.

Bu projede amacımız depremzedelerin seslerinin lokalizasyonunu yapmaktır. Geliştirilen dairesel mikrofon dizilimi mekanizması ile aynı ortamda farklı lokasyonlarda kaydedilen 8 kanallı ses veritabanı kullanılacaktır. Amacımız "Ses Kaynağı Lokalizasyonu (SSL)" olarak bilinen problemi çözmek için farklı yaklaşımlar kullanmaktır. Makine öğrenmesi ve Derin Öğrenme yöntemleri uygulanacak ve doğrulukları ve performansları karşılaştırılacak ve değerlendirilecektir.

İnsan ve hayvanların kulaklarının doğal yeteneklerine dayanarak, çeşitli SSL yöntemleri geliştirilmiş ve etkileşimli uygulamalara uygulanmıştır. SSL işleminin doğruluğu, faz ve genliğe dayalı binaural özelliklerle ilgilidir. Bu özellikler ILD ve ITD bilgileridir. ILD bilgisi genlik farklılıklarını temsil ederken, ITD faz farkını temsil eder.

Bu projede ITD özelliği kullanılmıştır. 8 kanallı mikrofon girişi ön işlemden geçirilmiş ve geleneksel algoritmaların kullandığı benzer yöntem kullanılarak her 2 kanal için faz farkı dizisi elde edilmiştir. Bu bilgi ile Destek Vektör Regresyonu ve Rastgele Orman Regresyonu gibi karmaşık özellikler gerektiren makine öğrenmesi yöntemleri başarıyla

uygulanmıştır. Ayrıca CNN yöntemi çok boyutlu verileri işleyebilmesi ve sağlamlığı nedeniyle kullanılmıştır. CNN'in bir diğer avantajı ise bu projede kullanılan özellikler için daha az karmaşık Ağ yapısı gerektirmesidir.

Sonuç olarak, bu problem için önerilen tüm çözümler yükseklik kutup ekseni için daha iyi doğruluğa sahiptir. Ancak azimutta, GCC-PHAT yöntemi daha iyi doğruluğa sahiptir.

Anahtar Kelimeler: Konumlandırma, Binoral, SSL, Genlik, Faz, ILD, ITD, Derin Öğrenme, CNN, Makine Öğrenmesi, SVR, Random Forest Regression

1
INTRODUCTION

Natural disasters, especially earthquakes, can be devastating for the countries that do not have proper infrastructures and precautions. When this disasters occur, stranded people should be rescued from the disaster area. Finding victims in wreckage is challenge for rescue missions. Disaster victims are usually in critical health condition due to harsh environmental conditions and injuries. So search and rescue missions are race against time situations. Since every second is important, rescue teams must be guided well because disaster may spread large area. As an example, The earthquake disaster in Turkey on February 6, disaster spreads 11 province and 108.812km^2 size area. So transportation problems and lack of rescue team personnel is a challenge for operations. In these situations, rescue teams use equipments that leads the to victims such as sonar based localization devices, thermal measurement devices. In this project, our goal is that localization of the earthquake victims sounds.

1.1 Project Goal

The main goal of the project is to localization of sound source. 8-channel sound database will be used which is recorded by developed circular microphone array mechanism in the same medium at different locations. Our goal is to use different approaches to solve problem known as "Sound Source Localization(SSL)". Machine learning and Deep Learning methods will be applied and their accuracies and performances will be compared and evaluated.

1.2 Preexamination

SSL is important for humans and some animals for interacting with the environment. All healthy human and most of the animal has SSL abilities. Based on this natural ability, various SSL methods developed and implemented to interactive applications. Accuracy of SSL process is related to binaural features that based on phase and amplitude. This features are ILD and ITD information.

The ILD information is represents amplitude difference between two sensors. Visual Representation of ILD can be seen in 1.1.

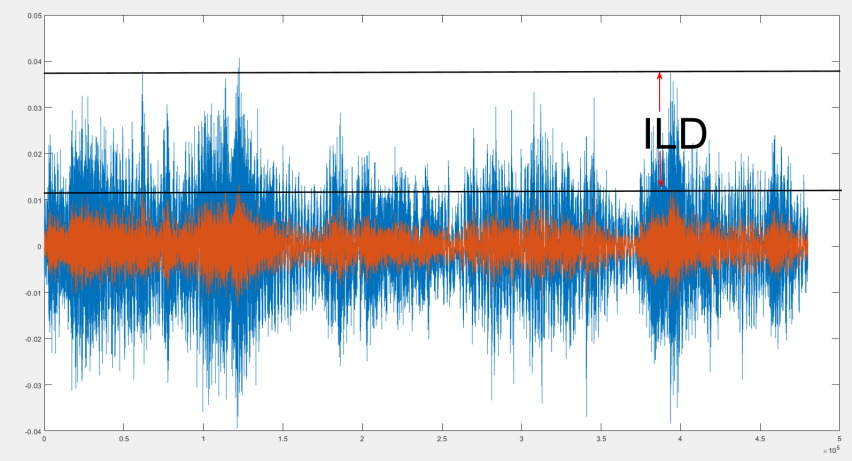


Figure 1.1 ILD representation

On the other hand ITD information is about delay of the signal. If a sound source is not in the equal distance to all sensors, their phases will be different. Visual representation of ITD is given in 1.2.

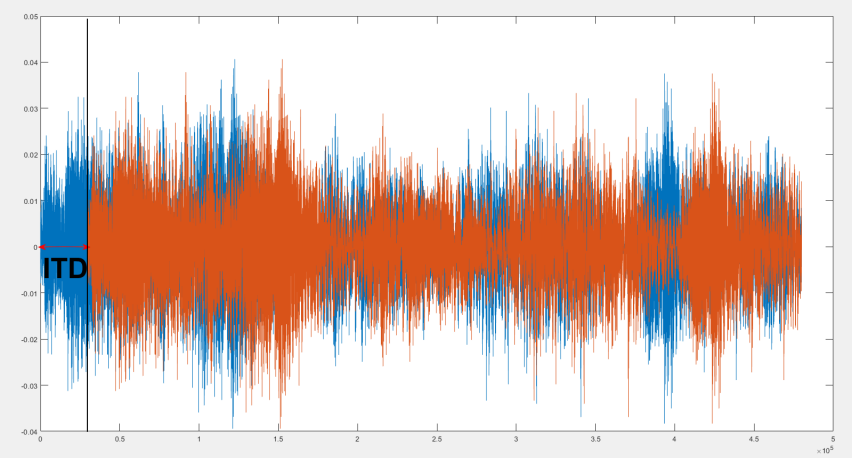


Figure 1.2 ITD representation

In this project, ITD based feature matrix will be created for every record. 8x8 feature matrix given to the machine learning and deep learning models.

All sound records in the dataset is recorded in same room and predefined different locations. When model trained, models performance will be measured with the allocated test records.

2
LITERATURE ANALYSIS

2.1 Traditional Method in Sound Source Localization

Sound source localization is the process of determining the location of a sound source using one or more microphones. This technology has wide applications in many fields such as robotics, automation, security, healthcare, and entertainment. Traditional sound source localization methods are usually based on the physical properties of sound waves and estimate the location of the source using various algorithms.

2.1.1 Time Difference Based Method (Time Difference of Arrival - TDOA)

TDOA is one of the most widely used traditional methods[1] . This method determines the position of the source using the difference between the times the sound wave reaches different microphones. After propagating from the source, sound reaches different microphones at different times. These time differences can be used to calculate the relative distance of the source to the microphones.

The advantages of the TDOA method are:

- Low computational complexity.
- Suitable for real-time applications.
- Hardware requirements are relatively low.

However, the TDOA method also has some disadvantages:

- Performance may be reduced in noisy environments.
- Echo and reflections can make accurate localization difficult.
- Correct synchronization of microphones is important.

2.1.2 Methods Based on Amplitude Difference

Amplitude difference based methods use the difference in amplitude levels of the sound wave at different microphones[1]. The microphone closer to the sound source detects a higher amplitude level. These amplitude differences can be used to determine the location of the source.

Advantages of these methods:

- Simple and easy to implement.

Disadvantages

- Sensitive to noise and ambient conditions.
- Not as accurate as TDOA for distance estimation.

2.1.3 Beamforming Methods

Beamforming is a technique that aims to focus in a specific direction using an array of microphones and suppress sounds from other directions. This method can be used to determine the direction of the sound source.

Advantages of beamforming methods:

- Noise and echo suppression.
- Provides high directionality.

Disadvantages

- Computational complexity is high.
- Requires a large number of microphones.

2.1.4 Limitations of Traditional Methods

Traditional methods can be effective under certain conditions, but they have some limitations:

- **Noisy environments:** Noise can reduce localization accuracy by making it difficult to accurately calculate time and amplitude differences.
- **Echo and reflections:** In enclosed spaces, sound waves can reflect off walls and other surfaces, creating multiple paths. This can negatively affect the performance of traditional methods.
- **Microphone array:** The number and arrangement of microphones affects localization accuracy. More microphones usually give better results, but increase the complexity and cost of the system.

2.2 Deep Learning Based Audio Source Localization

Overcoming the restrictions of conventional audio source localization methods, as well as providing more stable solutions, requires extensive application of deep learning techniques for the last few years [2]. Because the deep learning algorithms are strong in pattern recognition and can operate from enormous datasets, they are able to place sounds not only in a clean but also in a noisy environment where disturbances, echoes, and reflections may occur.

2.2.1 Lightweight R-CNN Model

The "Lightweight R-CNN" model proposed by Mehmet Cem Catalbas and Simon Dobrisek is one of the important breakthroughs in deep learning-based sound localization. This model is to provide very accurate sound location with low computing complexity, which is a key task in disaster situations where computing power and speed are most demanding. [3].

2.2.1.0.1 Motivation and Development Catalbas and Dobrisek present an approach that re-examines the classic iterative SSL problem by converting it into a regression problem, thus making it easier to compute and decreasing latency. This model makes the application of conventional, low-cost microphone arrays to capture a moving source's localization with a high level of accuracy very feasible, and thus very useful in real-time. Using traditional, cost-effective microphone arrays, this model ensures high positional precision in the 3D space with moving sound sources, which is perfect for time-critical applications. Figure 2.1 is a graphical illustration of the main perceptual model of SSL that shows how this R-CNN system facilitates sound localization, which is the main intention of this technique.

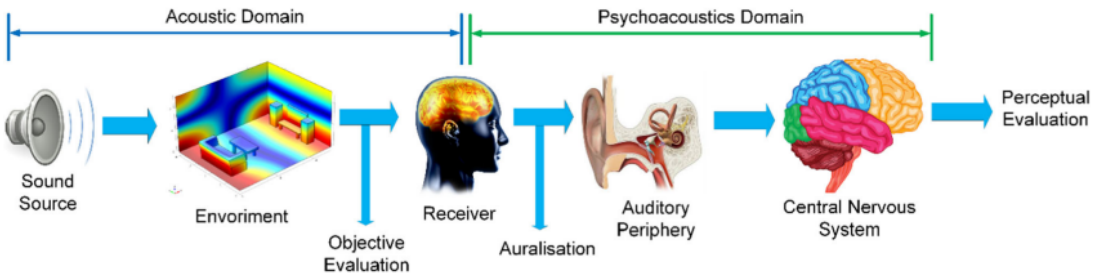


Figure 2.1 General perceptual model of sound source localization as proposed by Catalbas and Dobrisek.

2.2.1.0.2 Technical Details of the Lightweight R-CNN The Lightweight R-CNN model employs a novel deep regression network, which processes distance matrices (derived from time delays between microphones). This network is specifically trained to map these matrices to 3D coordinates of the sound source. Figure 2.2 illustrates the architecture of the lightweight R-CNN, showcasing its convolutional layers and regression output structure. These elements are optimized for SSL in resource-constrained settings; however, the effectiveness of the model can vary significantly depending on several factors. Although the architecture is efficient, it may still face challenges because of the inherent complexity of sound localization tasks.

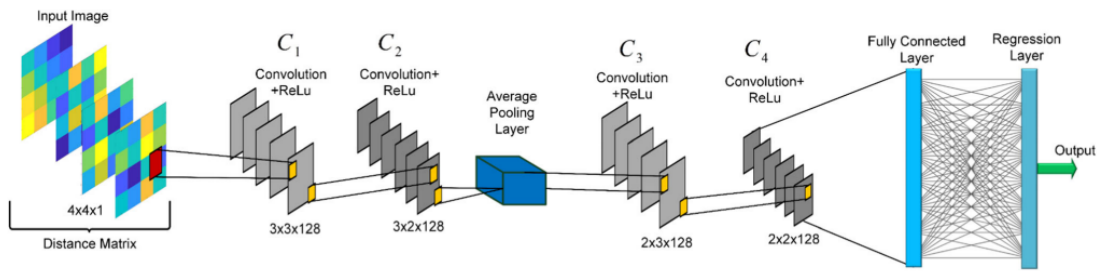


Figure 2.2 Architecture of the lightweight R-CNN model for SSL as presented in the study.

2.2.2 Performance Evaluation of the Lightweight R-CNN

The research conducted by Catalbas and Dobrisek revealed a substantial accuracy improvement thanks to the applied localization technology compared to the typical TDOA approach. The R-CNN which operates at a very light-weight level reduced the average localization error from 45.826 cm in the TDOA method to 16.298 cm, thus it allowed a more precise real-time positioning. The figure 2.3 The performance in this case was actual and predicted positions and the proposed model's accuracy was enhanced which thus made it so.

Test locations	TDOA (cm)			Proposed approach (cm)			Real coordinates (cm)		
	x	y	z	x	y	z	x	y	z
1	339.029	16.060	181.165	351.602	1.650	160.116	342	0	153
2	64.040	543.008	215.735	34.536	539.072	171.765	38	530	172
3	166.054	644.906	218.125	129.046	637.821	178.102	160	660	172
4	559.080	590.779	210.728	526.983	603.505	170.535	530	600	172

Figure 2.3 Comparison of localization accuracy between TDOA-based and Lightweight R-CNN approaches.

2.2.3 Applications in Disaster Response Scenarios

The research of Catalbas and Dobrisek shows that there has been a considerable enhancement in localization accuracy as compared to earlier TDOA-based approaches. The miniaturized R-CNN, in turn, decreased the average localization error from 45.826 cm (in the TDOA approach) to 16.298 cm, thus making a more exact localization real-time possible. The figures presented in Figure 2.3 compare the real and predicted positions, hence wrongness is being propelled less than before. Not to be outbid, the lightweight R-CNN model's efficient architecture and high localization accuracy make it a suitable candidate in the disaster response scenarios. In scenarios of earthquakes, where the locations of the victims have to be determined quickly to save lives, the devices that use this sound localization algorithm can be installed to search and rescue operations teams to effectively navigate vast and challenging environments.

2.2.3.0.1 Dataset and Testing in Real Environments As part of this study, a database of authentic user acoustic 3D signals was created by making recordings from 14 of the predetermined locations. These delay matrices, converted from the voice recordings, simulate a dynamic environment, which is typical of the disaster situations encountered. The success in such tests of this model is a strong indicator of its capability in real-life applications, especially in situations when resources are scanty.

2.2.4 Advantages and Limitations of Lightweight R-CNN

The Lightweight R-CNN model provides several advantages over traditional and other deep learning-based localization methods:

- **High Accuracy in Complex Environments:** Achieves reliable localization even in acoustically challenging environments.

- **Low Computational Complexity:** Designed for efficiency, making it feasible for deployment on devices with limited processing power.

However, the model also has limitations:

- **Challenges with Multiple Sources:** Performance may decrease when multiple sound sources are present.

In summary, Catalbas and Dobrisek's lightweight R-CNN model advances SSL technology, providing an effective, low-complexity solution for real-time applications in resource-limited environments.

2.3 Challenges in Audio Source Localization

Sound source localization is a very complicated process that is influenced by many factors. In situations when the source is singular and at a standstill, as well as when there is no noise or echo in the environment, sound source localization is actually performed much easier. On the other hand, these ideal cases are rarely available in practical conditions, thus, many issues regarding sound source localization have to be faced.

2.3.1 Major Challenges

- **Noise:** Background noise is one of the crucial challenges in sound source localization. Noise distorts the signals received at the microphones, therefore it is difficult to measure the differences in time and amplitude signals which can be synthesized by the microphones. Consequently, it negatively affects the accuracy of localization.
- **Echo and Reflections:** In closed regions, sound rays may bounce off the walls and surface of the area, thus producing several alternatives. These paths of sound may reach microphones at different times with different amplitudes, therefore the place of the sound source is miscalculated.
- **Multiple Sources:** If multiple audio signals are currently active, their signal superposition will show interference. If not, the sources could be separated from each other to make them more distinguishable.
- **Moving Sources:** Should both the sound sources and the microphone be in motion the localization problem becomes even more acute? This is mainly because time differences and amplitude differences are constantly in flux.

- **Microphone Array:** The number and array of microphones also contribute to the localization accuracy. More mics normally give better pictures, but a complex and expensive setup might be needed.

- **Low Cost and Reliability:** The systems that allow for sound localization, particularly during emergency scenes, should be affordable and reliable. The reason is that those places might not have a great deal of resources and yet, they need immediate assistance.

- **Real Time Operation:** In countless cases, the sound source localization which is part of the task must be accomplished in real-time. Slow processing and high latency must be fixed to keep up the speed.

2.3.2 Methods Developed to Overcome the Challenges

- **Noise reduction techniques:** Different signal processing technologies can be utilized to exhibit less noise. For instance, Wiener filtering, spectral subtraction, and independent component analysis (ICA) can be used to suppress the noise.

- **Echo and reflection suppression:** Several techniques have also been crafted to minimize the effect of echo and reflection. For example, beamforming and multichannel echo cancellation can be used.

- **Blind source separation:** Blind source separation techniques can be applied in order to split the multiple sources' signals. These methods target to isolate the signals of sources with no need for previous information.

- **Deep learning:** The ability to detect complex patterns and learn from large datasets that are very deep enables deep systems to perform localization with nearly perfect accuracy both in clean and noisy environments and even in cases of echo problems/reflections.Research In Lightweight R-CNN Model, Catalbas and Dobrisek (2023) put forward effective solutions that allow for a wider and more resilient system built on challenges such as noise and echo[3].

3
SYSTEM ANALYSIS AND FEASIBILITY

In this chapter, project requirements will be specified, main elements and functions will be stated. In addition, in feasibility, the process and the needs of the project will be planned to meet the requirements.

3.1 System Analysis

In general, the system basically takes an 8 channel input and gives $2x1$ output vector which includes azimuth and elevation coordinates of the polar coordinates system as seen in 3.1.

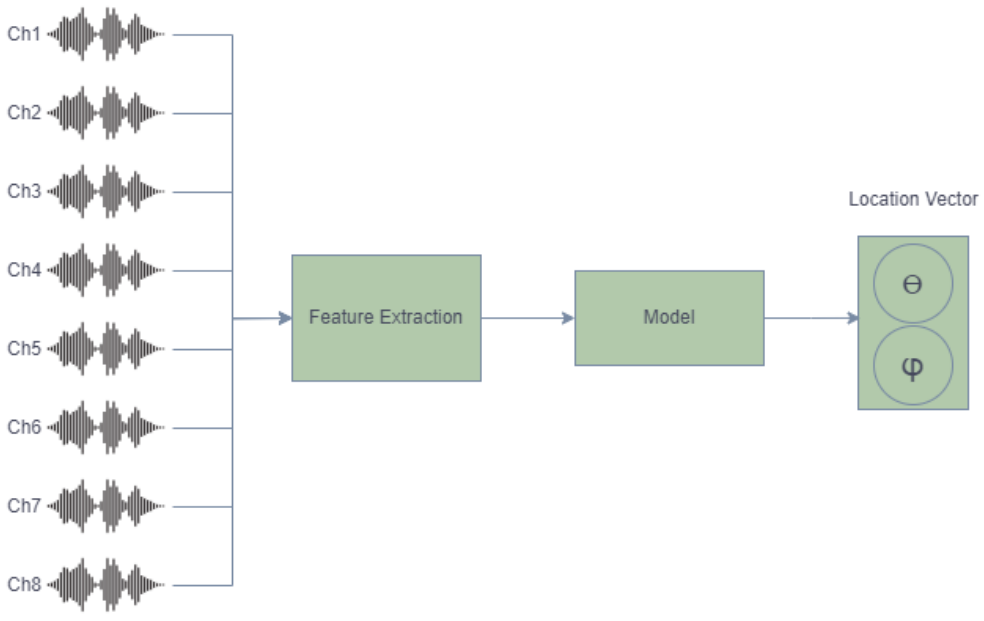


Figure 3.1 Block Diagram of the System

The training and test data are recorded in the same room. Every audio data recorded from predefied locations. 3.2 shows the environment that data are recorded and the local axis of the microphone array system for localization.

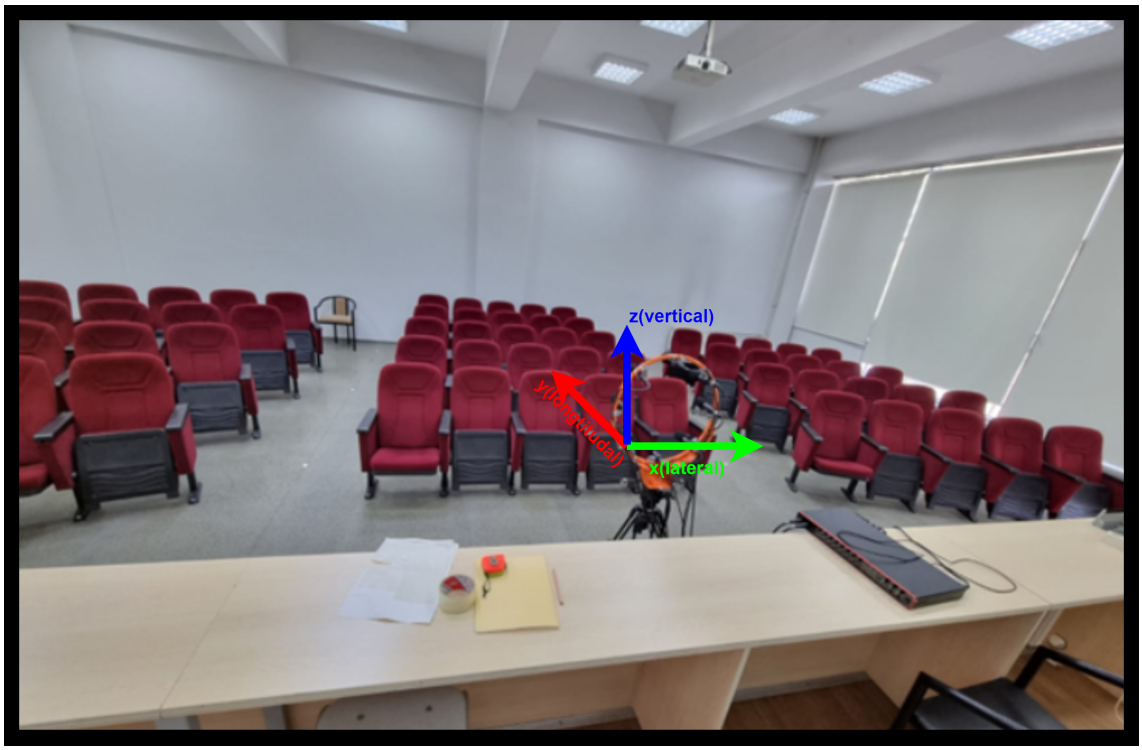


Figure 3.2 Sound Record Environment

3.1.1 Requirement Analysis

The requirements of the developed model are given below:

- Machine Learning and Deep Learning methods will be used and compared. The main comparison metrics are accuracy and performance.
- The models that will be developed will have 2D localization ability. The sound sensor mechanism that is used is compatible for 2D localization because all microphones are in the same plane.

For the hardwares that will be used in project, this specifications should be met:

- 8-channel input
- High frequency(96 kHz)

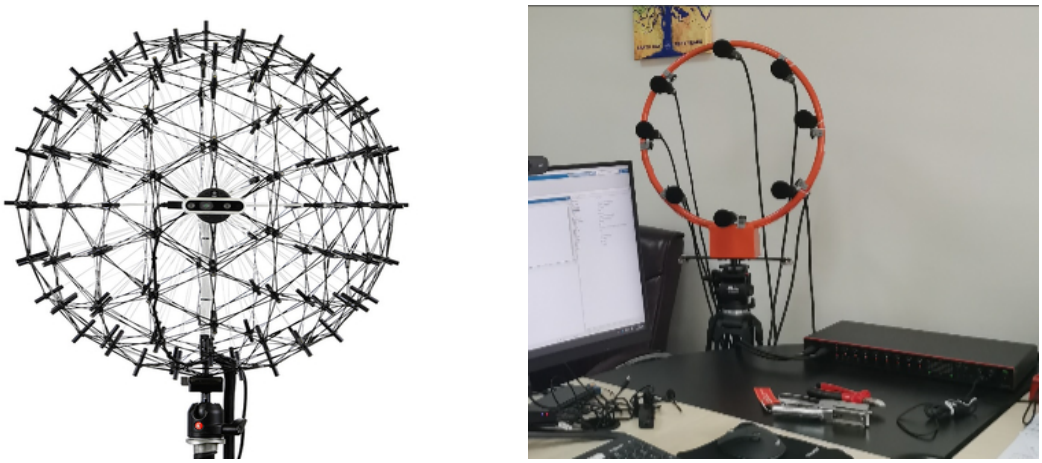
3.2 Feasibility

3.2.1 Technical Feasibility

In this section, hardwares and softwares that are used in this project is stated.

3.2.1.1 Hardware Feasibility

1. Microphone Array Commercial devices for this purpose are usually spherical, 3-dimensional and has lots of MEMS microphone that are integrated directly to the designed hardware. Due to these systems are expensive, cost effective new mechanism with 8 microphone that are lined up circular is developed. This two mechanism can be seen in 3.3



(a) (b)

Figure 3.3 (a)Commercial Products , (b)Designed Mechanism

- Focusrite Scarlett 18i20 Sound Card : For providing requirements such as getting 8-channel high frequency synchronous sound input, an advance sound card should be used. Details:
  - Multichannel XLR microphone input
  - Sound gain control
  - USB Interface
  - High Sampling Frequency up to 192 kHz.

  As can be seen in details, this sound card fulfills the requirements for the project.
- Chassis: Basketball Hoop and a tripod are used as chassis for microphone array system. The main reason for basketball hoop to be chosen is being circular and economic. Tripod used to keep microphones stand still and localize in vertical axis.
- Microphone: For the commercial products, usually MEMS microphones used and integrated within designed hardware. Because of the hardware design s not included in this project, any microphone that have compatible interface can be used.

2. Computer: Developed deep learning and machine learning models will be executed local machines. Because of the models and features will be used are not large, the computer which specs are given below is used in project.

    - Monster Abra A5 v15.5
    - Intel i7 10750H
    - 1 TB SSD
    - 16 GB RAM

3.2.1.2 Software Feasibility

- MATLAB 2024b Academic[4]
- MATLAB/Simulink Phased System Toolbox
- Phyton[5]
- Scikit-Learn[6]
- Tensorflow[7]
- Keras[8]

3.2.2 Legal Feasibility

The softwares that are used for project development are open source except MATLAB. For MATLAB, Academic License used. All rights of hardwares and dataset belongs to advisor of the project.

3.2.3 Economic Feasibility

Table 3.1 ECONOMIC FEASIBILITY

Item	Price
Focusrite Scarlett 18i20	$27000TL$
Basketball Hop	$900TL$
Tripod	$400TL$
Microphone($x8$)	$100-300TL$
Computer	pre-owned

3.2.4 Workforce and Time Feasibility

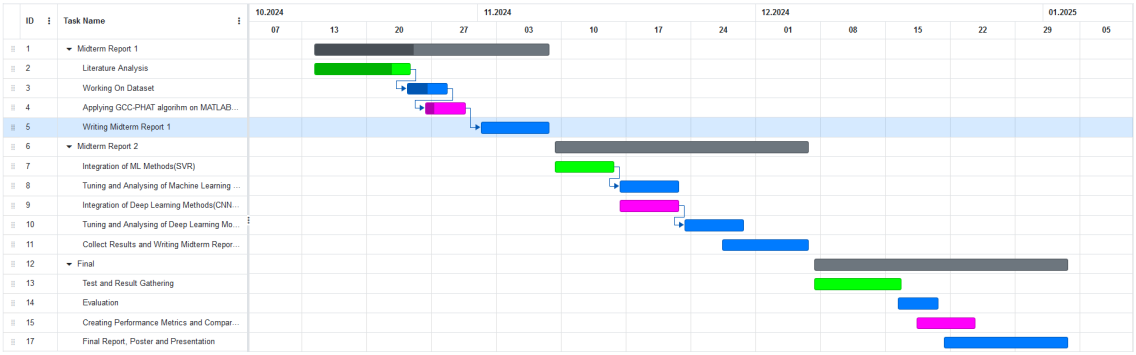


Figure 3.4 Gantt Diagram

4
SYSTEM DESIGN

4.1 Material and Dataset

For training model that solves the SSL problem, dataset should have multichannel sound inputs and all records are labeled with the locations of the sound source. The sound database has 441 records taken from different locations. All data are recorded with the mechanism that is mentioned in Section 3.2.1.

Every data in this dataset is a ".mat" file that includes a 490000x8 sized matrix. Every column represents a different channel. The data have a 96kHz sample rate. Data distribution of the total dataset is given in figure 4.1.

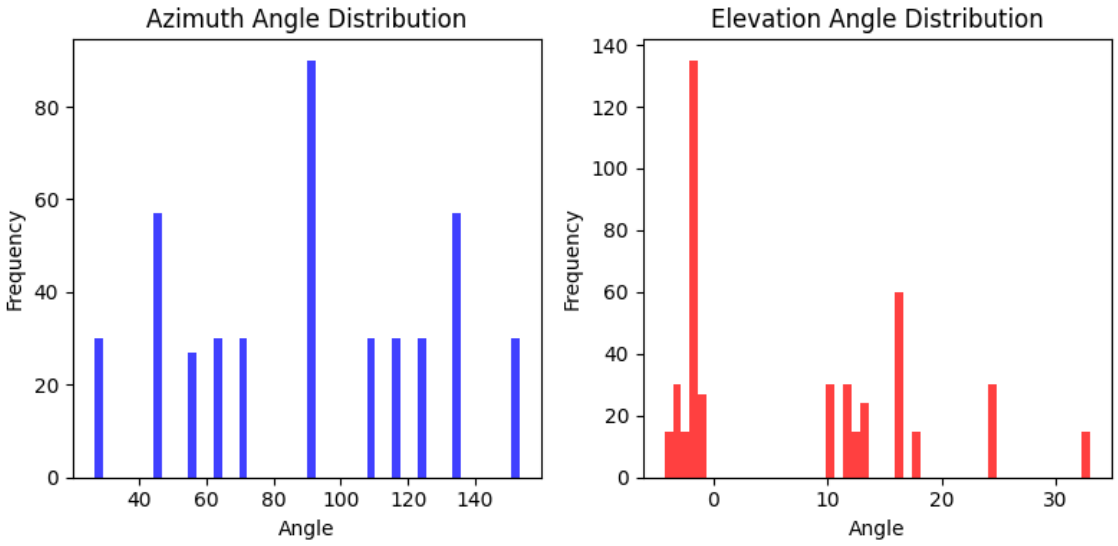


Figure 4.1 Angle Distributions of Elevation and Azimuth

4.2 Methods

4.2.1 Feature Extraction

ITD information between channels will be the base feature. To extract this information, cross-correlation method will be used. This method also used in

traditional methods.

Firstly, Fast Fourier Transform applied to every channel input. Then Cross-Correlation applied every two channel combination possible. IFFT operation applied to the array obtained with cross-correlation and lastly, a piece sized of 2231 is taken from final array arround maximum.

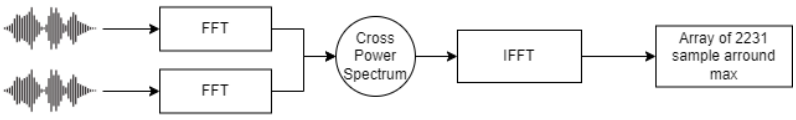


Figure 4.2 Feature Extraction

With the algorithm given in **??** maksimum time delay array between x and y channel extracted and $E_{28x2231}$ cross-correlation matrix that is mentioned in 4.1. Number of row represents combination of two of 8 channel.

$$M[x,y] = \begin{vmatrix} M_{1,1} & M_{1,2} & . & . & . & . & . & . & M_{1,2231} \\ M_{2,1} & M_{1,2} & . & . & . & . & . & . & M_{2,2231} \\ . & . & & & & & & & . \\ . & . & & & & & & & . \\ . & . & & & & & & & . \\ . & . & & & & & & & . \\ . & . & & & & & & & . \\ M_{28,1} & . & . & . & . & . & . & . & M_{28,2231} \end{vmatrix} \tag{4.1}$$

4.2.2 Machine Learning Approach

4.2.2.1 Support Vector Regresssion

For solving SSL problem with machine learning approach, Support Vector Machines is used. The main reason why SVMs to prefered is:

- Handles non-linear data better due to kernels.
- Effective with small dataset.
- Robust to overfitting problem.
- Effective in high dimensional spaces.

Despite these advantages, it is sensitive to noise, requires careful tuning of hyperparameters and computitionally expensive.

Between SVM methods, Support Vector Regression(SVR) will be used in this problem. Even if classification could be implemented in small range as well, SSL problem is likely a regression problem. So SVR basially solves the problem given in equation 4.2.

$$\begin{aligned} \min_{w,b,\zeta,\zeta^*} \frac{1}{2} w^T w + C \sum_{i=1}^{n} (\zeta_i + \zeta_i^*) \\ \text{subject to } y_i - w^T \phi(x_i) - b \leq \varepsilon + \zeta_i, \\ w^T \phi(x_i) + b - y_i \leq \varepsilon + \zeta_i^*, \\ \zeta_i, \zeta_i^* \geq 0, i = 1, ..., n \end{aligned} \tag{4.2}$$

In this equation all sample predictions penaltized if the prediction at least epsilon distant far from the correct value[6].

4.2.2.2 Random Forrest Regression

Random Forrest Regressor is preferred as second Machine Learning approach to the SSL problem. The main reasons are:

- Understands non-linear relationships.
- Robust to overfitting problem.

But the main challenge is it requires properly preprocessed features. Because of we have preprocessed data, Random Forrest Algorithm can be succesfully applied.

The Random Forest Regressor is a simply a decision tree. With the random forest architecthture, it provides multiple decision tree and combines their result for more precise results[9]. This called bagging or boostrap aggregation that is shown in figure 4.3.

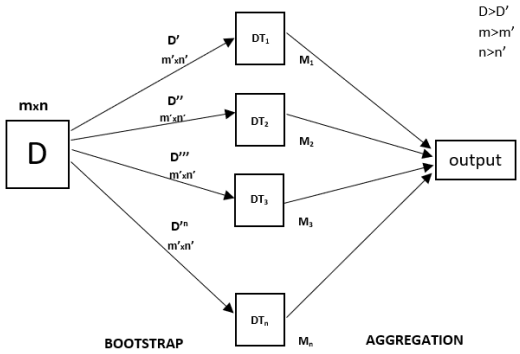


Figure 4.3 Bootstrap Aggregation in Random Forest[9]

4.2.3 Deep Learning Approach

Deep Learning achieves state-of-art performance and versatility at different domains such as speech recognition, computer vision, autonomous systems etc. It is also used for many localization problems. For SSL case, CNN architecture will be used due to ability of handle multi-dimensional input, efficiency and robustness.

For convolution layer, it is planned to use 2 convolution blocks with ReLu activation function. Because of the feature data is an 8x8 matrix, 2 blocks with 3x3 kernel expected to be sufficient for model to handle. If overfitting occurs due to small data size, number of convolutional block will be reduced to 1. The architecture of the designed model could be seen in 4.4

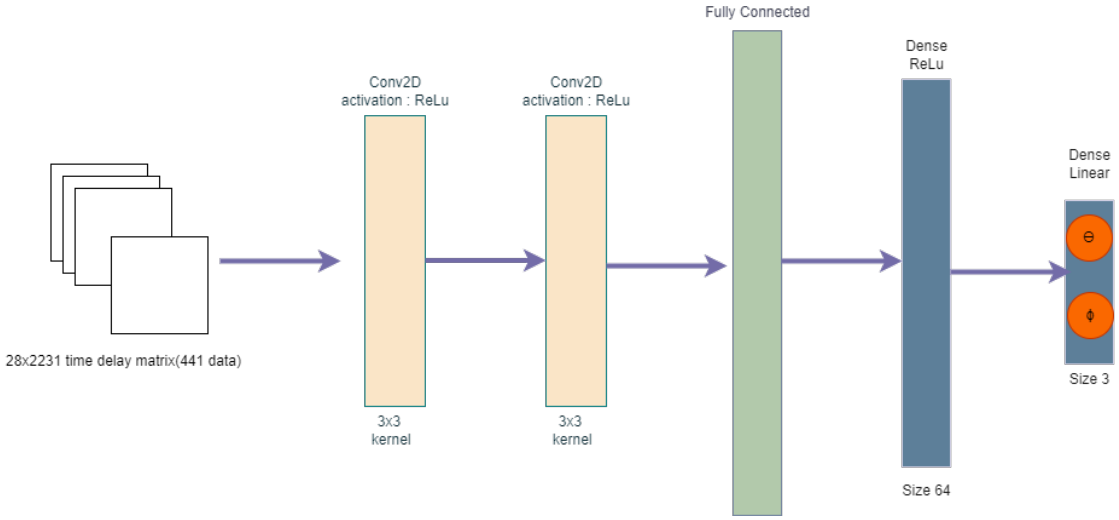


Figure 4.4 The architecture of designed CNN model

4.2.4 Evaluation Methods

Two base methods are used to evaluate success of the models:

- Mean Square Error(MSE): MSE is the average squarred distance between predicted and actual value. It magnifies the penalty score of larger errors than smaller ones which is highlight significant deviations.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4.3}$$

- Mean Absolute Error(MAE): MAE gives the average absolute error between predicted and actual value. It is easy to interpret because penalty has the same unit with the prediction and actual value.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4.4}$$

This metrics are also commonly used in regression problems. They are easy to interpret and combination of this two method gives clue about the source of errors.

5
EXPERIMENTAL RESULTS

5.1 Performance Analysis

5.1.1 CNN

Metric	Value
Average MSE	6.4419
Average MAE	1.6904
Azimuth MSE	11.1696
Azimuth MAE	2.3665
Elevation MSE	1.7142
Elevation MAE	1.0143

Table 5.1 Performance Results of the CNN

Relevant Visuals:

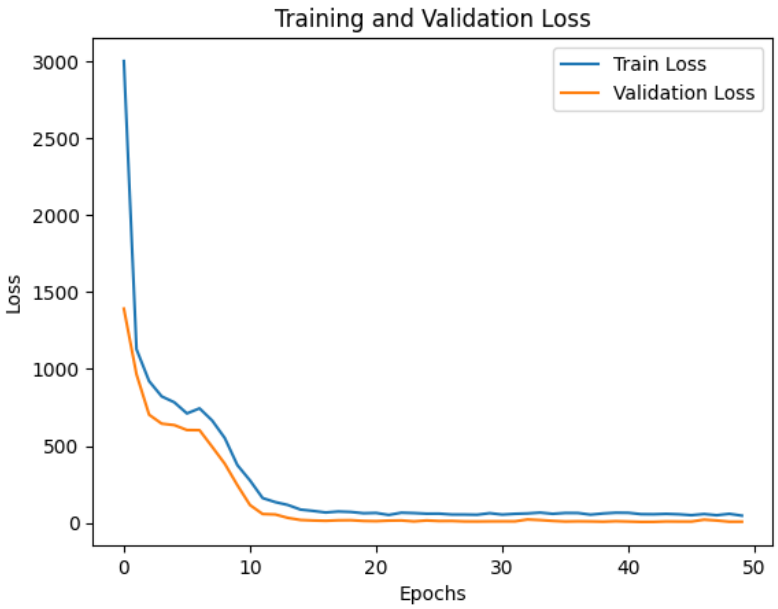


Figure 5.1 Training and Validation Loss for CNN

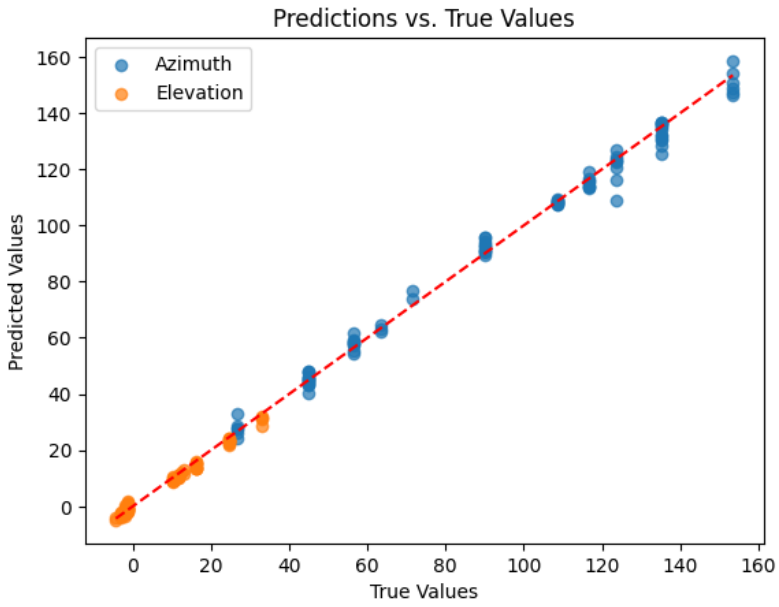


Figure 5.2 Predictions vs. True Values for CNN

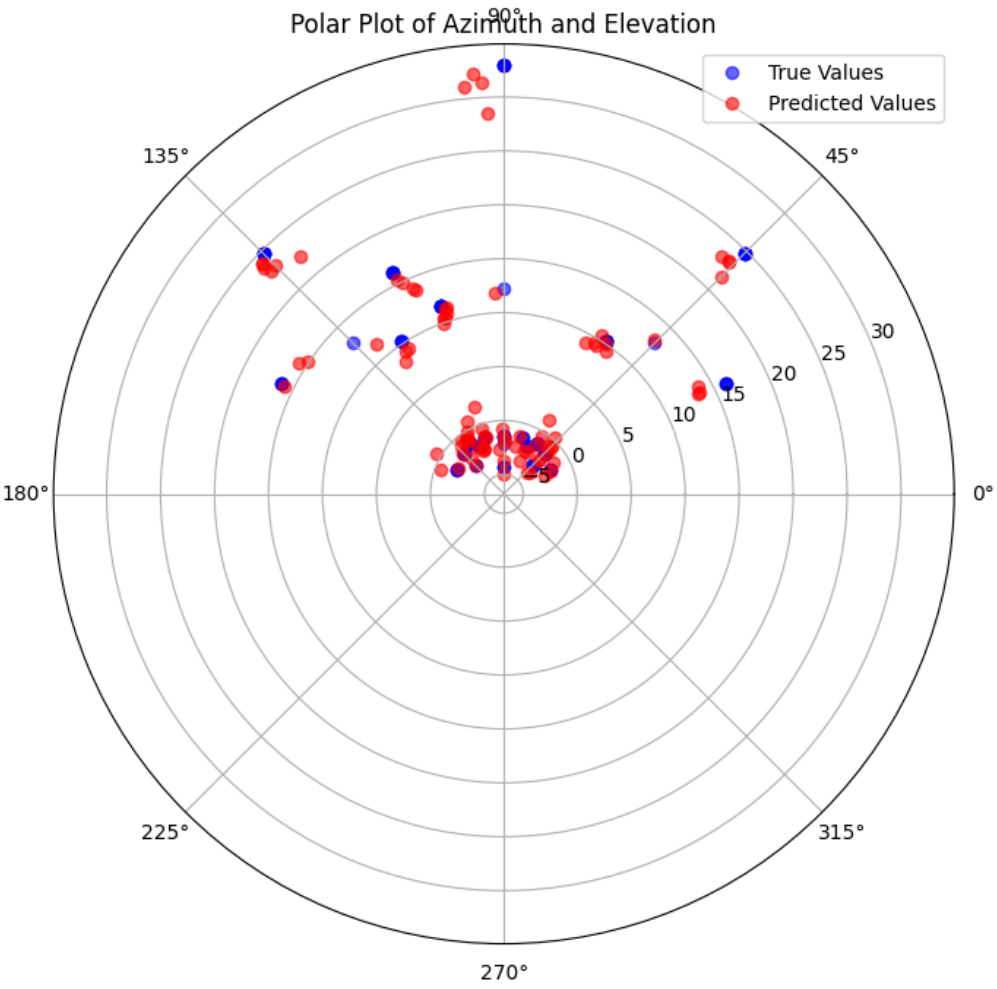


Figure 5.3 Polar Plot of Azimuth and Elevation

5.1.2 CNN with Cross-Validation

Fold No	MSE	MAE	Azimuth MSE	Elevation MSE	Azimuth MAE	Elevation MAE
1	5.6647	1.7542	9.3715	1.9579	2.3963	1.1121
2	4.9386	1.6613	7.3581	2.5192	2.1684	1.1542
3	9.6957	2.0770	15.5601	3.8312	2.6363	1.5176
4	21.5674	3.2267	39.2359	3.8989	4.7942	1.6592
5	15.7823	2.5266	27.5102	4.0545	3.4243	1.6290

Table 5.2 Performance Scores for Each Fold of CNN Cross-Validation

Metric	Value
Average MSE	11.5297 (±6.3)
Average MAE	2.2492 (±0.5)
Azimuth MSE	19.8072
Azimuth MAE	3.0839
Elevation MSE	3.2523
Elevation MAE	1.4144

Table 5.3 Average Performance Results of CNN Cross-Validation

Relevant Visuals:

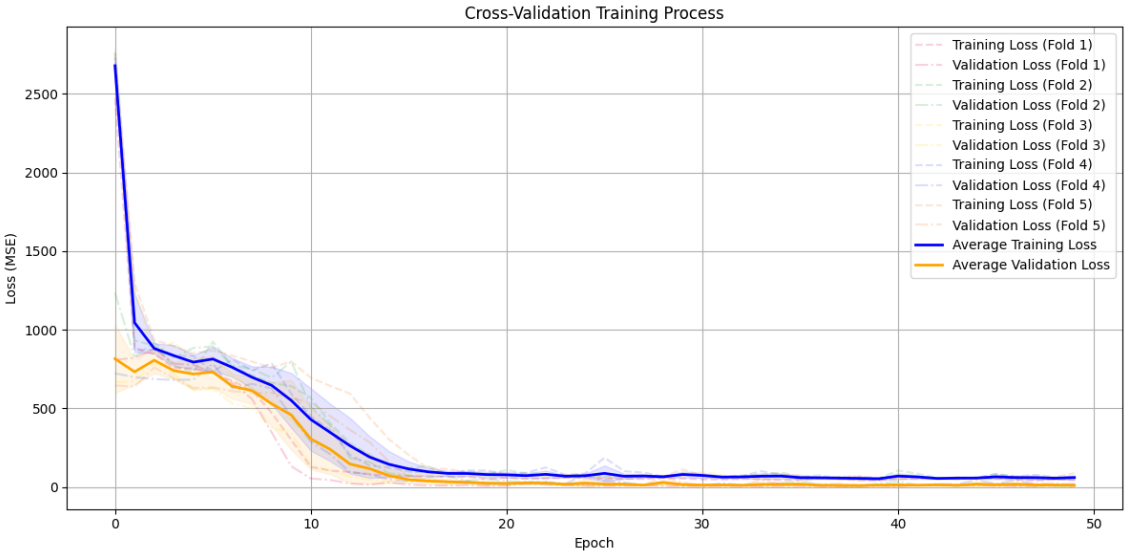


Figure 5.4 Cross-Validation Training Process

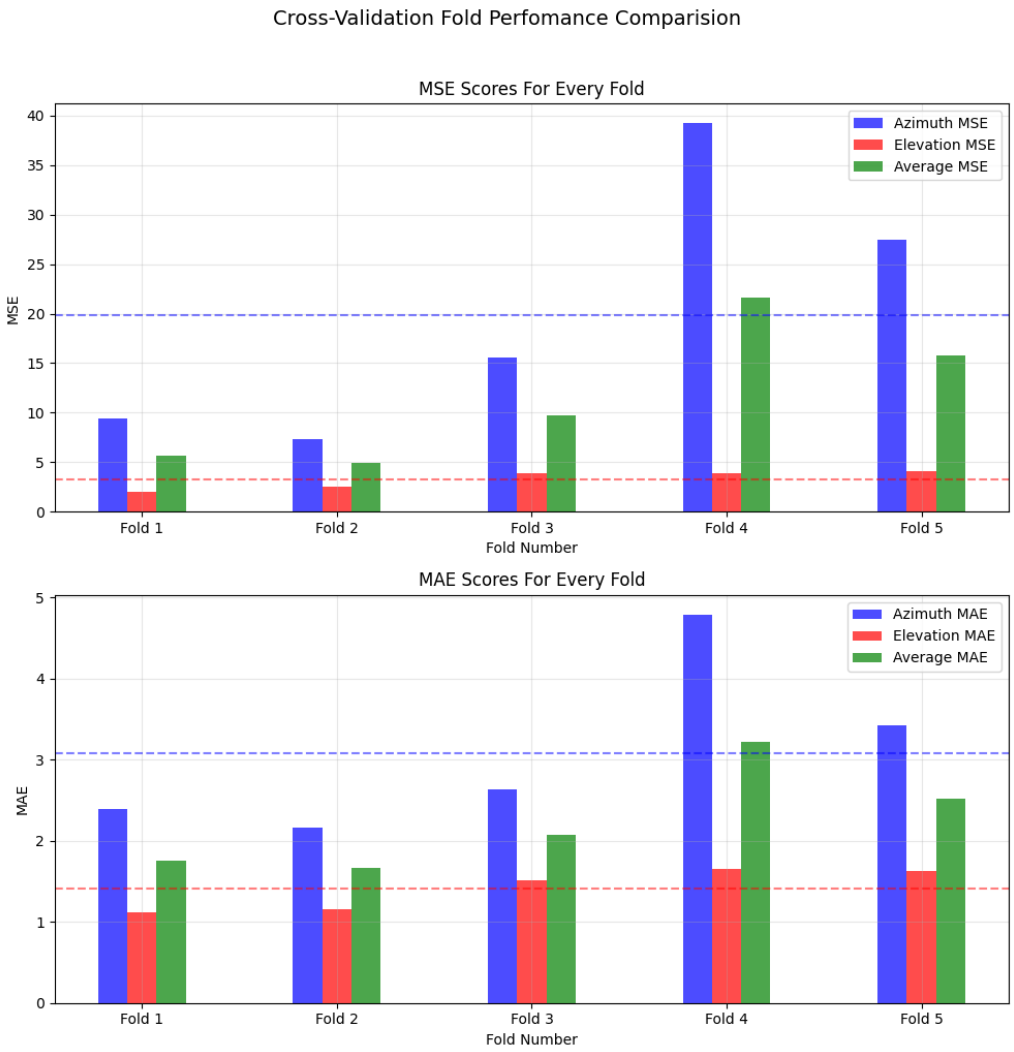


Figure 5.5 MSE and MAE Scores Across All Folds

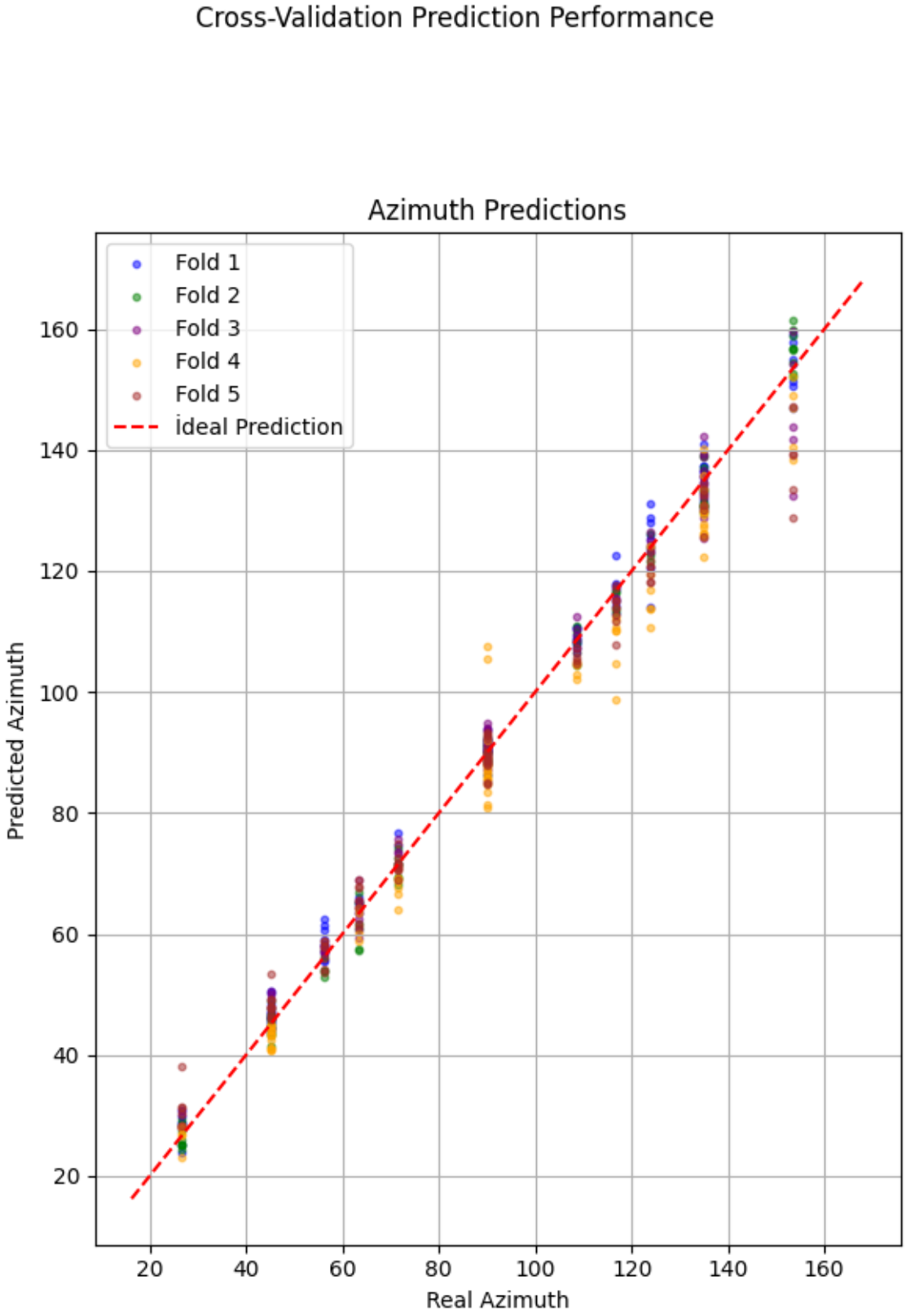


Figure 5.6 Azimuth Predictions

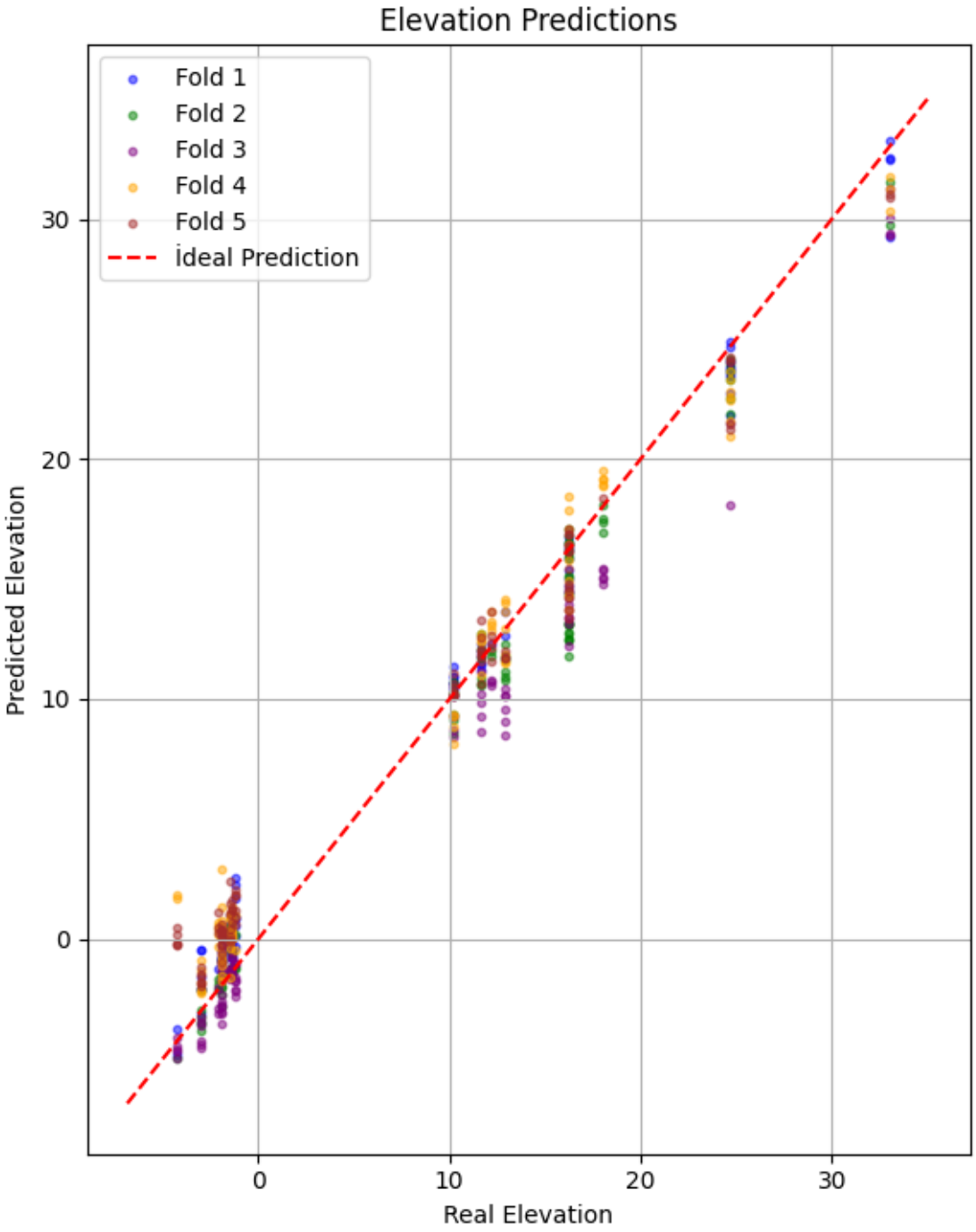


Figure 5.7 Elevation Predictions

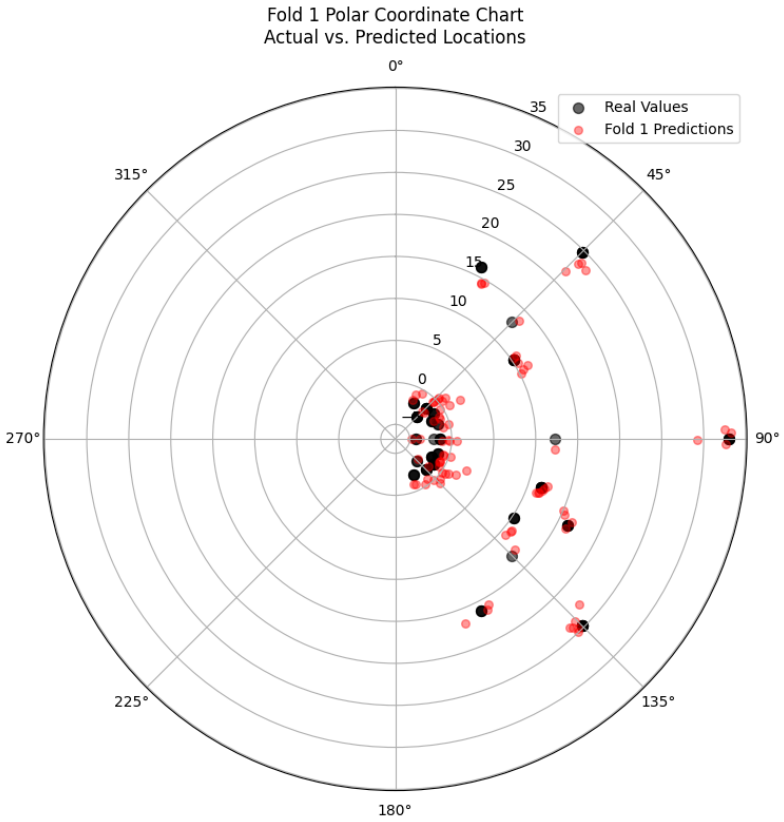


Figure 5.8 Polar Coordinate Chart - Fold 1

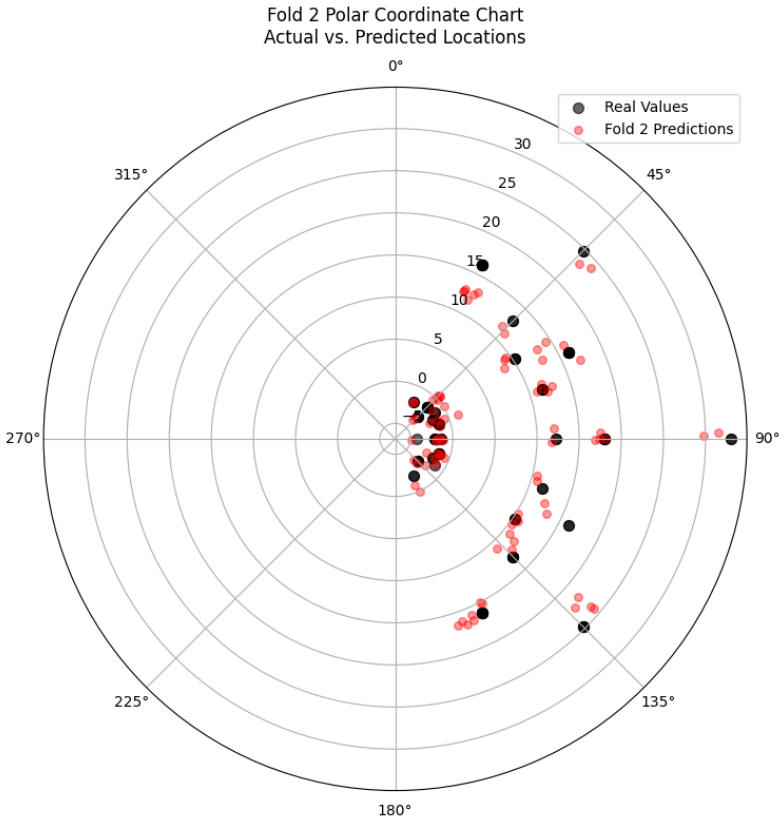


Figure 5.9 Polar Coordinate Chart - Fold 2

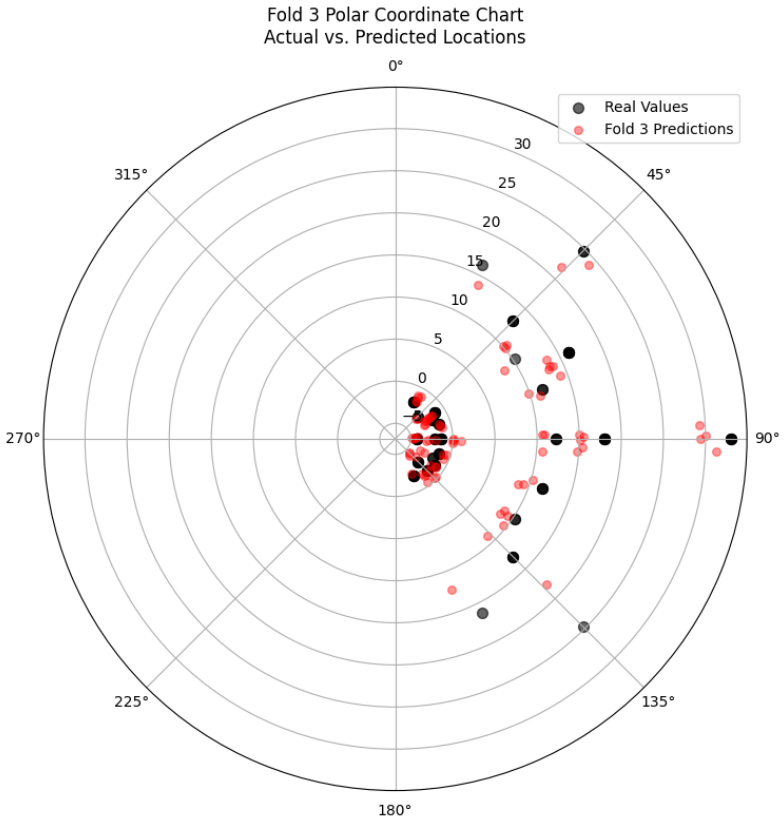


Figure 5.10 Polar Coordinate Chart - Fold 3

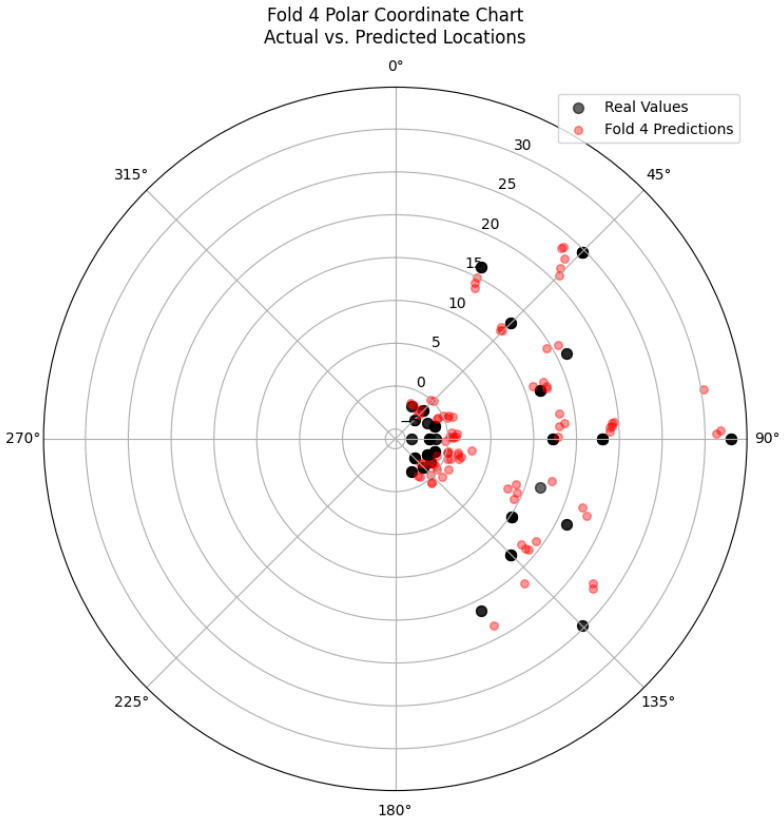


Figure 5.11 Polar Coordinate Chart - Fold 4

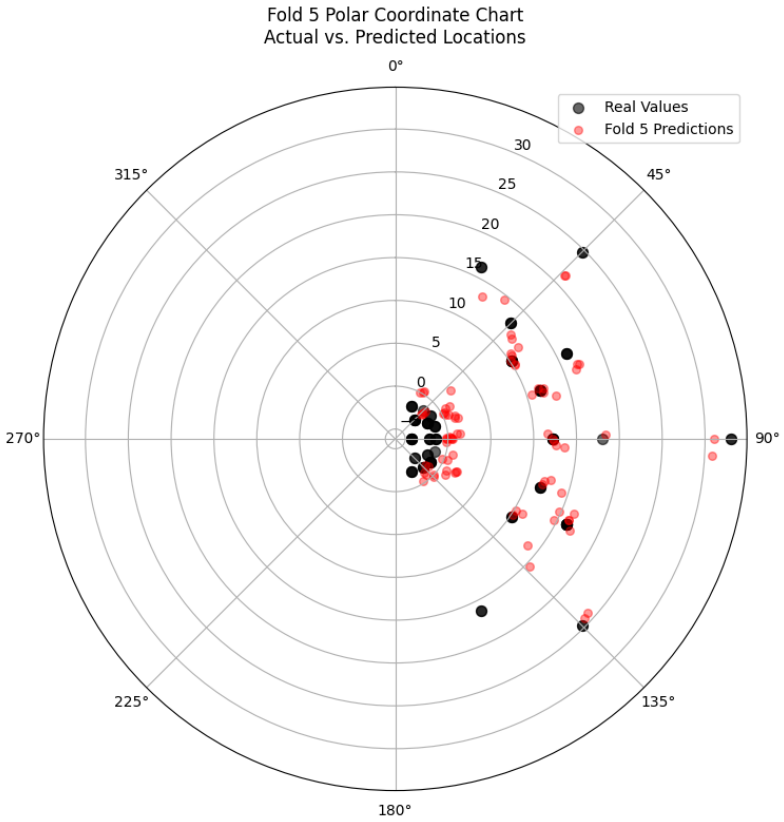


Figure 5.12 Polar Coordinate Chart - Fold 5

5.1.3 CNN with Residantial Connection

Metric	Value
Test Loss MSE	10.8464
Average MAE	2.2312
Azimuth MSE	18.7513
Azimuth MAE	3.0938
Elevation MSE	2.9417
Elevation MAE	1.3686

Table 5.4 Performance Results of CNN with Residential Connection

Relevant Visuals:

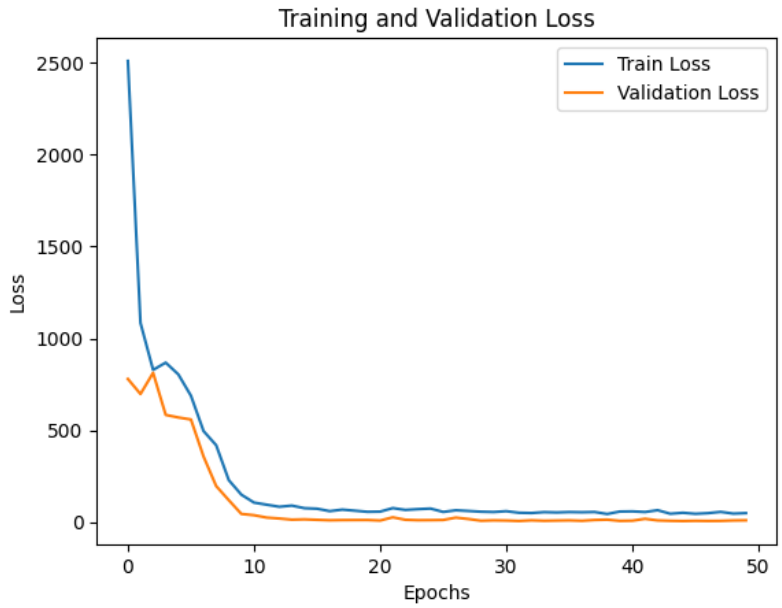


Figure 5.13 Training and Validation Loss for CNN with Residential Connection

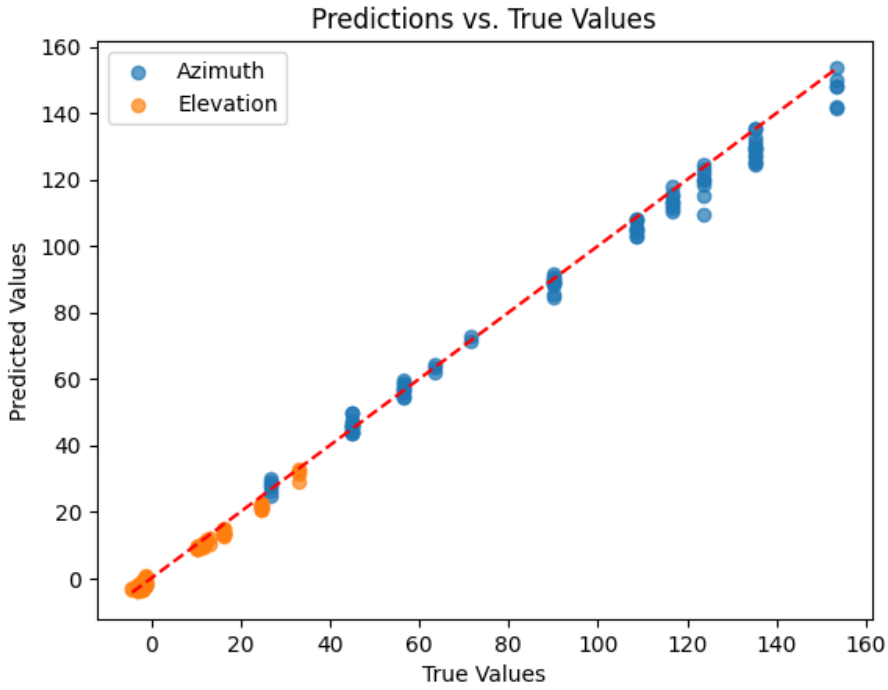


Figure 5.14 Predictions vs. True Values for CNN with Residential Connection

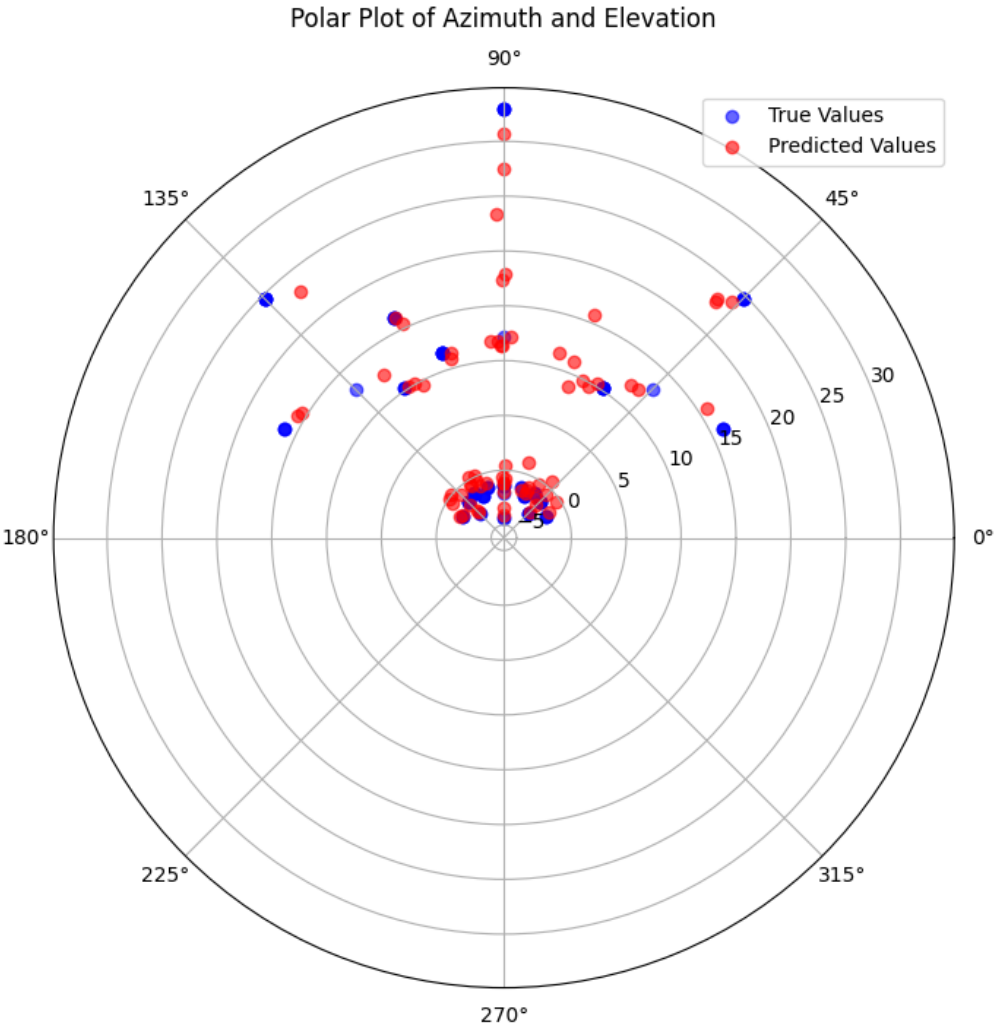


Figure 5.15 Polar Plot of Azimuth and Elevation for CNN with Residential Connection

5.1.4 Random Forest Regressor

Metric	Azimuth	Elevation
MSE	25.7683	3.6532
MAE	4.0019	1.2803

Table 5.5 Performance Results of Random Forest

Relevant Visuals:

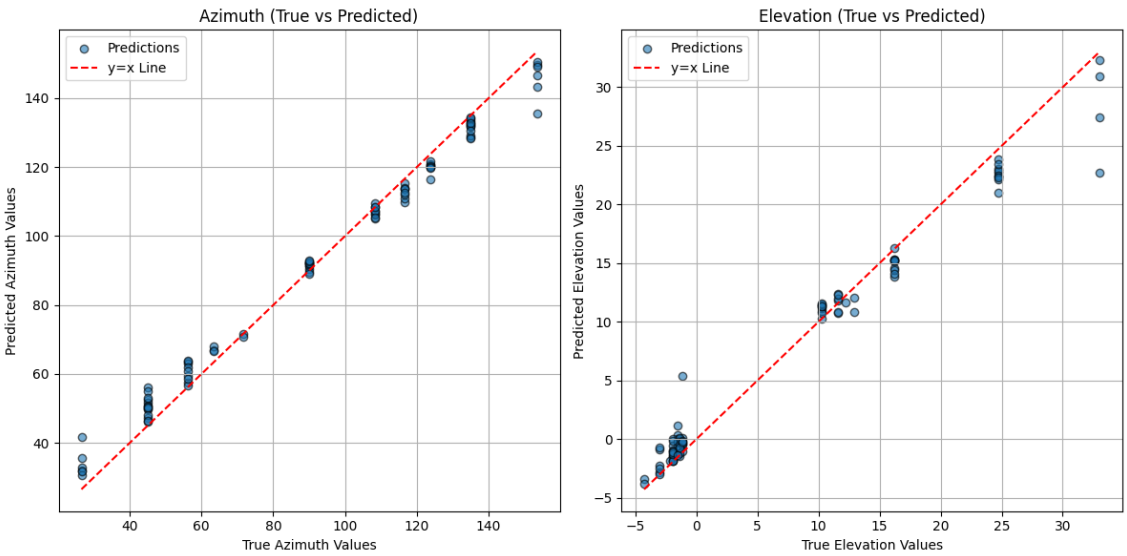


Figure 5.16 True vs. Predicted Elevation and Azimuth for Random Forest

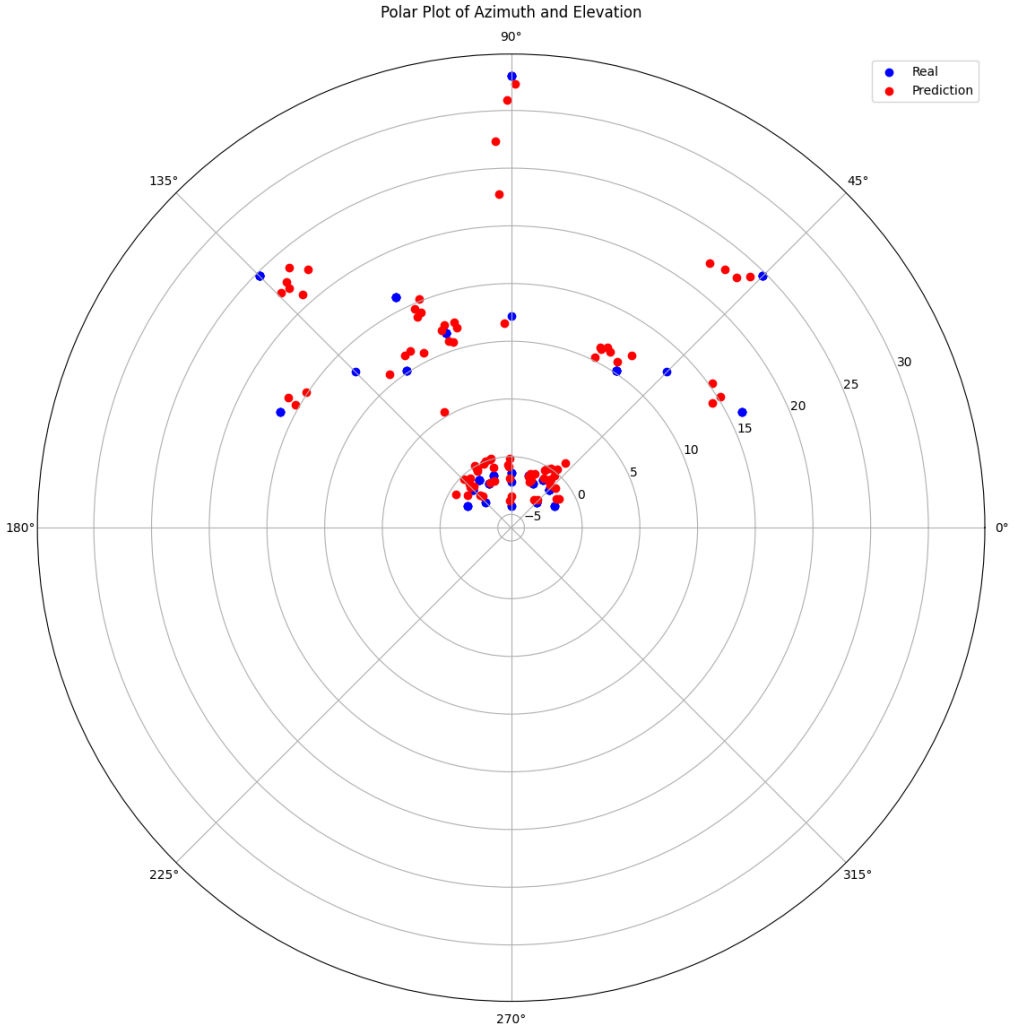


Figure 5.17 Polar Plot of Azimuth and Elevation for Random Forest

5.1.5 Support Vector Regression (SVR)

Metric	Azimuth	Elevation
MSE	15.2703	1.5814
MAE	2.8657	0.8804

Table 5.6 Performance Results of Support Vector Regression (SVR)

Relevant Visuals:

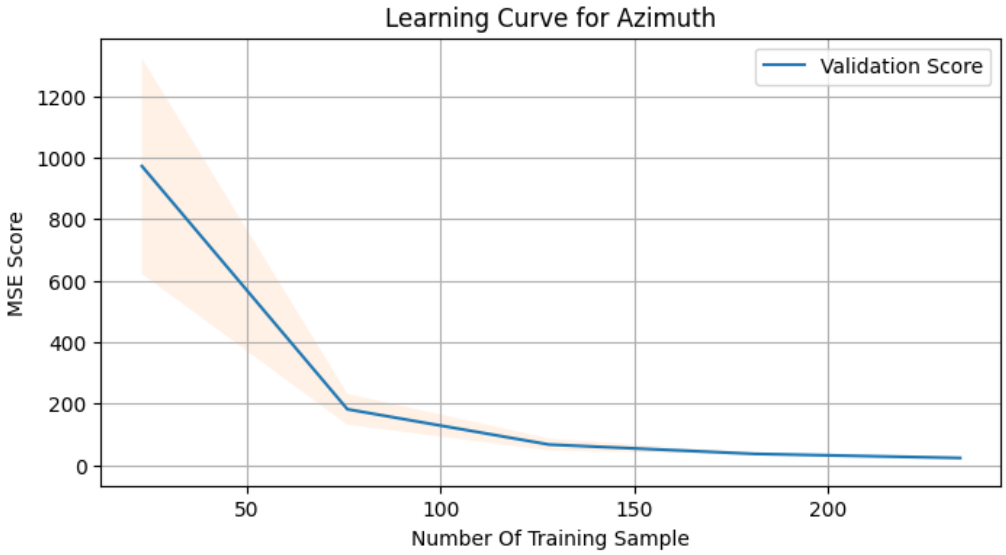


Figure 5.18 Learning Curve for Azimuth in SVR

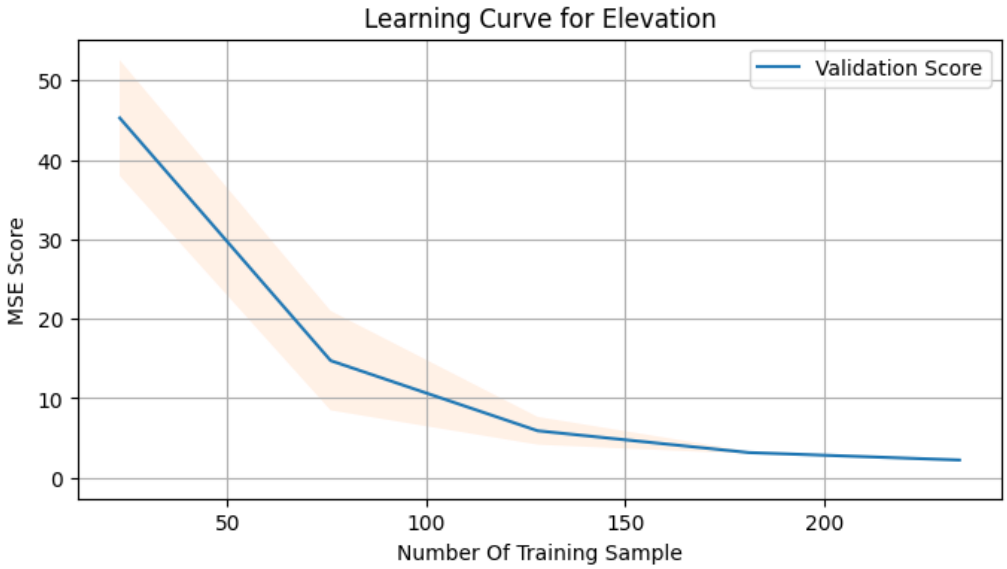


Figure 5.19 Learning Curve for Elevation in SVR

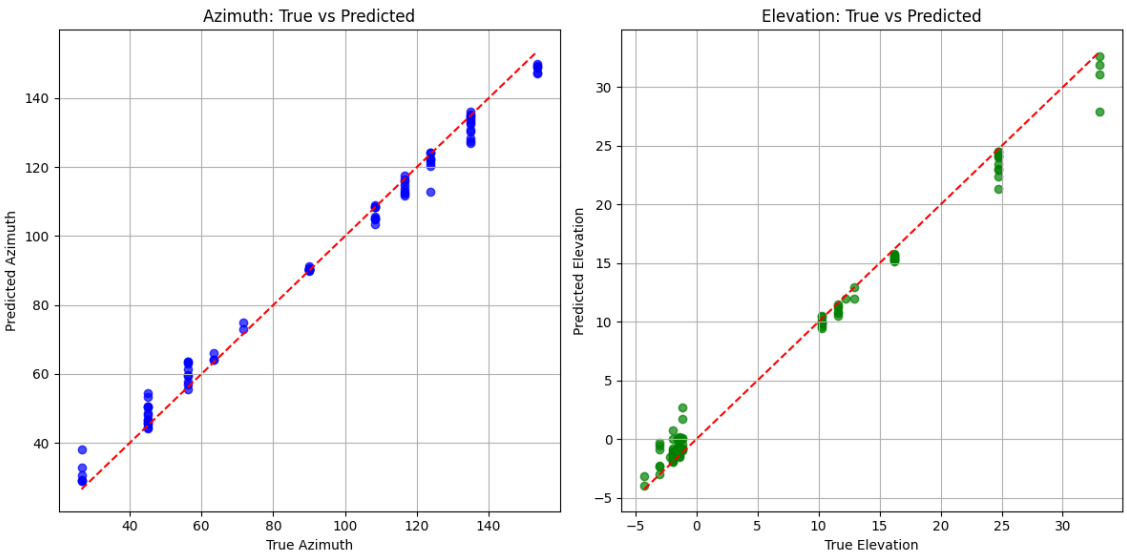


Figure 5.20 True vs. Predicted Azimuth and Elevation for SVR

Figure 5.21 Polar Plot of Azimuth and Elevation for SVR

5.2 Comparative Analysis

In this section, we analyze the performance of different models, namely CNN, CNN with Cross-Validation, CNN with Residual Connections, Random Forest, and Support Vector Regression (SVR), based on metrics such as Mean Squared Error (MSE) and Mean Absolute Error (MAE).

5.2.1 Overall Performance

Model	Average MSE	Average MAE	Azimuth MSE	Azimuth MAE	Elevation MSE	Elevation MAE
CNN	6.4419	**1.6904**	11.1696	2.3665	1.7142	1.0143
CNN (Cross Validation)	11.5297	2.2492	19.8072	3.0839	3.2523	1.4144
CNN (Residual Connections)	10.8464	2.2312	18.7513	3.0938	2.9417	1.3686
Random Forest	14.71075	2.6411	25.7683	4.0019	3.6532	1.2803
SVR	8.42585	1.87305	15.2703	2.8657	**1.5814**	**0.8804**
GCC-Phat	**6.415**	1.735	**6.37**	**1.61**	6.46	1.86

Table 5.7 Model Performance Comparison

5.2.2 Key Observations

CNN:

- Has an average MSE of 6.4419 and an average MAE of 1.6904, showing outstanding overall performance.
- It provides the smallest Average MAE of all the models, although it does not get the absolute lowest Average MSE (the GCC-Phat is marginally lower at 6.415).
- With an elevation MSE of 1.7142 and an azimuth MSE of 11.1696, it is competitive from both perspectives.

CNN (Cross Validation):

- Comparatively to the normal CNN, a larger Average MSE (11.5297) and MAE (2.2492) are shown most likely due of fold variability.
- Although cross-validation usually improves robustness, it seems less effective than the baseline CNN in this dataset.

Random Forest:

- The error rates are relatively high, with an average MSE of 14.71075 and MAE of 2.6411.

- Its Azimuth MSE is particularly high (25.7683), indicating a lower appropriateness compared to neural network-based techniques.

SVR:

- Provides a moderate average MSE of 8.42585 and MAE of 1.87305.
- Notably, it outperforms the CNN in elevation measures (1.5814 MSE, 0.8804 MAE). However, it is still less effective overall than CNN and GCC-Phat.

GCC-Phat:

- Achieves the lowest average MSE (6.415) among all tested models and excels in azimuth metrics (6.37 MSE, 1.61 MAE).
- Its Elevation performance (6.46 MSE, 1.86 MAE) is lower, bringing its Average MAE to 1.735, little more than that of the CNN.

5.2.3 Conclusion

The GCC-Phat and the standard CNN both seem like excellent choices. CNN performs well in both azimuth and elevation, while GCC-Phat has the lowest average MAE and the highest azimuth accuracy. SVR performs worse than CNN and GCC-Phat when all metrics are averaged, although showing promise in the Elevation dimension. By contrast, Random Forest generates the highest total number of mistakes.

6
CONCLUSION AND DISCUSSION

In conclusion, in this research project, different approaches on SSL problem are applied and compared with traditional GCC-PHAT method. SVR and Random Forest Regression methods are implemented with different parameters and optimal results acquired. Also CNN model was created and applied with different training methods and different connections such as cross validation training and residual connections. Addintionally complex network structures like CNN-LSTM hybrid model and extra layers like batch normalization are implemented. Mostly when model been too complex, training section encountered with overfitting problem.

Overall, CNN performs well in both azimuth and elevation, while GCC-Phat has the lowest average MAE and the highest azimuth accuracy. SVR performs worse than CNN and GCC-PHAT when all metrics are averaged, although showing most precise result in the Elevation dimension. On the other hand Random Forest has the worst performance on azimuth.

As a result, the CNN model remains a strong and dependable option, especially if lower mean absolute error is prioritized. If minimizing MSE or maximizing the azimuth dimension is crucial, GCC-Phat has a modest advantage. For specific circumstances when elevation accuracy is critical, SVR's results may be useful.

References

[1] G. Meng, C. Yang, H. Guo, and Y. Wang, "Method and practice of microphone array speech source localization based on sound propagation modeling," *Applied Mathematics and Nonlinear Sciences*, vol. 9, Oct. 2024. DOI: 10.2478/amns-2024-2681.

[2] J. Xu, B. Li, Y. Zhao, and S. Xue, "Sound source localization based on data and neural network model," *Journal of Physics: Conference Series*, vol. 2816, p. 012 085, Aug. 2024. DOI: 10.1088/1742-6596/2816/1/012085.

[3] C. Çatalbaş and S. Dobrišek, "Dynamic speaker localization based on a novel lightweight r-cnn model dynamic speaker localization based on a novel lightweight r-cnn model," *Neural Computing and Applications*, Jan. 2023. DOI: 10.1007/s00521-023-08251-3.

[4] *Matlab*, https://www.mathworks.com/products/matlab.html, Accessed: 2024-7-19.

[5] *Python*, https://www.python.org, Accessed: 2024-8-6.

[6] *Scikit-learn*, https://scikit-learn.org/stable/, Accessed: 2024-8-7.

[7] *Tensorflow*, https://scikit-learn.org/stable/, Accessed: 2024-8-7.

[8] *Keras*, https://keras.io, Accessed: 2024-8-6.

[9] e. a. Comment, *Random forest regression in python*, GeeksforGeeks, Sep. 2024. [Online]. Available: https://www.geeksforgeeks.org/random-forest-regression-in-python/.

Curriculum Vitae

FIRST MEMBER

Name-Surname: Umut GÜZEL
Birthdate and Place of Birth: 10.09.2000, İstanbul
E-mail: umut.guzel@std.yildiz.edu.tr
Phone: 0545 337 95 67
Practical Training: SANLAB SİMÜLASYON Gömülü Sistem Ekibi

SECOND MEMBER

Name-Surname: Mehmet ÇALOĞLU
Birthdate and Place of Birth: 17.04.2000, Ordu
E-mail: mehmet.caloglu@std.yildiz.edu.tr
Phone: 0546 744 76 15
Practical Training: Nevalabs Frontend Ekibi

Project System Informations

System and Software: Windows İşletim Sistemi, Python, tensorflow, NumPy, Scikit Learn
Required RAM: 2GB
Required Disk: 2.5GB